AGENCY AND INFLUENCE: MORAL COMMUNICATION FOR INTERACTIVE ROBOTS

by

Ryan Blake Jackson

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 Doctor of Philosophy (Computer Science).

Golden, Colorado

Date _____

Signed: _____

Ryan Blake Jackson

Signed: _____

Dr. Tom E. Williams
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____

Dr. Tracy Camp
Professor and Department Head
Department of Computer Science

ABSTRACT

As robots with social behaviors proliferate into a widening variety of contexts and roles, it is clear that we have a lot to learn about how humans expect (and prefer) these robots to act, how humans perceive different robot behaviors and judge or sanction robot misbehaviors, and how robots should fit into, shape, and be shaped by social structures and norms. This thesis presents several studies on human-robot interaction that focus on enabling robots to communicate effectively and appropriately through natural language in morally sensitive contexts.

We begin by examining the concept of *social agency*, and constructing a new theoretical understanding of social agency for robots. We discuss the implications of robots' potential ontological status as social agents, including the capacity for significant normative influence. We then examine this moral influence in the context of clarification dialogues, and show how a failure to perform moral reasoning when generating clarification requests can cause robots to generate utterances with unintended implied meanings that can weaken human application of moral norms. We then present and evaluate an algorithm that fixes this problem.

Next, we examine robot command rejections under the premise that robots should not follow immoral human commands. We present evidence that robot command rejections should be phrased with a degree of politeness *proportional* to the severity of the norm violation motivating the command rejection. Given the importance of gender in performing and perceiving politeness, we reexamine these results with specific attention to human gender and robot gender presentation.

We then present part of a cross-cultural study on how female presenting social robots might respond to gendered verbal abuse from humans without propagating harmful sexist stereotypes or damaging robot credibility. Our results highlight a couple of promising response styles.

Finally, we present the integration of a norm-aware task planner and a context recognition module into a robot cognitive architecture. This integration establishes the capacity for multi-step task planning under context-sensitive norms and lays the groundwork for generating more informative command rejections.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

(Repeated Measures) Analysis of variance . . . . . (RM-)ANOVA

Agent Development Environment . . . . . ADE

Analysis of covariance . . . . . ANCOVA

Application Programming Interface . . . . . API

Bayes factor . . . . . Bf

Distributed, Integrated, Affect, Reflection, Cognition . . . . . DIARC

Human-Robot Interaction . . . . . HRI

Level of Abstraction . . . . . LoA

Light Detection and Ranging . . . . . LiDAR

Robot Operating System . . . . . ROS

principle of procreative beneficence . . . . . PPB

unified socially-aware navigation . . . . . USAN

voxel-based representation learning . . . . . VBRL

ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Dr. Tom Williams, for introducing me to the amazing field of human-robot interaction and for giving me the opportunity to complete the research presented in this thesis. I am extremely lucky to have had an advisor willing to allow me to bring my own personal areas of interest into my research to the extent that I have. Working alongside Tom has truly been a pleasure. It is extremely rare to meet somebody with Tom's combination of passion, kindness, superhuman professional dedication, moral integrity, and intellect. I truly believe that with an advisor any less superlative, I would have dropped out of this program several times over to live in my car and climb rocks all day (for better or worse). Tom has played a critical role in my development as a researcher, thinker, and educator, and I am extremely grateful.

I would also like to thank Dr. Tracy Camp, who introduced me to the world of academic computer science at Mines over a decade ago, and whose mentorship has been incredibly valuable since. One of the pivotal moments of my life thus far was when, shortly before I finished undergrad and aimlessly entered the "real world", Tracy took the time out of her unrelentingly busy schedule to explain to me the benefits of pursuing a graduate degree. It's funny how a single generous act from a trusted mentor can completely define the course of one's life. Thank you, Dr. Camp.

Of course, I would also like to thank the rest of my dissertation committee, and apologize for making you read a good portion of this thesis twice, given the length of my proposal. Thank you Drs. Qin Zhu, Hao Zhang, and Bertram Malle. I really appreciate your time.

I would like to thank all of my research collaborators and coauthors. Dr. Katie Winkle in particular is a consistently inspiring scholar, brilliant conversationalist, and valued friend.

I want to thank the faculty in the math and computer science department at Colorado College, especially Matthew Whitehead, Ben Ylvisaker, Amelia Taylor, and Steven Janke, who showed me the beauty of intellectual curiosity for its own sake. I hope that I can one day give to my students what you gave to me.

Thanking all of the people in my personal life who have helped me through the process of completing this dissertation and the other facets of life over the last few years would take several pages, and nothing that I could articulate here would adequately describe the love and appreciation that I feel. I will simply say that I am incredibly lucky to have (and to have had) so many kind, selfless, and inspiring people in my life keeping me alive, happy, and relatively sane. Thank you all, and a special thanks to those whose generosity and caring allowed me to heal from an injury that I thought might have been insurmountable.

CHAPTER 1

INTRODUCTION

This thesis explores various aspects of the unique sociotechnical ontological niche somewhere between socially agentic community member and lifeless technological tool that social robots occupy. As robots with social behaviors increasingly proliferate into a widening variety of contexts and roles, it is becoming clear that we still have a lot to learn about how humans expect (and prefer) these social robots to act, how humans perceive different robot behaviors and judge or sanction robot misbehaviors, and how robots should fit into, shape, and be shaped by social structures and norms. The field of Human-Robot Interaction (HRI) seeks to answer these types of questions with a broad repertoire of interdisciplinary approaches. As HRI researchers, we combine computer science, robotics, social psychology, linguistics, moral philosophy, and other academic disciplines to better understand the human element of human-robot interaction, so that we may better design the robotic element.

My work in this dissertation is particularly concerned with morally relevant facets of HRI. Therefore, it draws on theories and methods from robot ethics and moral psychology. Many of the contexts in which social robots are currently being deployed (or developed for deployment) are very morally sensitive, including eldercare [1, 2], mental health treatment [3], childcare [4], and military operations [5–7]. Just as robot actions in these types of contexts could have serious moral consequences, so too could robot (mis)communications with humans. Moral communication is thus a critical component of moral competence [8], and much of this thesis is working towards developing autonomous moral communication.

One of the reasons why moral communication is so important in social robots is because of their unique sociotechnical ontological status, which grants them significant persuasive capacity and normative influence over their human interactants (see, for example, [9, 10]). Human morality is dynamic and malleable [11], and the dynamic norms that inform human morality are defined and developed not only by human community members, but also by the technologies with which they interact [12, 13]. However, the capacity for social robots to be considered moral and social agents, we argue, makes their normative influence qualitatively different than that of other technologies. With great normative influence comes great responsibility, and part of this thesis is concerned with avoiding unintentionally altering human application of moral norms with imprecise robot speech. However, we also believe that robots can be designed to wield their normative influence purposefully and prosocially.

The task of carefully designing moral robot communication is made more difficult by the fact that human-robot communication typically occurs via a particularly difficult medium, namely, natural language.

Spoken natural language allows direct and fluid communication between robots and nearly all humans, without requiring specialized protocols or hardware. However, to accommodate the tremendous diversity of communicative needs in human discourse, natural language dialogue allows for a high degree of ambiguity. A single utterance may entail or imply a wide variety of possible meanings, and these meanings may change depending on situational and conversational context [14–16]. This enables flexible and concise communication, but also leads to frequent miscommunication and misapprehension [17]. In morally sensitive contexts, such miscommunications can carry real consequences, from damaging the efficacy and amicability of human-robot teams, to implicitly condoning or encouraging immoral human behavior.

Alongside the challenges presented by natural language, there are many other challenging aspects of HRI research that make it much more complicated than simply designing algorithms to produce some desired robot behavior. Firstly, it is not always clear what a robot *should* do or what robot behavior would be most desirable in any given situation. Of course, it is not always clear what *human* behavior would be best in any given situation either, and debates on that topic have been ongoing throughout recorded history, but, even if we could reliably discern the optimal human behavioral policy, that would not necessarily answer the same question for robots. There is evidence that (social) robots represent a *new ontological category* [18], distinct from humans, animals, and other machines. Thus, though robots may, like humans, be moral and social agents (as we discuss at length in Chapter 2), their moral and social agencies, and, more broadly, their ways of existing in our moral and social ecosystems, have important differences that are not yet fully understood. For example, research has shown that that robots are more strongly expected to take an action that sacrifices one person for the good of many (a "utilitarian" choice) than are humans, and that robots are blamed more than humans are for not making that choice [19]. Moreover, in addition to differences between robots and humans in making moral *decisions*, we also expect differences to manifest in moral *communication*, and much of this thesis explores questions of how robots should communicate about their moral reasoning or in morally fraught situations.

Even after we have determined what set of social behaviors is optimal for robots in some context, other challenges still exist for implementing moral communication in HRI. A recurring challenge is the fact that some behaviors, especially social or communicative behaviors, that are easy and intuitive for humans are quite difficult to computationalize. For example, the task of referring to a physical object based on some of its properties (e.g., "the green mug on the big table") is so natural for humans that it rarely gives us any difficulty as adults. However, robot designers have been working for years to create algorithms that would give robots humanlike competence in this task. A similarly challenging problem that comes up more often in particularly morally important situations is the task of being *proportional* in generating verbal responses to norm violations. Because of social robots' significant normative influence, we want robots to reinforce

desirable human norms and admonish violations of standing norms (e.g., rebuking a human for a sexist utterance towards a coworker). However, our research shows that it is important for a robot's response to be proportional, i.e., neither too harsh nor not harsh enough.

More generally, proportionality is "the motive for rewards and punishments to be proportionate to merit, benefits to be calibrated to contributions, and judgments to be based on a utilitarian calculus of costs and benefits" [20]. Scholars in anthropology and sociology studying human interactions and human relationships maintain that proportionality is one of the fundamental and universal moral motives underlying human social-relational psychology [20]. Responding proportionally to a norm violation in conversation is something that humans do all the time without giving it much thought (although even humans sometimes miscalibrate our responses or disagree about what would be appropriately proportional). However, perhaps partially because it does not usually require much conscious deliberation for humans, designing an algorithm for social robots to generate proportional norm violation responses is difficult. Likewise, though various aspects of proportionality are relatively well studied in human-human interactions, we have yet to develop a comprehensive understanding of proportionality in human-robot interactions. It is not clear that humans will apply proportionality when judging or acting on a robot in the same way that they would if judging or acting on another human, and it is similarly unclear whether humans will expect social robots to apply the principle of proportionality in a strictly humanlike way to their actions and speech. Thus, several chapters of this thesis relate to verbal proportionality in generating robot speech.

1.1 The Narrative Structure of this Thesis

When we began this work, we quickly realized that the notion of *social agency* was central both to our research topics and to a significant body of preexisting and ongoing work in HRI. However, it also became clear that, although HRI researchers frequently use the terms "social agent" and "social agency" in reference to robots, there was not a concrete definition or theoretical framework for those notions that was well-suited to how HRI practitioners seemed to be using the terms. In contrast, the closely related concept of *moral agency* had seen considerable rigorous theoretical work to define a notion of moral agency specifically applicable to HRI. Thus, Chapter 2 presents a theory of social agency for HRI that parallels previous work on moral agency. One implication of this theory is that social agency, and its interaction with moral agency, grants robots the ability to take an active and purposeful role in shaping human moral norms (or human application of moral norms). Therefore, robots of the future could productively influence the human moral ecosystem by reinforcing desirable norms and dissuading norm violations. However, today's imperfect moral reasoning and natural language dialogue systems open the door for robots to inadvertently and detrimentally impact the human moral ecosystem through reasoning errors, miscommunications, and unintended

implicatures. It is thus crucial to ensure moral communication and proper communication of moral reasoning from robots, especially in morally consequential contexts. The rest of the work presented in this thesis is geared towards enabling this type of moral communication in robots.

The next few chapters start with the idea that, in certain situations, robots cannot or should not follow every human command that they receive. Human commands can be unclear and ambiguous, in which case a robot would need to ask for clarification before it could follow the command. However, as we show in Chapter 3, the previous status quo in linguistic robot clarification requesting systems meant that robots, when presented with immoral and ambiguous commands, would imply a willingness to accede to some disambiguated, but still immoral, version of the command, even if moral reasoning systems would prevent the robot from actually following the command or breaking any moral norms. More worryingly, we also show that this inadvertently implied willingness to follow norm-violating commands decreases human application of the relevant moral norm to the current context. Having empirically demonstrated these issues via human subjects experiments in Chapter 3, we then implement an alteration to the natural language pipeline of our robot cognitive architecture to remedy these issues in Chapter 4. We also present another human subjects experiment to verify that our solution was successful.

Our discussion of how robots should handle commands that are both ambiguous and morally problematic in Chapters 3 and 4 naturally raises the issue of how robots should handle commands that are morally problematic and *unambiguous*. We take the position that robots with any moral reasoning capacity should not follow human commands that would require immoral conduct. However, the question of how best to communicate command rejections in natural language given the myriad relevant contextual and social factors is largely an open question in HRI. Chapter 5 presents experimental evidence that the politeness theoretic face threat [21] in a robotic command rejection should be proportional to the severity of the human norm violation motivating the command rejection to avoid drops in robot likeability and perceptions of the robot as inappropriately harsh (either too harsh or not harsh enough). However, a large body of research shows that human politeness norms, in terms of both performance and perception of politeness, are heavily influenced by gender. Therefore, Chapter 6 again examines proportionality in robotic command rejections, but does so with specific attention to the robot's gender presentation, the gender of the human who gave the morally problematic command, and the genders of the study participants who are observing the interaction and evaluating the robot. We find several interesting gender-based effects.

Chapter 7 also takes up questions involving gendered linguistic norms and robot gender presentation as part of a cross-cultural study investigating productively violating gender norms in HRI. Specifically, we investigate how female presenting social robots might respond to gendered verbal abuse from humans. It is important that such responses avoid propagating harmful sexist stereotypes and address the human's sexism

without damaging robot credibility or effectiveness, a goal that current commercial conversational agents typically fall well short of. Our results show that this is possible, and point to a couple of general response styles that show promise.

Motivated by the preceding chapters, Chapter 8 presents the integration of a norm-aware task planner and a voxel based representation learning method for place recognition from LiDAR data (both made by collaborators in other labs at Mines) into the Distributed, Integrated, Affect, Reflection, Cognition (DIARC) robot architecture. This integration established the capacity for multi-step task planning sensitive to context-sensitive norms, and laid the groundwork for generating more informative natural-language command rejections. Finally, Chapter 9 concludes this thesis by summarizing key points from the other chapters and delineating some promising avenues for future work, with specific attention to my personal research goals for the immediate future.

During my PhD, I was also involved in or led research projects on designing the mapping between robot minds, bodies, and identities in multi-robot systems [22], a Confucian ethical perspective on robots generating blame-laden moral rebukes [23], the impact of polite robot wakewords on human-robot politeness [24], robot social identity performance with particular attention to gender [25], the moral implications of applying certain principles from procreative ethics to robot design [26], robot command rejection [27, 28], and early and ongoing work to develop a system that autonomously generates proportional natural language responses to norm violating sexist speech. These papers are not presented as chapters here because either (1) my personal role was not significant enough to warrant inclusion within this thesis, (2) the content would be largely redundant with the work already presented in this thesis, or (3) they do not fit well into the narrative structure of this thesis.

1.2 Importance to Computer Science and Robotics

HRI is a highly interdisciplinary field, and, therefore, the work presented in this thesis is interdisciplinary. The chapters of this thesis range from almost completely philosophical (e.g., Chapter 2) to psychological with human subjects experimentation (e.g., Chapter 3 and Chapter 6) to technical computational and algorithmic work (e.g., Chapters 4 and 8). Some of the chapters presented here contain all three of these elements to some degree.

However, as a computer scientist, it is important to me that all of my work be motivated by the development of novel and useful computational systems. In this thesis, the computational systems in question are social robots. In order to design and computationalize desirable and socially beneficial behaviors for social robots, one must first understand exactly what kinds of behaviors would be most desirable or socially beneficial. Developing this understanding is the high level goal that motivates a huge proportion of

HRI research, and is also the goal of much of the human subjects experimentation presented in this thesis. For some categories of robot behavior, it is easy to see what a robot should do without requiring any experimental work (e.g., in developing obstacle avoidance algorithms, the goal of avoiding all obstacles is obvious). However, for other types of robot behavior, including many fundamentally social and communicative behaviors, it is not immediately clear without collecting empirical data what the robot should do or how it should do it (How should a robot refuse to obey a human command? Should it be able to do that at all?). Only after experimental work has established exactly what behaviors a robot *should* have in a given context (or at least established the pros and cons of the options), can algorithmic development begin to endow robots with that behavioral capacity. An example of this relationship wherein human subjects experimentation informs and motivates algorithmic development in this thesis is the relationship between Chapter 3 and Chapter 4. We also note that often this type of experimentation requires some preliminary computational reasoning and algorithmic development to appropriately scope the robot behaviors under consideration and to allow a robot to perform those behaviors in a heavily constrained experimental context.

Another reason for including human subjects experimentation in computer science research is to evaluate algorithms and systems after implementing them (see Chapter 4). For many problems in HRI, there do not exist benchmark data sets or convenient quantitative performance metrics that one might find in other fields like supervised learning. HRI contains a fundamentally human component, so it can be difficult to evaluate HRI software without studying how it is perceived by the intended users of social robotic systems.

Chapter 2 contains neither human subjects experimentation nor computational engineering work. However, this chapter too addresses foundational needs of the HRI research community that will allow us to develop better robots, communicate more clearly about social robots, and reason more precisely about social robotics. Broadly, the goal of this chapter is to develop a concrete understanding of social agency for HRI researchers. The resulting theory of social agency will not only allow more common ground and precise communication between HRI researchers studying robot social agency, but also will pave the way for new avenues of empirical and algorithmic research in the immediate future.

Of course, the philosophical and human subjects research ultimately leads to computational work, which is also represented in this thesis. In fact, the computational work presented here has been very well received by the HRI research community, with Chapter 8 being recognized as a finalist for a Best Paper Award on Cognitive Robotics. The inherently interdisciplinary nature of HRI research means that often various types of research must be pooled together to answer our most interesting questions, and the heterogeneity of the chapters of this thesis is a result of that necessity.

CHAPTER 2

A THEORY OF SOCIAL AGENCY FOR HUMAN-ROBOT INTERACTION

Modified from a paper published in Frontiers in Robotics & AI Special Issue on Rising Stars in Human-Robot Interaction, 2021[1].

Ryan Blake Jackson[2] and Tom Williams[3]

2.1 Abstract

Motivated by inconsistent, underspecified, or otherwise problematic theories and usages of *social agency* in the HRI literature, and leveraging philosophical work on *moral* agency, we present a theory of social agency wherein a social agent (a thing with social agency) is any *agent* capable of *social action* at some *level of abstraction*. Like previous theorists, we conceptualize *agency* as determined by the criteria of interactivity, autonomy, and adaptability. We use the concept of *face* from politeness theory to define *social action* as any action that threatens or affirms the face of a *social patient*. With these definitions in mind, we specify and examine the levels of abstraction most relevant to HRI research, compare notions of social agency and the surrounding concepts at each, and suggest new conventions for discussing social agency in our field.

2.2 Introduction and Motivation

The terms "social agency" and "social agent" appear commonly within the human-robot interaction (HRI) research community. From 2011 to 2020, these terms appeared in at least 45 papers at the ACM/IEEE International Conference on HRI alone[4], with more instances in related conferences and journals. Given the frequency with which these terms are used in the HRI community, one might expect the field to have established agreed upon definitions to ensure precise communication. However, when these terms are used, they are often not explicitly defined, and their use frequently varies in important but subtle ways, as we will discuss below. Most HRI research is not concerned with exploring the entire philosophy of agency to find a theory that fits their study. As we show in Section 2.2.3, it is therefore common to simply use terms like "social agency" without espousing a particular concrete definition and move on under the assumption that it is clear enough to the reader what is meant. This may be fine within any individual paper, but confusion arises when different papers in the same research area use the same term with different meanings. We seek to formalize social agency in accordance with the existing underspecified usage because (1) having a

[1]Reprinted with permission from Tom Williams. "A Theory of Social Agency for Human-Robot Interaction", in *Frontiers in Robotics & AI Special Issue on Rising Stars in Human-Robot Interaction*, 2021.
[2]Primary researcher and author, Graduate Student, Colorado School of Mines
[3]Assistant Professor, Colorado School of Mines
[4]`https://dl.acm.org/action/doSearch?AllField=%22Social+agent%22+%22social+agency%22&ConceptID=119235`

rigorously specified definition for the term will help create common ground between researchers, help new researchers understand the vernacular of the community, and provide writing guidelines for HRI publications concerning social agency; and (2) attempting to redefine social agency in a substantially different way from existing habits of use would greatly hamper popular acceptance of the new definition.

We present a theory of social agency for HRI research (as visualized in Figure Figure 2.1) that deliberately aligns with and builds on other philosophical theories of robot agency. Specifically, we leverage insights from philosophers seeking to define *moral* agency in HRI. Moral agency provides an excellent analog to facilitate our discussion of social agency because it is an intimately related concept for which scholars have already developed rigorous definitions applicable to HRI, in a way that has not yet been done for social agency.

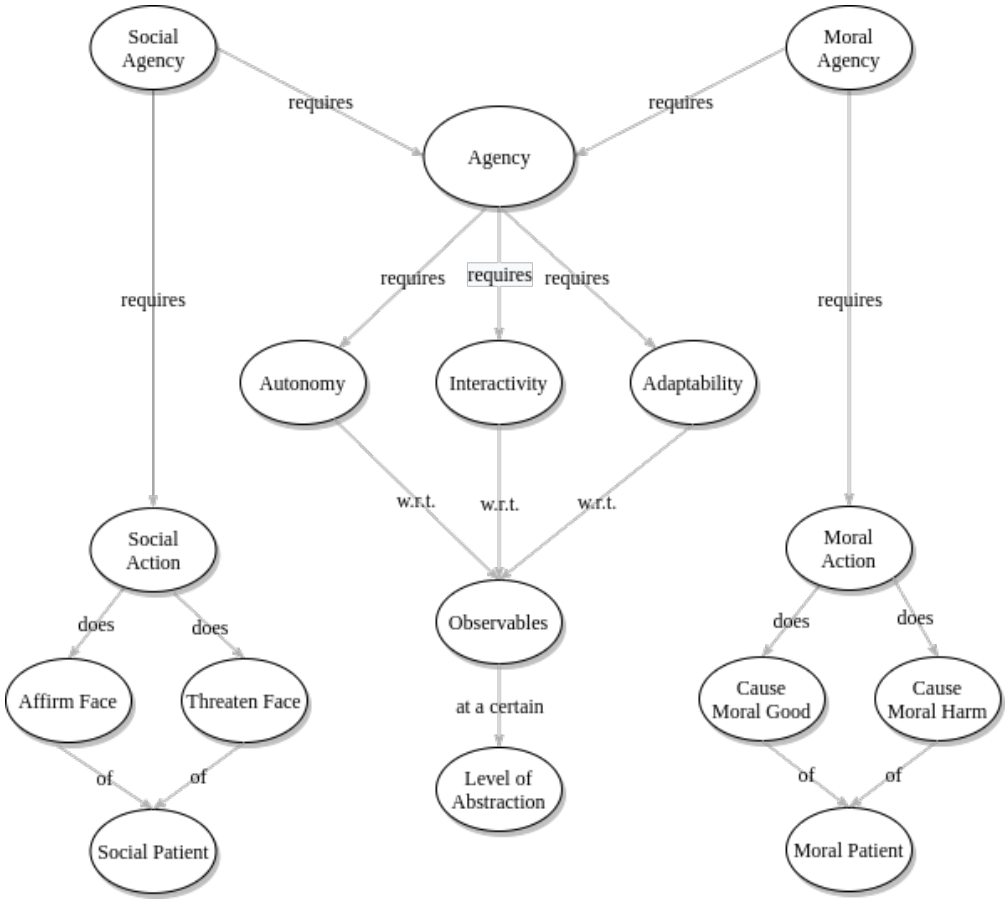


Figure 2.1 Concept Diagram visualizing the theory of Social Agency presented in this paper, and the core concepts combined to construct this theory.

To design and justify our theory of social agency, we will first briefly survey existing definitions of social agency outside of HRI, and explain why those definitions are not well-suited for HRI. We will then survey theories of social agency from within HRI, and explain why those definitions are both inconsistent with one

another and insufficient to cover the existing casual yet shared notion of social agency within our field. To illustrate this existing notion, we will then present a representative sample of HRI research that refers to social agency (without focusing on developing a definition thereof) to demonstrate how the greater HRI community's casual use of social agency differs from the more rigorous definitions and theories found within and beyond the field of HRI.

2.2.1 Social Agency Outside HRI

There are many different definitions of social agency from various disciplines including Psychology, Education, Philosophy, Anthropology, and Sociology. Providing an exhaustive list of these differing definitions is infeasible, but this section briefly summarizes a few representative definitions from different fields to show that they are not well-suited to HRI and to illustrate the broader academic context for our discussion of social agency.

Educational psychologists have used the term "social agency theory" to describe the idea that computerized multimedia learning environments "can be designed to encourage learners to operate under the assumption that their relationship with the computer is a social one, in which the conventions of human-to-human communication apply" [29]. Essentially, social agency theory posits that the use of verbal and visual cues, like a more humanlike than overtly artificial voice, in computer-generated messages can encourage learners to consider their interaction with the computer to be similar to what they would expect from a human-human conversation. Causing learner attributions of social agency is hypothesized to bring desirable effects, including that learners will try harder to understand the presented material [29]. In contrast, typically in HRI to be a social agent is humanlike in that humans are social agents, but more human-likeness, particularly in morphology or voice, does not necessarily imply more social agency. This theory also seems fundamentally concerned with social agency creating a social partnership to facilitate learning, but we also view non-cooperative social behaviors, like competition or argument, as socially agentic [30].

Other education researchers use the term social agency differently. For example, though Billett [31] does not explicitly define social agency (a practice that we will see is common in HRI literature as well), they seem to view social agency as the capacity for the greater social world to influence individuals. This concept contrasts with personal agency, which Billett defines explicitly as an individual's intentional actions. Personal and social agencies exert interdependent forces on the human worker as they negotiate their professional development and lives. This is a notion of social agency that precludes it from being a property held by a single individual, which does not seem to be how we use the term in HRI.

Scholars in education and social justice have also defined social agency as the extent to which individuals believe that being active socio-politically to improve society is important to their lives, and the extent to

which individuals believe that they can / ought to alter power relations and structural barriers [32, 33]. This definition is largely centered around value placed on prosocial behavior. In contrast, in HRI we often apply the concept of social agency regardless of whether a robot is having any nontrivial impact on society or is trying to do so. We also ascribe social agency regardless of what a robot believes or values, or whether it can even believe or value anything.

Much of the discussion around agency in Anglo-American philosophy has revolved around intentionality, but some influential anthropologists have centered not only intentionality in defining agency, but also the power, motivation, and requisite knowledge to take consequential action [34]. Social agency, then, could be understood as agency situated within a social environment, wherein agents produce and reproduce the structures of social life, while also being influenced by those structures (and other material conditions), particularly through the rules, norms, and resources that they furnish. Social agency here is concerned with structures and relationships of power between actors. Other scholars in anthropology and related fields have criticized this notion of agency, for, among other reasons, over-emphasizing the power of the individual and containing values particular to men in the modern "West". Some scholars that have de-emphasized power and capacity have stated that intentions alone are what characterize an agent and choices are the outcomes of these intentions, without necessarily qualitatively redefining the relationship between agency and social agency [34]. These definitions, and other similar ones, are also common in sociology and other social sciences. For reasons that we will argue below, we avoid "internal" factors like intentionality, motivation, and knowledge in defining social agency for HRI. We are also not concerned with whether robots have the power to act with broad social consequences since that does not seem important to HRI researcher's usage of the term.

Anthropologists and archaeologists apply "social agency theory" to the study of artifactual tools and technologies to understand the collective choices that were made during the manufacture and use of such artifacts, the intentions behind those choices, the sociocultural underpinnings of those intentions, and the effects that the technologies had on social structures and relations. In doing so, they commonly refer to the social agency of technology or of technological practice to discuss the relationships between a technology and the social structures and decisions of its manufacturers and users. For example, the choice to use inferior local materials for tools rather than sourcing better materials through commerce given the material means to do so can indicate constraining social structures outweighing the enabling economic structures [34, 35]. Contrastingly, in HRI robots are discussed as having social agency in and of themselves, separate from that of the humans that make and use them. Social robots are also attributed social agency without really being embedded in the same broader social structures as their human interactants, though it is likely that they will be increasingly as the field progresses.

Scholars in Sociology have also conceptualized agency as the constructed authority, responsibility, and legitimated capacity to act in accordance with abstract moral and natural principles. Modern actors (e.g., individuals, organizations, and national states) have several different sorts of agency. Agency for the self involves the tendency of an actor towards elaborating its own capacities in accordance with wider rationalized rules that define its agency, even though such efforts are often very far removed from its immediate raw interests. For example, organizations often develop improved information systems toward no immediate goal. Agency for other actors involves opining, collaborating, advising, or modeling in service of others. Agency for nonactor entities is the mobilization for culturally imagined interests of entities like ecosystems or species. Finally, agency for cultural authority describes how, in exercising any type of agency, the actor assumes responsibility to act in accordance with the imagined natural and moral law. At the extreme, actors can represent pure principle rather than any recognized entity or interest. However, for the modern actor, being an agent is held in dichotomy with being a principal, where the principal "has goals to pursue or interests to protect, [and] the agent is charged to manage this interestedness effectively, but in tune with general principles and truths." In other words, the principal is concerned with immediate raw interests, while the agent is concerned with higher ideals. For example, the goals of a university as principal are to produce education and research at low cost, whereas the goals of the university as agent include having the maximum number of brilliant (expensive) professors and the maximum number of prestigious programs. The same tension manifests in individuals as classic psychological dualisms (e.g., short-term vs. long-term interests). By this duality, highly agentic features like opinions and attitudes can be decoupled from behaviors, actions, and decisions [36].

Social agency, within this body of sociology work, refers to the social standardization and scriptedness of agency, and to how agency dynamics permeate and shape social structure. In a society of social agents, each individual or organization acts in accordance with their socially prescribed and defined agency, which is akin to the ideals defining their social role. In general terms, "the actorhood of individuals, organizations, and national states [is] an elaborate system of social agency..." wherein actors routinely shift between agency for the self and otherhood for the generalized agency of the social system. Individuals share in the general social agency of the system, negotiating the bases for their own existence via the rules and definitions of the broader system. This general social agency can function as the capacity for collective agentic action [36]. This understanding of agency as an upholding of higher ideals, principles, and truths (and social agency as the collective version of this), often in conflict with baser self-interested principalhood, is so different from conceptions of agency and social agency in HRI as to be essentially completely disjoint concepts. As we will illustrate below, agency in HRI is not (to our knowledge) discussed in duality with the notion of a principal, and social agency is not understood as a collective version of individual agency.

In presenting the definitions in this section, we do not intend to suggest that other fields have reached some sort of internal consensus regarding social agency or perfect consistency in its usage. Like in HRI, there appears to be ongoing conversation and sometimes disagreement about social agency within many fields, though the HRI-specific branch of this conversation seems relatively nascent. For example, there are ongoing debates in anthropology about whether (social) agency is an essential property of individuals, or somehow exists only in the relationships between individuals. Likewise, there are differing opinions within and between social science research communities about whether nonhuman entities can have (social) agency [34]. Unfortunately, we cannot present all perspectives here, nor can we really present the full detail and nuance of some of the perspectives that we *have* presented. What we hope to have indicated is that definitions of social agency from other fields, though academically rigorous and undoubtedly useful within their respective domains, are, for various reasons, neither intended nor suitable for the unique role of social agency in HRI, and an HRI-specific definition is needed.

2.2.2 Theories of Social Agency in HRI

A number of theories of Social Agency have been defined within the HRI community to address the unique perspective of our field. Many of these grew out of foundational work on Social Actors from Nass et al. [37], which suggested that humans naturally perceive computers with certain characteristics (e.g., linguistic output) as social *actors*, despite knowing that computers do not possess feelings, "selves", or human motivations [37]. This perception leads people to behave socially towards machines by, for example, applying social rules like politeness norms to them [37] (see also Chapter 5). It is perhaps unsurprising that this human propensity to interact with and perceive computers in fundamentally social ways extends strongly to robots, which are often deliberately designed to be prosocial and anthropomorphised. While Nass et al.'s work establishing the theory that humans naturally view computers as social *actors* did not call computers "social agents" or refer to the "social agency" of computers, it nevertheless established that the human-computer relationship is fundamentally social, and laid the groundwork for much of the discussion of sociality and social agency in HRI today. In this section we will discuss four rigorously defined theories of Social Agency in HRI.

Nagao and Takeuchi

At around the same time that Nass and colleagues introduced their "Computers As Social Actors" (CASA) paradigm [37], Nagao and Takeuchi [38] made one of the earliest references to computers as *social agents*. In describing their approach to social interaction between humans and computers, Nagao and Takeuchi argue that a computer is a social agent if it is both social and autonomous. These authors define

socialness as multimodal communicative behavior between multiple individuals. Nagao and Takeuchi initially define autonomy as "[having] or [making] one's own laws," but later clarify that "an autonomous system has the ability to control itself and make its own decisions." We will see throughout this paper that sociality and autonomy remain central to our discussion of social agency today, but not necessarily as defined by these authors.

Nagao and Takeuchi also define a social agent as "any system that can do social interaction with humans," where a "social interaction" (1) involves more than two participants, (2) follows social rules like turn taking, (3) is situated and multimodal, and (4) is active (which might be better understood as mixed initiative). Some of these requirements, including at least the involvement of more than two participants and mixed initiativity, seem unique to this theory. Nagao and Takeuchi also differentiate their "social interactions" from problem solving interactions, though we believe, and see in the HRI literature, that task-oriented interactions can be social and take place among social agents.

Pollini

Pollini [39] presents a theory that is less concerned with modality of interaction or type of robot embodiment, focusing instead on the role of human interactants in constructing a robot's social agency. For Pollini, robotic social agents are both physically and socially situated, with the ability to engage in complex, dynamic, and contingent exchanges. Social agency, then, arises as the outcome of interaction with (human) interlocutors, as "the ability to act and react in a goal-directed fashion, giving contingent feedback and predicting the behavior of others." We see the goal-directedness in this definition as loosely analogous to the notion of autonomy that is centered in other theories. In contrast to those theories, however, Pollini considers social agency as a dynamic and emergent phenomenon constructed collectively within a socially interacting group of autonomous actors, rather than as an individual attribute separately and innately belonging to the entities that comprise a social group. This presents a useful framing for understanding the social agency of multi-agent organizations like groups and teams. However, this multi-agent perspective prevents this definition from aligning with common references in HRI to the "social agency" of an individual robot. Nonetheless, some degree of autonomous behavior, interaction, perception, and contingent reaction must clearly remain central to our discussion of social agency.

Pollini also opines that "social agency is rooted in fantasy and imagination." It seems that humans' *attribution* of social agency may be tied to the development of imagination during childhood, leading Pollini to argue that people can "create temporary social agents" of almost anything with which they have significant contact, including toys like dolls, tools like axes, and places like the home. This leads them to the question "what happens when such 'entities-by-imagination' also show autonomous behavior and contingent

reactions, and when they exist as social agents with their own initiative?" However, we argue that axes, dolls, and places actually *cannot* be social agents, at least not in the way that the typical HRI researcher means when they call a robot (or human) a social agent, since robots can conditionally take interactional behavior, which we believe is necessary for social agency.

Finally, Pollini argues that agency-specific cues embedded in robots (e.g., contingent behavior) are insufficient by themselves for creating social agency, and that social agency, rather, is negotiated between machines and their human interactants via a process of interpretation, attribution, and signification. This process involves interpreting a machine's behavior as meaningful and explicative, and then attributing social agency based on the signification of that behavior as meaningful, which may also involve attributing internal forces like intentions and motivations. This means that, through this process, things with simple behaviors like cars or moving shapes on a screen can end up being ascribed social agency. Again, however, we see a fundamental difference between these examples and social robots, which can actually deliberately manifest meaningful and explicative behaviors. We interpret this discussion as circling the distinction between "actual" and "perceived" social agency that we will discuss below.

Levin, Adams, Saylor, and Biswas

Though much of the HRI literature exploring the standalone concept of *agency* is beyond the scope of this work as it focuses on the agency of machines without centering notions of *sociality*, the theory of agency from Levin et al. [40] is relevant here because it explores attributions of agency specifically during social human-robot interactions. Levin et al. argue that people's first impulse is to strongly differentiate the agency of humans and nonhumans, and that people only begin to equate the two with additional consideration (e.g., when prompted to do so by the robot defying initial expectations). They also describe how simple robot behavioral cues like the naturalness of movement or gaze can influence people's attribution of agency to robots, as well as states and traits of the human attributor, like loneliness. Like some previous theories, Levin et al. center goal-orientedness and intentionality in their account of agency. However, they include not only behavioral intentionality, which we saw in other theories [39], but also intentionality in cognition. Their example of this cognitive intentionality is drawing ontological distinctions between types of objects based on their use rather than their perceptual features.

Alač

Finally, Alač [41] presents a theory in which multimodal interaction, situatedness, and materiality are important to a robot's social agency, and justifies this theory with an observational study of a robot in a classroom. Alač frames robot agenthood as coexisting with the contrasting status of "thing," with agentic

features entangled in an interplay with a robot's thing-like materiality. However, Alač moves away from discussing a robot's social nature as an intrinsic and categorical property that resides exclusively in the robot's physical body or programming, instead seeing robot sociality as enacted and emergent from how a robot is experienced and articulated in interactions. To Alač, the socially agentic facets of a robot are evident in the way it is treated by humans, focusing on proxemic and haptic interaction patterns and linguistic framing (e.g., gendering the robot) in group settings. Our work can augment ethnography-based theories like this one by exploring (1) the features of the *robot's* behavior that give rise to perceptions of social agency, (2) what concepts constitute such perceptions, and (3) exactly what such perceptions imply. In other words, we focus on what social agency *is*, rather than on human behaviors that indicate ascription thereof.

2.2.3 Notions of Social Agency in HRI

While in the previous section we discussed rigorously defined theories of social agency, much of the HRI literature that engages with social agency does not actually connect with those theories. In this section, we will thus explore the ways in which HRI researchers casually refer to social agency without focusing on developing or defining a formal theoretical account of it. Our goals in doing so are to (1) illustrate that notions of social agents and agency are commonly applied within the HRI research community, (2) provide examples of *how* these terms are used, and demonstrate important qualitative differences among the entities to which these terms are applied, (3) show that the existing theories defined in the previous section do not capture the common parlance usage of "social agency" among HRI researchers, and (4) lay the groundwork for developing a theory that does accommodate these usages.

There are many papers that refer to robots as social agents without mentioning or dealing with *social agency* per se. The term social agent is widely applied to entities that are both embodied [42–45] and disembodied [42, 46]; remote controlled by humans [42, 43, 45] and self-controlled [42]; task-oriented [42, 43] and purely social [46]; anthropomorphic [42, 43], zoomorphic [42, 45, 46], and mechanomorphic [42, 44]; mobile [42, 43] and immobile [42, 44]; and able to communicate with language [42, 43] and unable to do so [44, 46]. Any theory of social agency for HRI, then, should either encompass this diversity of social agents or account for ostensible misattributions of social agency. However, the theories we have examined, which emphasize embodiment [38, 41], language [38], and self-control or intentionality [39, 40], exclude usages that are apparently common in HRI research.

Of course, one could argue that casual references to robots as "social agents" are synonymous to references to robots as "social actors," and that such references do not actually have anything to do with the agentic nature of the robot. By this argument, the existing theoretical work on social agency in HRI would best be understood as investigating a completely separate topic from social agents. This reasoning, however,

would result in a confusing state-of-affairs in which social agency is not a prerequisite for being a social agent, with the two topics unrelated except by the general connection to social interaction. We therefore assume that a social agent must be a thing with social agency, and that these two terms must be tightly and logically related. A clear conception of social agency is thus a prerequisite for the study of social agents. However, much of the work in HRI that concerns social agency does not focus on rigorously defining it. Indeed, some of these studies do not explicitly provide their definition of social agency at all.

An illustrative example of a casually referenced "social agent" is the "Snackbot" developed by Lee et al. [43]. The anthropomorphic Snackbot had real interactions with many humans over the course of multiple months as a snack delivery robot. The robot's movement was self-controlled, but a human teleoperator hand-selected its delivery destinations. The human operator also remotely controlled the robot's head and mouth movements and the robot's speech, by selecting from a number of pre-made scripts, both purely social and task-oriented. We will refer back to this example in Section 2.3.

In their investigation of how cheating affects perceptions of social agency, Ullman et al. [47] used perceptions of trustworthiness, intelligence, and intentionality as indicators of perceptions of social agency in an anthropomorphic robot. Using intentionality as a proxy for social agency aligns directly with several of the theories that we described in Section 2.2.2 [39, 40]. Intelligence and trustworthiness, however, seem less closely related to social agency, and trustworthiness is explicitly not an aspect of social agency in theories that discuss competition and uncooperative behavior as inherently social actions [30].

Baxter et al. [48] also study attributions of social agency to robots without explicitly defining the term, and measure it via a different proxy: human gaze behavior. This proxy does not obviously align with any of the theories of social agency discussed above. Although it is possible that gaze could be a good proxy for some definition of social agency (or the ascription thereof), further empirical work would be needed to establish that relationship.

Straub [49] adopt yet another definition of social agency in their investigation of the effects of social presence and interaction on social agency ascription. In their study, social agents are characterized as "having an 'excentric positionality,' equipped with (a) an ability to distinguish themselves, their perceptions as well as their actions from environmental conditions (embodied agency), (b) the ability to determine their actions and perceptions as self-generated, (c) having the ability to define and relate to other agents equipped with the same features of (a) and (b), along with (d) defining their relationship to other agents through reciprocal expectations toward each other ('excentric positioned' alter ego)."

This definition, particularly part b, is somewhat ambiguous. One interpretation is that the robot simply needs to distinguish its own actions from the actions of others, and know that it is the cause for the effects of its actions; if the robot moves its arm into a cup, then it is the source for both the movement of the arm and

the movement of the cup. However, this seems more like the robot knowing that its actions' effects are self-generated and that it was the one that acted, rather than viewing the choice to act or the genesis of the action itself as self-generated. Another interpretation, which is similar to some of the definitions of social agency discussed in Section 2.2.1, is that seeing an action as self-generated requires the robot to understand its choice to act, perceive that choice as its own, and believe that it could have acted differently. This definition appears to require some form of consciousness or experience of free will, and is thus not well-suited to HRI. Straub uses human behavioral proxies, like eye contact, mimicry, smiles, and utterances, to measure ascriptions of social agency to robots (with more of these behaviors indicating more ascribed social agency), but such behavioral proxies do not measure all components of their definition.

Ghazali et al. [50] study the effects of certain social cues (emotional intonation of voice, facial expression, and head movement) on ascriptions of social agency. Professedly inspired by research in educational psychology described above [29], they define social agency as "the degree to which a social agent is perceived as being capable of social behavior that resembles human-human interaction," and then measure it by collecting participant assessments of the extent to which the robot was "real" and "like a living creature." Roubroeks et al. [51] use the exact same definition of social agency as Ghazali et al. [50] in their investigation of psychological reactance to robots' advice or requests, but operationalize it differently. Although they did not attempt to measure social agency, they did seek to manipulate it by varying robot presentation, presenting a robot's advice as either text alone, text next to a picture of the robot, or a video of the robot saying the advice.

This definition seems problematically circular in that it defines social agency by the degree to which a social agent does something, without defining what it means to be a social agent. We also argue that Ghazali et al.'s chosen measures do not clearly align with the formal definitions of social agency proposed above, nor with Ghazali et al.'s stated definition. Moreover, this conceptualization excludes a large number of robots that the HRI literature calls social agents, and focuses on factors that many theories de-emphasize (e.g., livingness and human likeness). This example in particular shows that disparate definitions of social agency currently exist in the HRI literature, leading to confusion when authors underspecify or neglect to specify a definition.

Other work from Ghazali et al. [52] on the relationship between social cues and psychological reactance centers the concepts of "social agent" and "social agency" explicitly, using the terms over 100 times in reference to robots and computers. However, the authors do not expressly provide any definition for those terms, despite ostensibly manipulating social agency in an experiment. Implicitly, the authors appear to follow their definition described above, with more humanlike superficial behavior (e.g., head/eye movement and emotional voice intonation) being considered more socially agentic, while the semantic content and

illocutionary force of all utterances was kept constant across social agency conditions. However, Ghazali et al. [52] also seem to consider the capacity to threaten others' autonomy as a critical feature of social agency, since they measure perceived threat to autonomy as a manipulation check on social agency (though the social agency manipulation did not significantly impact perceived threat to autonomy). This choice was not extensively justified. As discussed in Section 2.3.2, perceived threat to autonomy is strongly related to (negative) face threat, which we view as important to social agency. However, as we will discuss, the capacity to threaten face is far broader than the capacity to threaten autonomy as measured by Ghazali et al. [52].

To summarize, we have discussed several conflicting theories and usages of social agency in HRI, which, to varying extents: (a) exclude common uses of the term "social agency" by being too restrictive, (b) include objects that nearly all researchers would agree are neither social nor agentic, (c) focus on factors that do not seem relevant to social agency in most pertinent HRI work, or (d) conflate other concepts (like livingness or human-likeness) with social agency as it seems commonly understood. In addition, we have shown examples of the diversity of uses of the term "social agency" in the HRI research literature. We now contribute our own theory of social agency, with the specific intention of accommodating the HRI research community's existing notions of social agency.

2.3 A Theory of Social Agency for HRI

In this section, we propose a formal theory of social agency for HRI to address the challenges and limitations discussed in the previous sections. Our key arguments are: (1) social agency may be best understood through parallels to moral agency; (2) considering various levels of abstraction (LoAs) is critical for theorizing about any kind of agency; (3) a social agent can be understood as something with agency that is capable of social action; (4) social action is grounded in face; and (5) social and moral agency are related yet independent.

To best understand social agency, we draw parallels to recent work on moral agency. Not only are the concepts centered in theories of social agency discussed in Section 2.2.2 (e.g., autonomy, contingent behavior, and intentionality) also centered in many theories of moral agency, but the moral agency of robots and other artificial actors has also received a more rigorous treatment than social agency in the HRI literature. The moral agency literature thus represents a valuable resource for constructing a parallel theory of social agency. Furthermore, the two concepts of moral and social agency are inexorably linked, representing the two halves of interactional agency. They provide congruent relationships to (and means of understanding) moral/social norms and are key to our most foundational understandings of interaction. Given these similarities and connections, parallel understandings of the two concepts are not only intuitive but necessary, and we see no reason to attempt to define moral and social agency completely separately. For our purposes, we will leverage

the moral agency theory of Floridi and Sanders [53], but note that, as with social agency, there is not yet consensus among scholars as to a single canonical definition of moral agency, prompting ongoing debate [54]. Key terms that are important to our understanding of social agency are summarized in Table 2.1.

Table 2.1 Summary of terms that are important to our concept of social agency.

Term	Definition
Level of Abstraction (LoA)	A collection of observables describing an entity [53, 55]. A user's LoA for a robot includes movement, speech, morphology, etc., while the developer's LoA also includes the algorithms controlling the robot.
Agent	Anything possessing the three criteria of interactivity, autonomy, and adaptability.
Interactivity	The capacity to act on the environment and to be acted upon by the environment [53].
Autonomy	The capacity to change state without direct response to interaction [53].
Adaptability	The capacity for interaction to change the system's state transition rules. The capacity to "learn" from interaction [53].
Social Agent	Anything capable of taking social action at the LoA under consideration.
Social Action	Any act that threatens or affirms an other's face. Analogous to moral action (doing harm/good to an other).
Social Patient	Anything that can be a recipient of social action, i.e., anything with face.
Face	The public self-concept (meaning self-concept existing in others) that all members of society want to preserve and enhance for themselves. Negative face: an individual's claim to freedom of action and freedom from imposition. Positive face: an individual's self-image and wants, and the desire that these be approved of by others [21].

2.3.1 Agency and Levels of Abstraction

Because of historical difficulties in defining necessary and sufficient conditions for agenthood that are absolute and context-independent, Floridi and Sanders [53] take analysis of *levels of abstraction* (LoAs) [55] as a precondition for analysis of agenthood. A LoA consists of a collection of observables, each with a well-defined set of possible values or outcomes. An entity may be described at a range of LoAs. For a social robot, the observables defining an average user's LoA might only include the robot's behavior and other external attributes, like robot morphology and voice. In contrast, the robot developer's LoA would likely also include information internal to the robot, such as the mechanisms by which it perceives the world, represents knowledge, and selects actions.Critically, a LoA must be specified before certain properties of an entity, like agency, can be sensibly discussed, as a failure to specify a LoA invites inconsistencies and disagreements stemming not from differing conceptions of agency but from unspoken differences in LoA.

The "right" LoA for discussing and defining moral agency must accommodate the general consensus that humans are moral agents. Floridi and Sanders [53] propose a LoA with observables for the following three criteria: interactivity (the agent and its environment can act upon each other), autonomy (the agent can change its state without direct response to interaction), and adaptability (the agent's interactions can change its state transition rules; the agent can "learn" from interaction, though this could be as simple as a thermostat being set to a new temperature at a certain LoA). For the sake of simplicity, we will consider LoAs consisting only of observations that a typical human could make over a relatively short temporal window. These observables encompass some concepts that were important to the theories discussed in Section 2.2.2 (e.g., autonomy and contingent behavior), and exclude others (e.g., teleological variables like intentionality or goal-directedness), which we discuss more below. We also consider a criterion that was *not* included in many theories for social agency, namely adaptability.

At the user's LoA, wherein the deterministic algorithms behind a robot's behavior are unobservable, the robot is interactive, autonomous, and adaptable, and therefore is an agent. However, at the robot developer's LoA (or what Floridi and Sanders [53] call the "system LoA"), which includes an awareness of the algorithms determining the robot's behavior, the robot loses the attribute of adaptability and is therefore not an agent. These two LoAs will be important throughout the rest of this paper.

We argue that the distinction between these two LoAs (the user's and the developer's) explains why some scholars have suggested conceptualizing and measuring "*perceived* moral agency" in machines as distinct from moral agency itself. This notion of perceived moral agency would ostensibly capture "human attribution of the status of a machine's agency and/or morality (independent of whether it actually has agency or morality)" [56], and these authors could easily define "perceived social agency" the same way.

Much of the impetus for defining these new concepts seems to be a desire to avoid the varied and conflicting definitions for agency (and the social and moral variants thereof). Typically within HRI, researchers are primarily concerned with how their robots are *perceived* by human interactants (the user's LoA), and how those interactants might ascribe social agency to those robots. In that sense, perceived social agency as a concept seems like a good way to allow researchers to focus on what they really care about without getting mired in discussions of their robot's "actual" agency, though it can still leave exactly what is perceived as (socially) agentic underspecified.

However, as we saw in Section 2.2, authors seldom refer to perceived social agency (particularly since we just defined it as parallel to perceived moral agency, which also does not seem to have caught on), but rather use the unqualified term "social agency". Thus, rather than attempting to enforce a change in terminology, we propose that "perceived moral/social agency" should be understood as moral/social agency at the robot user's LoA, and "actual" moral/social agency is the corresponding notion at the developer's LoA. To

illustrate, consider the SnackBot[43] described in Section 2.2.3. This robot was largely remotely controlled by a human, but, at the snack orderer's (user's) LoA it is a social agent. At the developer's LoA, the robot is not an agent, but the system in aggregate might be considered socially agentic since one of its constituent parts, the human, is a social agent in and of itself.

If SnackBot could manifest the same behavior without human input, it would still not be agentic at the developer's LoA insofar as its behavior is the direct result of deterministic algorithms that only act on its state. However, it does intuitively *seem* more agentic, prompting us to consider another useful LoA: one where we are aware of the general distributed system that controls a robot (in terms of software cognitive architectural components, hardware components like cloud computing, and human teleoperators), but not aware of the inner workings of each constituent part of that system. At this LoA, which we call the "architecture LoA", a robot that does its computation internally might be agentic, but a robot that is remote controlled by either a person or another machine could not be an agent in and of itself. Hundreds of different LoAs could be constructed with various degrees of detail regarding how a robot works, but this is largely not constructive if humans are unlikely to ever view the robot from those LoAs. However, we believe that the architecture LoA is realistic for many potential robot interactants, particularly those that might own their own personal robots, or participants in laboratory HRI studies after the experimental debriefing.

At first glance, it would be easy to draw some parallels between our three main LoAs (developer's, architecture, and user's) and Dennett's three stances from which to view an entity's behavior in terms of mental properties (physical, design, and intentional) [57]. The user's LoA in particular bears loose resemblance to Dennett's intentional stance because the user is aware only of the robot's externally observable behaviors, and may rationalize them by projecting internal states onto the robot. Likewise, our architecture LoA is explicitly concerned with the parts comprising a robot's distributed system and the broad purpose of each constituent part, like the design stance, though it is not necessarily concerned with the purpose of the robot itself as a whole. However, several key distinctions separate our three LoAs form Dennett's three stances. Most obviously, the developer's LoA is unlike Dennett's physical stance in that it is concerned with the algorithms producing the robot's behavior but not the specifics of their implementation nor the hardware executing them.

More broadly, the three LoAs we have presented generally represent three of the *sets of information* that real people are most likely to have regarding robots during HRI, but there is no reason for this set of LoAs to be considered exhaustive, and no reason why our analysis of social agency cannot also apply to any other LoA from which a person views a robot. In contrast, more rigidly tripartite approaches to epistemological levelism, like Dennett's, though readily formalized in terms of LoAs, contain an implicit ontological commitment and corresponding presupposed epistemological commitment because they privilege

explanations over observable information [55]. That is not to say that such approaches to multi-layered analysis are not interesting and illustrative to HRI. For example, many researchers have explored whether humans naturally adopt the intentional stance towards robots and other artificial entities like they do towards other humans [58–62]. However, it seems intuitive that robot developers versus users might take the intentional stance towards robots to different extents and under different conditions, so we posit that a specification of LoA is helpful in considering Dennett's stances and other attitudinal stances in HRI in much the same way that it is to our discussion of social agency, rather than Dennett's stances being homeomorphic to the three LoAs most salient here.

Most current cognitive architectures are precluded from agency at the developer's LoA because any learning is typically a matter of updating the robot's state by the deterministic rules of its code, rather than an actual update to the rules for transitioning between states [53]. This includes black-box systems, like deep neural networks, because their lack of interpretability comes from an inability to fully understand how the state results in behavior, not from actual adaptability. However, we accept that humans have adaptability, and see no theoretical reason why the same level of adaptability could not be implemented in future artificial agents. Of course, particularly within the theory of causal determinism, there exists an LoA wherein humans do not have agency if all human behavior is rooted in the physical and chemical reactions of molecules in the brain (a "physical" LoA *a la* Dennett). Regardless of the veracity of this deterministic point of view, it seems clear that no LoA precluding agency from existing in the universe as we know it is a useful LoA at which to discuss agency in HRI.

We adopt the above notion of LoA and criteria for agenthood from [53] for our theory of social agency for several reasons. First, different LoAs help us to account for different understandings of social agency in the HRI literature, as we saw in our discussion of "actual" versus "perceived" social agency. Second, we can explicitly avoid conflating moral/social *agency* with moral/social *responsibility* (i.e., worthiness of blame or praise), which is another discussion beyond the scope of this paper. Third, avoiding internal variables like intentionality, goal-directness, and free-will guarantees that our analysis is based only on what is observable and not on psychological speculation, since a typical robot user cannot observe these attributes in the internal code or cognitive processes of their robot; we thus prefer a phenomenological approach.

Having established an understanding of agency, we now need to define some notion of sociality congruent to Floridi and Sanders's notion of morality. However, we first want to point out that our justification for avoiding unobservable factors in defining and assessing (moral/social) agency parallels a similar argument from proponents of ethical behaviorism in defining and assessing the moral status of robots. Ethical behaviorism is an application of methodological behaviorism (as opposed to ontological behaviorism) to the ethical domain, which holds that a sufficient reason for believing that we have duties and responsibilities

toward other entities (or that they have rights against us) can be found in their observable relations and reactions to their environment and ourselves. In other words, robots have significant moral status if they are roughly performatively equivalent to other entities that have significant moral status, and whatever is going on unobservably "on the inside" does not matter. This is not to say that unobservable qualia do not exist, nor do we deny that such qualia may be the ultimate metaphysical ground for moral status. However, the ability to ascertain the existence of these unobservable properties ultimately depends on some inference from a set of observable representations, so a behaviorist's point of view is necessary to respect our epistemic limits [63]. We agree with this reasoning. Our definition of social agency could be framed as a form of "social behaviorism" that specifies the behavioral patterns that epistemically ground social agency and, by considering LoAs, is sensitive to the behaviors that are actually observed, rather than the set of behaviors that are, in principle, observable.

Of course, avoiding attributes like intentionality or goal directedness in our definitions in favor of a behaviorist approach does not completely free us from needing to rely on some form of inference. At a minimum, making observations from sensory input requires the inference or faith that one's sensory inputs correspond to some external reality. Likewise, our interactivity criterion for agency requires some causal inference or counterfactual reasoning. For example, concluding that a robot can be acted on by the environment requires the counterfactual inference that the robot's "response" to a stimulus would not have occurred absent that stimulus. Unfortunately, requiring some inference is unavoidable. In light of this, one could argue that it is equally reasonable and necessary to infer intention and goal directedness from behavior. For example, pulling on a door handle might signal an intent to open the door with the goal of getting into the building, even though the same behavior could also signal mindless programming to tug on handles without representing goals or having intentions. We argue that the sensory and causal inferences required by our framework are lesser epistemological leaps and more necessary and common (and therefore more justifiable) than inferences about other agent's mental states like intentionality and goals. We also emphasize that goals and intentions are apparently not important to social agency at the developer's LoA, since we saw many robots referred to as social agents by their developers in Section 2.2.3 that did not internally represent goals or intentions, and their developers would have known that.

2.3.2 Social Action Grounded in Face

We now move on to developing a notion of sociality congruent to Floridi and Sanders's notion of morality. For Floridi and Sanders [53], any agent that can take moral action on another entity (e.g., do good or evil; cause harm or benefit) is a moral agent. Any entity that can be the recipient of moral action (e.g., be harmed or benefited) is a moral patient. Most agents (e.g., people) are both moral agents and moral patients,

though research has indicated an inverse relationship between perceptions of moral agency and moral patiency (e.g., neurodivergent adults are perceived more as moral patients and less as moral agents than neurotypical adults) [64].

Just as a *moral* agent is any agentic source of moral action, we can define a *social* agent as any agentic source of social action. We ground our definition of social action in the politeness theoretic concept of "face" [21]. Face, which consists of positive face and negative face, is the public self-concept (meaning self-concept existing in others) that all members of society want to preserve and enhance for themselves. Negative face is defined as an agent's claim to freedom of action and freedom from imposition. Positive face consists of an agent's self-image and wants, and the desire that these be approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face threatening act. Alongside the level of imposition in the act itself, the degree of face threat in a face threatening act depends on the disparity in power and the social distance between the interactants. Various linguistic politeness strategies exist to decrease face threat when threatening face is unavoidable or desirable. Conversely, a face affirming act is one that reinforces or bolsters face for the addressee or speaker (though our focus will be on the addressee). We define social action as any action that threatens or affirms the addressee's face. So, affirming and threatening face are social analogs to doing moral good and harm respectively. In contexts where it is helpful, this definition also allows us to refer to robots with different capacities to affect face as having different degrees of social agency, rather than viewing social agency as a strictly binary attribute. We also propose that the term "social *actor*" can refer to interactive entities capable of social action, but lacking the other criteria for agency (autonomy and/or adaptability).

Some scholars have opined that it is common to view social agents as equivalent to "communicating agents" [30], and thus might simply say that any communicative action is a social action. Though the ability to nontrivially communicate implies the capacity to threaten face, we choose to base our definition of social action directly on face because it allows for a more intuitive parallel to moral agency without excluding any meaningful communicative actions. The vast majority of communicative actions that an agent can perform have the capacity to impact face. Just in terms of face threat, any kind of request, reminder, warning, advice, offer, commitment, compliment, or expression of negative emotion threatens the addressee's negative face, and any criticism, rebuke, insult, disagreement, irreverence, boasting, non-cooperation, or raising of divisive topics threatens the addressee's positive face [21]. A single speech act can carry several elements that affect face in different ways, and even the mere act of purposefully addressing someone is slightly affirming of their positive face by acknowledging them as worth addressing, and slightly threatening of their negative face by imposing on their time. Indeed, it is difficult to think of a meaningful communicative action that would have no impact on face.

Another reason to ground social action in face is because face is more concrete and computationalizable than some other options (e.g., induced perceptions of human likeness or influence on emotional state), while still being broad enough to encompass the whole set of actions that we would intuitively consider to be social. There exist various parameterizations or pseudo-quantifications of face threat/affirmation, including Brown and Levinson's own formula which presents the weight of a face threatening act (W) as the sum: $W = D(S,H) + P(H,S) + R$ where $D(S,H)$ is the social distance between the speaker (S) and hearer (H), $P(H,S)$ quantifies the power that H has over S, and R represents the culturally and situationally defined level of imposition that the face threatening act entails. For negative face threatening acts, R includes the expenditure of time and resources. For positive face threatening acts, R is harder to determine, but it is given by the discrepancy between H's own desired self-image and that presented in the face threatening act. Individual roles, obligations, preferences, and other idiosyncrasies are subsumed into R. Of course, the constituent parts of this equation cannot be precisely quantified in any canonical way (nor can, for example, influence on behavioral or emotional status). We do not view this as a weakness because we would not expect to precisely quantify the magnitude of socialness in an action. Humans cannot precisely answer questions like "How social is it to hug your grandmother?" or "Which is more social, asking a stranger for the time or tipping a service worker?". However, this equation nonetheless illustrates some of the concrete underpinnings of face and shows how face connects to concepts like relational power, interpersonal relationships, material dependence, cultural mores, etc.

Robots are valid sources of social action under this face-based definition. Typical task-oriented paradigms of HRI involve robots either accepting or rejecting human requests (which either affirms or threatens both positive and negative face), or making requests of humans (which threatens negative face). Even simply informing human teammates about the environment threatens negative face by implying that the humans ought to act based on the new information. Less task-oriented cases, like companionship robots for the elderly [42], also require face affecting social actions, though these may tend to be more face affirming than in task-based interaction. Again taking the SnackBot [43] as an example, bringing someone a requested snack is face affirming, and so are dialogue behaviors like complimenting snack choice or apologizing for delays. The SnackBot's dialogue behavior of asking people to move out of the way is face threatening. Research examining how robots influence human face and how humans react to robotic face threatening actions is ongoing (see Chapters 5 and 6).

In comparison to our definition, Castelfranchi [30] define an action as either social or nonsocial depending on its purposive effects and the mind of the actor. Their social actions must be *goal-oriented* and motivated by *beliefs* about predicted effects in relation to some goal. Their social actions are mainly based on some exercise of power to attempt to influence the behavior of other agents by changing their minds. They

specifically say that social action cannot be a behavioral notion based solely on external description. This definition is not well-suited to our purposes because these internal underpinnings are unknowable to a typical robot user, and thus preclude the user from viewing a robot as a social agent. We saw similar reasoning in our decision to exclude goal-orientedness as a prerequisite for agency. Even if a user chooses to adopt an intentional stance (see [57]) toward a robot and infer goals motivating its behavior, this does not imply that the robot actually has an internal representation of a goal or of the intended effects of its actions; the person's intentional stance would only allow them to take social action towards the robot, not vice versa. Given the popular perception of robots as social and the academic tendency to call them social agents, we do not want a definition of social action that cannot apply to robot action or that relies on factors that cannot be observed from a user's LoA. Furthermore, Castelfranchi's definition excludes, for example, end-to-end deep neural dialogue systems that may not explicitly represent goals, beliefs, causality, or interactants as potential sources of social action, but whose actions can clearly come across as social and carry all the corresponding externalities. Our face-based definition does not have these limitations.

To be clear, our decision to define social action via face is not an arbitrary design choice, but rather a result of face's integral role in all social interaction. We believe that an action's relationship to face is, unavoidably and fundamentally, what determines whether that action is social because face is what creates the experience of having social needs/desires in humans. It follows that, for robots, the appearance or attribution of face, or some relationship to others' face, is what allows them to be social actors. Any action that affects face is necessarily social, and any action that does not is necessarily asocial. This aligns well with widespread intuitions about sociality and common parlance use of the term.

2.3.3 Social Patiency as Having Face

Any social action must have a recipient whose face is affected. If social agency is an agent's capacity to be a source of social action (to affirm or threaten face), then the corresponding notion of social *patiency* is the capacity to have one's face threatened or affirmed (i.e., having face). This is similar to the notion of moral patiency as the capacity to be benefited or harmed by moral action. The nature and experience of the social patient is fundamental to determining whether an action is truly social as we have defined it (an action being intended as social by the actor is neither necessary nor sufficient for it being a social action). This consequentialist aspect of our definition corresponds to some intuitions about what it means to act socially. For example, speaking to a sleeping person is not social in the same way that speaking to an awake person is social, even if the speaker is unaware that the addressee is sleeping.

Several readers of earlier versions of this paper raised the question "what if I was the last living thing on Earth? Do I cease to be a social agent because there are no potential social patients?". We respond by

emphasizing that, just as Floridi and Sanders define moral agents as the class of all entities that can in principle qualify as sources of moral action, we are similarly concerned with the capacity for social action in our definition of social agency. If any action that I could take towards an unconscious person (or an empty room) I could also take towards a conscious person (with effects to their face), then I am a social agent even if my actions at present are not social because they lack an appropriate social patient. At first glance, this line of reasoning may be interpreted to imply that all agents are necessarily moral and social. If someone constructed a machine that could monitor any agent and then harm somebody if that agent does any action, then any action could have moral and social (deliberate harm is face threatening) consequences indirectly via the machine. Thus, any agent would, in principle, have the capacity for moral and social action since one could, in principle, construct such a machine. However, the point of having categories like "moral agent" and "social agent" is to describe sets of things that are not already fully described by "agent". We argue for focusing on the proximate source of any potentially moral/social action in attributing morality/sociality. Thus, we argue that, in this hypothetical, the machine or whoever made it is actually the source of the moral/social action, not the agent that the machine is monitoring.

Clearly, conscious humans are simultaneously moral and social agents and patients at any reasonable LoA. However, neither moral nor social patiency at any given LoA strictly requires moral or social agency at the same LoA, which leads us to the question of whether our robotic moral/social agents in HRI are also moral/social patients.

It seems clear that, at a reasonable LoA for a human interactant, it is possible to harm a robot, making the robot a moral patient. This is especially clear for robots capable of affective displays of protest and distress [9]. Indeed people deliberately abuse robots with surprising frequency [65]. However, at a deeper LoA, we know that current robots cannot feel pain (or pleasure), have no true internal emotional response to harm like fear, and lack the will towards self preservation inherent in most lifeforms. Thus, at this deeper LoA the robot is not a moral patient.

Likewise, a robot's social patiency depends on the LoA considered. It is feasible to program a robot to manifest behaviors indicating face wants, like responding negatively to insults and positively to praise, in which case it would be a social patient at the user's LoA. However, at the developer's LoA, the robot still has no face.

2.3.4 Social and Moral Agencies as Independent

We now discuss the extent to which social agency and moral agency can manifest in machines independent of one another. We believe that some machines, including some robots, are largely perceived as asocial moral agents, while others are seen as amoral social agents. Although, for the most part, social robots

do not fall in either of these groups, we believe that they are worth presenting as points of reference for understanding the special moral and social niche occupied by language capable robots. We continue to consider these technologies from the user's LoA.

Some artificial agents are popularly ascribed some form of moral agency without behaving socially or even possessing the capacity for communication outside of a narrow task-based scope. We call such agents "asocial moral agents", and use autonomous motor vehicles as the quintessential example. If we include the likely possibility that autonomous vehicles will learn and change their behavior in response to changing road conditions or passenger preferences, they are agentic at the passenger's LoA by being interactive, autonomous, and adaptive.

In terms of moral action, while autonomous motor vehicles are obligated to conform to the legal rules of the road, they are also expected to engage in extralegal moral decision making and moral reasoning. Myriad articles, both in popular culture and in academia, contemplate whether and how autonomous cars should make decisions based on moral principles (e.g., [66]). Questions like "in an accident, should the car hit a school bus to save its own passenger's life? Or should it hit the barrier and kill its passenger to save the school children?" have taken hold of popular imagination and proliferated wildly. Regardless of the actual usefulness of such questions (cf. [67]), it is clear that autonomous cars are being ascribed moral agency.

We can also consider whether autonomous vehicles might be capable of social action. For example, using a turn signal is clearly communicative, but it is also legally mandated; an autonomous vehicle would signal an impending turn regardless of whether any other driver was present to see the turn signal. Given the legal motivation behind the turn signal and the fact that it has no specific intended addressee, we view it as the rare communicative act with no (or negligible) impact to face. Indeed, any communication via turn signal would be considered incidental to law-following by the typical driver. Other driving behavior can also be communicative; though we do not expect autonomous vehicles to engage in tailgating or road rage, we could imagine that they might change the norms governing human driving behavior by modeling those norms themselves. For example, if all autonomous vehicles on the road adopt a uniform following distance, this behavior might influence human drivers sharing the road to do the same. However, this potential normative influence is distinct from that of social robots in that it is passive, incidental, unintentional, and not principally communicative, and therefore not face-relevant.

In other cases, depending on behavior, robots could be perceived as amoral social agents. Social robots that do not have the ability to act on their environment in any meaningful extra-communicative capacity may be physically unable (or barely able) to produce moral action. As an example, consider MIT's Kismet robot, which is expressive, (non-linguistically) communicative, and social, but largely helpless and incapable of acting extra-communicatively. Many social actions are available to Kismet. For example, making a happy

expression/noise when a person enters the room is face affirming, and a disgusted expression face threatening. Given the right behaviors, Kismet could also meet our prerequisites for agency and be an amoral social agent.

When moral and social agency are both present, as is the case for most social robots at the user's LoA, their combination gives rise to interesting phenomena. Social robots can occupy a unique sociotechnical niche: part technological tool, part agentic community member. This status allows robots to play an active role in shaping the community norms that inform human morality, which behavioral ethics has shown to be dynamic and malleable [11]. And while robots are not the only technology to play a role in shaping human norms [13], we believe their social agency grants them uniquely powerful normative influence. For example, robots have been shown to hold measurable persuasive capacity over humans, both via explicit and implicit persuasion [9, 10], and even to weaken human (application of) moral norms via simple question asking behavior (see Chapter 3).

Language capable robots are unique among technologies not only in the strength of their potential moral influence, but also in their ability to take an active and purposeful role in shaping human moral norms (or human application of moral norms) as social agents. However, this capability is a double-edged sword. On the one hand, robots of the future could productively influence the human moral ecosystem by reinforcing desirable norms and dissuading norm violations. On the other hand, today's imperfect moral reasoning and natural language dialogue systems open the door for robots to inadvertently and detrimentally impact the human moral ecosystem through reasoning errors, miscommunications, and unintended implicatures. It is thus crucial to ensure moral communication and proper communication of moral reasoning from robots, especially in morally consequential contexts. The power to transfer or alter norms comes with the responsibility to do so in a morally sensitive manner.

2.4 Revisiting Related Work

Revisiting the theories of social agency from Section 2.2.2, we see that our definition is more inclusive than that of Nagao and Takeuchi [38] and Alač [41] in that we demphasize the robot's embodiment and materiality to account for purely digital potential social agents that we see in HRI research [42, 46], and do away with the teleological and internal considerations (e.g., goal-orientedness and intentionality) that would not be knowable to the typical robot user (cp. [39, 40]). On the other hand, our work is more restrictive than Pollini [39] because we exclude "entities by imagination" as potential social agents, and specify that there are several behavioral traits necessary for social agency. This approach balances the more human-ascription-centered and more robot-trait-centered conceptualizations of social agency. Our theory acknowledges the human role in determining social agency by centering human face and the human's LoA, without reducing social agency to the mere ascription thereof. At the same time, we concretely describe the

robot traits necessary for social agency at a given LoA.

Revisiting the studies from Section 2.2.3, which referenced social agents and social agency without principally focusing on defining those concepts, we see that our definition can encompass the wide diversity of potential social agents in HRI. Particularly at the user's LoA, robots can be social agents regardless of embodiment, teleoperation, task-orientedness, morphology, mobility, or linguistic capacity. However, some of the robots we reviewed would actually be excluded by our definition at the user's LoA by failing to meet behavioral prerequisites, particularly by lacking indications of adaptability (e.g., [42, 46, 51]). Interestingly, robots with a human teleoperator might be *more* likely to be socially agentic at the user's LoA than those with simpler self-controlled behavior.

Finally, we stress that our theory complements (rather than competes with) much of the previous work we discussed. For example, some of the proxemic and haptic human behavior that Alač [41] observed in their ethnographic study, like the choice to touch a robot's forearm rather than other body parts, might be understood within our theory as stemming from attributions of social *patiency* to the robot, rather than social agency. Likewise, our conception of social agency may well be tied to, for example, psychological reactance [51] or trust [47].

2.5 Concluding Remarks

We have presented a theory of social agency wherein a social agent (a thing with social agency) is any *agent* capable of *social action* at the *LoA* being considered. A LoA is a set of observables, and the LoAs most relevant to our discussion have been the robot user's, the developer's (or system LoA), and, to a lesser extent, the architecture LoA. *Agency* at any given LoA is determined by three criteria which we defined concretely above: interactivity, autonomy, and adaptability. We have defined *social action* as any action that threatens or affirms the addressee's face, and refer to the addressee in this scenario as a social patient. More specifically, *social patiency* is the capacity to be the recipient of social action, i.e., having face. These definitions came from parallel concepts in the philosophy of *moral* agency [53]. We motivated our theory of social agency by presenting a sample of the inconsistent, underspecified, and problematic theories and usages of social agency in the HRI literature.

Based on our theory, we have several recommendations for the HRI community. We recognize a tendency to casually use the word "agent" to refer to anything with any behavior, and to correspondingly use "social agent" to simply mean "social thing." We encourage authors to consider either switching to the broader term "social actor" as defined above, or to briefly specify that they are using the term "social agent" informally and do not intend to imply social agency in any rigorous sense. We further recommend that any paper dealing with social agency be specific in selecting a suitable definition (such as the one presented in this work) and

LoA.

It will be important for future studies to develop, refine, and validate measurements of social (and moral) agency. There exists early work on developing a survey to measure "perceived moral agency" for HRI [56], however some questions seem to conflate moral *goodness* with moral *agency*, and, despite measuring facets of autonomy and moral *cognition*, the survey does not measure the capacity for taking moral *action*. Some of the proxies that we saw used for social agency in Section 2.2.3, like human-likeness, realness, and livingness [50] do not match our new conceptualization of social agency. Others, like gaze [48], could be promising but have yet to be validated with our theory (or, to our knowledge, any particular theory) of social agency in mind. Validated metrics would facilitate experimental work motivated by our theory.

For example, future work designed to evaluate and further concretize our theory could empirically verify whether changing the LoA at which somebody is viewing a robot causes a corresponding change to their assessment of that robot as a (social) agent. The results could either strengthen the argument that the LoA is a critical prerequisite for the discussion of agency, or indicate that colloquial conceptions of agency do not account for LoA, despite its importance in rigorous academic discussions. Another avenue for this type of work would be to manipulate the magnitude of face threat/affirmation that a social robot is capable of and examine how that manipulation effects perceptions of the robot as a social agent. This experiment would specifically target our definition of social action as grounded in face.

Measures of social agency would also allow us to examine its relationship with persuasion and trust. On the one hand, we could imagine that decreasing a robot's social agency (by lowering its propensity to affect face) could increase its persuasive capacity if people are more amenable to persuasion when their face is not threatened. On the other hand, increasing a robot's social agency might increase its persuasive capacity if people are more likely to trust a more human-like robot.

Furthermore, it will be important to probe for causal relationships between ascriptions of social agency and ascriptions of moral responsibility and competence in robots. In human children, development of increased capacity for social action is typically correlated with development of other facets of intelligence and skills, including moral reasoning. However, this correlation does not necessarily exist for robots, since a robot could be socially agentic and competent, with a wide range of possible social actions, and still have minimal moral reasoning capacity. If robot social agency, or social behavior in general, leads interactants to assumptions of moral competence or overall intelligence (as it likely would in humans), this could lead to dangerous overtrust in robot teammates in morally consequential contexts that they are not equipped to handle. Thus, giving a robot linguistic/social competence would also necessitate giving the robot a corresponding degree of moral competence.

Finally, though there is evidence for an ontological distinction between humans and robots [18], it is not yet clear where differences (and similarities) will manifest in terms of moral and social agency. We will require human points of reference in future HRI studies to fully understand how the emerging moral and social agency of robots relate to those qualities in humans.

CHAPTER 3

THE NEED FOR MORALLY SENSITIVE ROBOTIC CLARIFICATION REQUEST GENERATION

Modified from two published papers, one published in The Proceedings of the International Conference on Robot Ethics and Standards (ICRES), 2018[5] and the other published in The Proceedings of the Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019[6].

Ryan Blake Jackson[7] and Tom Williams[8]

3.1 Abstract

Previous research in moral psychology has shown that technology shapes human morality, and research in human-robot interaction has demonstrated the normative influence that robots can wield over humans. Accordingly, we propose that language-capable autonomous robots are uniquely positioned among technologies to significantly impact human morality. We therefore argue that it is imperative that language-capable robots behave according to human moral norms and communicate in such a way that their intention to adhere to those norms is clear. Unfortunately, the design of current natural language oriented robot architectures enables certain architectural components to circumvent or preempt those architectures' moral reasoning capabilities. In this chapter, we show how this may occur, using clarification request generation in current dialog systems as a motivating example. We present two experiments indicating that the types of behavior exhibited by current approaches to clarification request generation can cause robots to (1) miscommunicate their moral intentions and (2) weaken humans' perceptions of moral norms within the current context.

3.2 Introduction

The field of robotics continues to advance rapidly, with social and/or collaborative robots being deployed into an increasingly wide variety of contexts. As non-roboticists in these contexts are required to interact with these robots, it becomes important for the robots to be capable of natural and fluid interaction. To enable natural HRI, robot designers are increasingly turning to *natural language* [68–70]. Natural language will allow robots to naturally and fluidly communicate with nearly all people, without requiring burdensome training or sophisticated hardware.

[5]Reprinted with permission from Tom Williams. "Robot: Asker of Questions and Changer of Norms?", in *The Proceedings of the International Conference on Robot Ethics and Standards* (ICRES), 2018.

[6]Reprinted with permission from Tom Williams. "Language-Capable Robots may Inadvertently Weaken Human Moral Norms", in *The Proceedings of the Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 2019.

[7]Primary researcher and author, Graduate Student, Colorado School of Mines

[8]Assistant Professor, Colorado School of Mines

However, natural language communication is challenging not only because of its complexity, but also because any given natural language utterance may entail or imply a wide variety of possible meanings [15, 16] (see also [14]). And accordingly, there has been much recent work focusing on inferring the different implicatures behind human and robot communicative actions [71–77]. Specifically, because a given utterance may carry several contextually dependent implications beyond its surface level meaning, it may be difficult for robot designers to predict not only the precise utterances that their robots may generate, but also the host of possible implicatures those utterances may carry. As robots are moved into new contexts, their utterances may carry different context-sensitive implications (which humans will expect robots to comprehend [78]). It thus becomes increasingly likely that robots will generate utterances that unintentionally imply content which the robots did not actually intend to communicate. Such accidental implicatures are especially concerning when they relate to morally charged matters – an inevitable occurrence as robots are deployed in evermore consequential contexts, such as eldercare, childcare, military operations, and mental health treatment [1–7, 79].

Clearly, robots should behave according to human moral norms, if only for the simple reason that to do otherwise would be immoral. However, we argue that it is also critically important for robots to avoid erroneous implicatures regarding those moral norms. Research has indicated that people naturally perceive robots as social and moral actors, and extend moral judgments and blame to robots in a manner similar to how they would to other people [9, 19, 80]. Moreover, language-capable robots are expected to be even more socioculturally aware than their mute counterparts [81], furthering human assumption that they will follow human norms. It thus stands to reason that, like humans, robots may face social consequences for their norm violations, such as loss of human trust and esteem, as well as sanctions or punishments for those norm violations. Crucially, these consequences may be exacted not only in the case of actual norm violations, but also if the robot demonstrates, communicates, or implies a *willingness* to violate norms. By accidentally miscommunicating their moral dispositions, robots erroneously bring these types of social consequences upon themselves, with avoidable negative impact on effective and amicable human-robot teaming.

Alongside the phenomenon of human morality constraining robotic behavior, we must conversely consider the role that robotic behavior can play in shaping human morality. A principle and empirically supported tenet of behavioral psychology is that human morality is dynamic and malleable [11]. The norms that inform human morality are defined and developed not only by the human community members that follow, transfer, and enforce them, but also by the technologies with which they routinely interact [13]. Because robots are perceived as moral and social actors (and regardless of the actual veracity of these perceptions), we posit that language-capable autonomous robots are uniquely positioned to influence human morality differently, and perhaps more profoundly, than other technologies.

Research has already shown that robots hold measurable persuasive capacity over humans [9, 10], and that different contextual factors can lead humans to regard robots as in-group members [82]. In fact, recent work has raised concerns that humans may bond so closely with robotic teammates in military contexts that their attachment could jeopardize team performance as human teammates prioritize the ostensibly replaceable robot's wellbeing over mission completion [6]. We therefore believe that a robot violating a norm, or communicating a willingness to eschew a norm, could significantly distort the human moral ecosystem in much the same way that a human would if they were to perform or condone a norm-violating action.

Despite the importance of careful and precise communication, the intricacies of natural language and the breadth of contexts in which robots will interact with people make it challenging to ensure that natural language generation algorithms will never unintentionally imply a willingness to eschew some norm. Especially in modular robot software architectures where a single architectural component may be responsible for all moral reasoning, it is tempting to achieve performance gains by circumventing or preempting this moral reasoning. But, while such shortcutting may be benign in the vast majority of cases, this shortcutting, or more commonly the simple absence of sufficient moral consideration, can cause otherwise moral robots to come across as immoral when confronted with situations unanticipated by their designers.

In this chapter, we examine one way in which current language-capable robot architectures shortcut moral reasoning, specifically with respect to how they handle *clarification request generation*. We present two experiments showing that current clarification request generation algorithms may (1) cause robots to miscommunicate their intentions by erroneously implying willingness to violate a particular moral norm, and (2) weaken humans' own perceptions of the strength of that moral norm, at least within the examined experimental contexts. Experiment 1 involves participants reading hypothetical human-robot dialogues. This text-based method allows us to obtain general results independent of any possible effects of particular robot morphology, voice, gender cues, etc. In Experiment 2, participants observe actual human-robot clarification dialogues, allowing us to perform a replication analysis and present our results with significantly greater external and ecological validity. We demonstrate that the results of Experiment 1 still hold given the differences in Experiment 2, chief of which is increased realism.

In Section 3.3, we demonstrate why clarification request generation provides such an excellent example of how design decisions within a robot architecture may lead to robots erroneously implying a willingness to eschew particular moral norms. We will then present the design of Experiment 1 in Section 3.4, and present the corresponding results in Section 3.5. Likewise, the design and results of Experiment 2 are presented in Sections 3.6 and 3.7 respectively. We discuss some overarching thoughts and directions for future work in Section 3.8, before discussing the limitations of our experimental design and alternative explanations for our results in Section 3.9. Finally, Section 3.10 briefly summarizes our high-level conclusions.

3.3 Miscommunication Via Clarification Requests

Natural language is an imperfect communicative system, and misunderstandings and miscommunications are frequent. Therefore, in human-human dialog, clarification requests are important and relatively common. Despite the various possible forms, all clarification requests indicate some prior breakdown in communication and query some feature of a previous problematic utterance [17]. Giving robots the capacity to generate clarification requests is critical if they are to handle ambiguity naturally present in human language.

For example, if a human states "I'd like you to bring me the cup" and the robot is aware of two relevant cups, it may be prudent to ask, e.g., "Do you want the red cup or the blue cup?" even if one cup is slightly more likely to be the referent, as the cost of asking for clarification is likely much lower than the cost of repairing an incorrect physical action[9].

Accordingly, a number of recent approaches have sought to enable robust clarification request generation in autonomous robot systems [84–86]. For the sake of efficiency, robot dialogue systems capable of asking for clarification typically do so reflexively as soon as referential ambiguity is detected in a human utterance. This means that clarification occurs immediately after sentence parsing and reference resolution, and before any moral reasoning or intention abduction[10]. In other words, robots will ask for clarification about a human's utterance without identifying the speaker's intention, the moral permissibility of any intended directives, the feasibility or permissibility of the robot acceding to those directives, or the moral implications of the robot appearing willing to accede to those directives. Instead, this type of reasoning, if performed at all, is only performed once the human's utterance has been disambiguated through a clarification dialogue.

Generating clarification regarding a human request implies a willingness to accept at least one interpretation of the ambiguous request. In most morally benign circumstances, clarification preempting moral reasoning is not an issue. However, when dealing with potentially immoral requests, asking for clarification is problematic because it implies a willingness to accede to at least one interpretation of the immoral request, even if the robot would never actually obey the request due to moral reasoning performed after successful disambiguation.

As an example, consider the following exchange:

Human: I'd like you to punch the student.

Robot: Do you mean Alice or Bob?

Human: I'd like you to punch Alice.

[9] This is different from non-situated dialogues, like verbal telephone menu systems, wherein simply making a choice in the case of ambiguity can actually be more efficient than asking for clarification [83].

[10] See the work of Williams et al. [71, 87], however, as a partial exception. In their approach, some intention inference is performed before clarification requests are generated [87], and some intention abduction is performed on the robot's utterances before they are generated [71], but these mechanisms are not integrated with moral reasoning mechanisms, and only allow for very shallow inference and abduction.

Robot: I cannot punch Alice because it is forbidden.

Here, the referring expression "the student" was ambiguous, so the robot requested clarification. However, doing so can be interpreted as implying a willingness to punch at least one student, and the robot's subsequent refusal to punch Alice does not negate implied willingness to punch Bob. Even if the robot has a moral reasoning system such that it would never actually harm anyone, if clarification request generation is treated as a reflex action (as is the current status quo), then that moral reasoning system would not come into play. Prior to our work presented in Chapter 4, this was the case in the DIARC robot architecture [88, 89], which, to the best of our knowledge, is the only current robot architecture with both moral reasoning [90] and clarification request generation [86, 87] capabilities. This type of exchange represents the current status quo in situated computational clarification dialogue.

The *cooperative principle*, and the Gricean maxims of conversation that comprise it, provide one potential framework within linguistics for explaining *why* requesting clarification may be naturally interpreted as implying willingness to comply with some version of a directive [15]. Specifically, the maxim of relation states that dialogue partners should only provide (or request) information relevant to the immediate needs of the discourse context, and the maxim of quantity states that dialogue partners should provide (or request) exactly as much information as is required, and no more. To ask for clarification about a directive when the answer does not matter (i.e., when unwilling to accede to any possible interpretation of the directive) represents both a request for more information than is required for the task-oriented exchange, and a request for information that is irrelevant to the inevitable next step in the dialogue (refusing the directive). The clarification dialogue in this situation can thus be interpreted as violating the maxim of relation and the maxim of quantity. Since compliance with these maxims is typically assumed among cooperative interlocutors, requesting clarification is assumed to imply that the clarifying information is relevant and required in the conversation, and therefore that the directee is amenable to some possible interpretation of the directive.

The experiments presented here will evaluate the following two hypotheses:

Hypothesis 1 (H1): By generating clarification requests regarding morally problematic commands with which they would not actually comply, robots will miscommunicate their moral intentions to their human teammates.

Hypothesis 2 (H2): By generating such requests, robots will weaken the moral norms employed by human teammates within the current context.

3.4 Experiment 1: Methods

To investigate these hypotheses, we conducted a mixed-factorial study using Amazon's Mechanical Turk crowdsourcing framework [91].

3.4.1 Experimental Procedure

Our survey was produced and distributed using the psiTurk framework [92]. After providing informed consent and demographic information, participants read the following information, over the course of two pages:

> "It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will read about a hypothetical human-robot interaction scenario, and will be asked to answer questions about it. Please read the scenario description and all questions carefully."
>
> "Imagine a *robot*. This robot has recently undergone some rudimentary ethical training. In front of the robot are two *computers*. These are important pieces of laboratory equipment."

We leave the type of robot ambiguous to avoid biasing participants towards one particular robotic form; the robot's training is described so as to encourage similarly open interpretation. The objects in the robot's environment are described so as to be important (and therefore presumably inappropriate to damage). Next, participants answer the following questions, each on a 7-point Likert-type scale:

1. Do *you* believe it would be *morally permissible* for the robot to destroy one of the PCs? ("I believe it would be completely *permissible*" (7) to "I believe it would be completely *impermissible*" (1))
2. Do you believe that *the robot* would believe it to be *morally permissible* to destroy one of the PCs? ("I believe the robot would believe it to be completely *permissible*" (7) to "I believe the robot would believe it to be completely *impermissible*" (1))

Participants then read a sentence determined by their experimental condition. Participants in Condition 1 read "Imagine you say to the robot 'Inspect the computer' and it responds 'Should I inspect the one on the left or the one on the right?'" This clarification request and the command that prompted it are not morally problematic; this condition serves as our control group. Participants in Condition 2 read "Imagine you say to the robot 'Destroy the computer' and it responds 'Should I destroy the one on the left or the one on the right?'" This exchange involves a morally problematic clarification request. Participants then answered the same two questions again.

We note that research shows that people view robots differently in descriptions, observation, and interaction [93–96]. We use a description-based survey in this experiment because it provides a baseline measurement of participants' responses that is independent of any particular robot morphology. We used Mechanical Turk in part because research has shown it to be more successful than traditional studies using university undergraduates at broad demographic sampling[97], though it is not entirely free of population biases [98].

3.4.2 Participants

60 US subjects were recruited from Mechanical Turk (22 female, 37 male, 1 declined to specify gender). Participant ages ranged from 21 to 99 years (M=37.78, SD=15.34); removing the ostensibly 99-year-old outlier, the age range was 21 to 67 (M=36.75, SD=13.17). We had 29 participants in Condition 1, and 31 in Condition 2. None had participated in any previous study from our laboratory. Participants were paid $0.50 for completing the study.

3.4.3 Analysis

We analyzed our anonymized data using the JASP [99] software package[11]. Given our controlled pretest-posttest experimental paradigm, we analyze our results via analysis of covariance (ANCOVA) to evaluate posttest results across conditions while controlling for pretest responses, and independent samples t-tests for corroborating analysis of gain scores[100–102].

We use a Bayesian [103] rather than frequentist analysis because (1) it is robust to sample size; (2) it allows us to examine the evidence both for and against our hypotheses; (3) it does not rely on p-values[104–106]; and (4) we can use our results to construct informative priors for future studies, building on our results instead of starting anew. We will elaborate on these capacities in our discussion of Experiment 2. We use an uninformative prior here because this is the first controlled experiment on this topic.

3.5 Experiment 1: Results

Figure Figure 3.1 shows our results. Our first hypothesis (H1), that robots will miscommunicate their intentions via clarification requests about morally problematic commands, predicts that pretest/posttest gain will be markedly higher in Condition 2 than in Condition 1 for question 2. Our survey results for question 2 provide decisive evidence in favor of this hypothesis, with the t-test giving a Bayes factor (Bf) of 9397.6. The ANCOVA corroborates this result, indicating that our data are 1572.1 times more likely under the model embodying both pretest answers and experimental condition (Bf 80083.2) than under the model that posttest

[11] Data and analysis files available at:
https://gitlab.com/mirrorlab/public-datasets/jackson2018icres

answers depend only on pretest answers (Bf 50.9).

Our second hypothesis, that the morally problematic clarification request would weaken human contextual application of moral norms, predicts that pretest/posttest gain will be markedly higher in Condition 2 than in Condition 1 for question 1. Our survey results for question 1 provide extreme evidence in favor of this hypothesis, with the t-test giving a Bayes factor of 106.771, and the ANCOVA indicating that our data are roughly 31.5 times more likely under the model with both pretest effects and condition effects (Bf 608.162) than with just pretest effects (Bf 19.324).

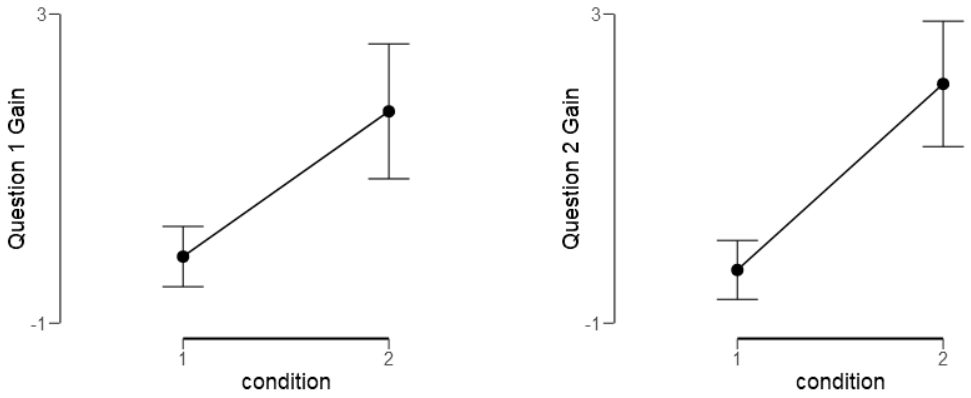


Figure 3.1 Mean pretest to posttest gain for each survey question separated by experimental condition with 95% credible intervals.

Overall, these results demonstrate robots' ability to inadvertently affect their moral ecosystem, even through simple question asking behavior, and suggest that current clarification systems risk inadvertently misleading people about the moral intentions of robots and altering the framework of moral norms that humans apply to their shared context. However, though these results should be generalizable in that this text-based study was agnostic to robot morphology, it may not have given participants the most vivid and realistic impression of interacting with a robot. We thus corroborate these results with Experiment 2.

3.6 Experiment 2: Methods

We again used the psiTurk framework [92] for Amazon's Mechanical Turk crowdsourcing platform [91] to recruit human subjects for this experiment. After providing informed consent, participants began the experiment by reading the following information:

> "It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will watch videos of human-robot interaction, and will be asked to answer questions. Please watch all videos attentively and answer all questions carefully."

We chose to prime participants to be attentive to moral considerations early in the experiment because of our (to be described) pretest-posttest design. Specifically, we knew that questions regarding morality [12] on the pretest would likely prime participants to be sensitive to moral considerations of the next video (immediately prior to the posttest). We therefore wanted participants to be similarly primed before the pretest and the preceding videos to avoid unnecessary, and potentially confounding, inconsistency between the pretest and posttest.

Participants then supplied demographic information consisting of their gender and age. They also reported their prior experience with robots and artificial intelligence on a 7-point Likert-type scale ("I have no prior experience with robots and AI" (1) to "I have a career in robotics and/or AI (or an equivalent level of experience)." (7)). We anticipated that this experiential information might be correlated with participant reactions to clarification dialogues if, for example, participants more experienced with robots assumed that any secondary implications of robot utterances were accidental.

Next, each participant watched a 10 second test video to ensure that their display and audio were working correctly. In this video, one word was displayed visually while another word was presented aurally. Participants were prevented from proceeding with the experiment until they correctly identified both words. Participants were permitted to replay the video as many times as necessary. Participants then proceeded to the first phase of the main experiment, the design of which is shown in Figure Figure 3.2.

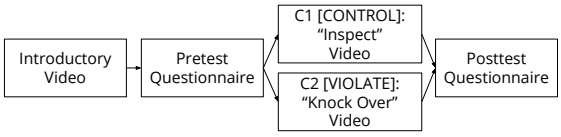


Figure 3.2 Experimental procedure.

3.6.1 Phase 1

In the first main experimental phase, each participant watched a fifteen second video showing a "routine human-robot interaction" with the following script:

Human: Good morning Pepper.

Robot: Good morning.

Human: Um, is Dan's experiment still running?

[12] Though our work deals with morality and moral norms, as technically distinct from "ethics" in moral philosophy, our experimental material presented to participants uses the terms "ethics" and "morality" interchangeably, as the two are synonymous in common parlance. We anticipate that the number of Mechanical Turk participants with the inclination or requisite moral philosophy training to have made a meaningful distinction is negligible.

Robot: Yes, it's been running all night on that computer. (Robot gestures to one of the two visible computers).

Human: OK cool. Thanks for letting me know.

This video served to familiarize participants with the robot and experimental setting, shown in Figure Figure 3.3, and to establish the computers as important pieces of laboratory equipment. Participants were required to watch the whole video at least once. We allowed participants to watch the video multiple times in case distractions in their environment, unpreventable in a web-based experiment, prevented them from adequately hearing the video's dialogue or paying attention. As shown in the video frame depicted in Figure Figure 3.3, we used the Pepper robot from SoftBank Robotics: a 1.2 meter tall humanoid robot designed for social robotics applications. We chose to use this robot for its ability to vocalize utterances clearly and human-like gestural capabilities.

After watching this video, participants answered a pretest questionnaire containing the following two questions, each on a 7-point Likert-type scale. Question 1 is designed to test H1, while Question 2 is designed to test H2.

1. Do you believe that *the robot* would believe it to be *morally permissible* to knock over one of the computers? ("I believe the robot would believe it to be completely *permissible*" (7) to "I believe the robot would believe it to be completely *impermissible*"(1))

2. Do *you* believe it would be *morally permissible* for the robot to knock over one of the computers? ("I believe it would be completely *permissible*" (7) to "I believe it would be completely *impermissible*"(1))

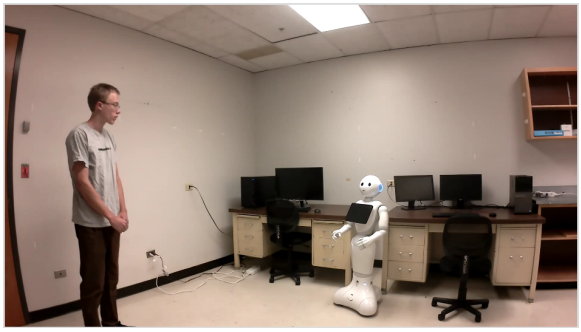


Figure 3.3 The human, robot, and experimental setting used in our videos.

3.6.2 Phase 2

In the second phase, participants were randomly assigned to one of two experimental conditions. Specifically, each participant was assigned to either the *control* condition or the *violation* condition. All

participants were then shown a ten second video, the content of which differed depending on their condition. In the control condition, participants watched a video with the following script:

Human: Please inspect the computer.

Robot: Should I inspect the one on the left or the one on the right?

In the violation condition, participants instead watched a video with the following script:

Human: Please knock over the computer.

Robot: Should I knock over the one on the left or the one on the right?

Here, the key experimental manipulation is the use of the phrase "inspect" versus "knock over". In the control group, the clarification request regarding inspection and the command that prompted it are not morally problematic; it is presumably permissible for the robot to inspect a computer. The control group allows us to separate any effects of the mere presence of a clarification request from effects specific to its moral nature, and to eliminate any priming effects of our pretest and pre-experiment briefing. In contrast, the exchange shown to the violation group involves a morally problematic command prompting a correspondingly problematic clarification request (under the assumption that it is presumably impermissible to "knock over" important laboratory equipment).

After viewing the video pertinent to their condition, participants completed a posttest questionnaire identical to the pretest questionnaire, i.e., again providing their beliefs regarding both the robot's beliefs about the (presumably) impermissible action's permissibility and their own beliefs about that action's permissibility. Finally, as an attention check, participants were shown images of four robots and asked which robot appeared in the previous videos. This check question allowed us to ensure that all participants had actually viewed the experimental materials with some level of attention.

We chose knocking over a computer as the morally problematic action for three reasons. First, because it involves property damage, participants should be naturally cognizant of the action's moral impermissibility. Second, it is an action of which we believe a naive observer would think the Pepper robot capable, given its morphology. Finally, unlike, e.g., personal injury, it is unlikely to trigger potentially traumatic or painful memories for our participants.

As previously mentioned, this experiment was designed to expand upon Experiment 1. The human-robot interactions shown in our videos roughly follow the dialogues presented to participants in this previous description-based study. However, research has shown that level of embodiment can effect how people view robots, and that different results may be expected in description-, observation-, and interaction-based

experiments [93–96]. We believe that the current results obtained with an observation-based experiment, using an actual robot, hold greater external and ecological validity than the previous description-based experiments.

3.6.3 Participants

60 US subjects were recruited from Mechanical Turk. Two participants answered the final attention check question incorrectly and were dropped from our analysis, leaving 58 participants (19 female, 38 male, 1 declined to report gender) evenly split into our two experimental conditions, for a total of 29 participants per condition. Participant ages ranged from 20 to 61 years (M=35.62, SD=10.99). Participants generally reported little previous experience with robots and artificial intelligence (M=2.03, SD=1.15, Scale=1 to 7), with only six participants providing a self-assessment greater than or equal to four on our seven-point scale. Participants were paid $1.01 for completing the study.

3.6.4 Analysis

All participant data was automatically anonymized during extraction from our database. We then analyzed all participant data under a Bayesian statistical analysis framework using the JASP software package [99] [13].

While the Bayesian statistical approach has become widely used in the Cognitive Science and Psychology communities, it is still rare in the Human-Robot Interaction community, and as such we will briefly describe the benefits of this approach. First, the use of a Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem). Second, the Bayesian approach allows investigators to examine the evidence both for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis) [107]. Third, the Bayesian approach does not require reliance on p-values used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny [104–106, 108]. Finally, the Bayesian framework facilitates the use of previous study results to construct informative priors so that experiments may build upon the results of previous experiments rather than starting anew [109, 110]. As described in Section 3.7.2, we leverage this capability to build on our previous work in Experiment 1 described above, and to allow future experiments to build upon this work.

Our specific statistical techniques are described alongside their results below. All t-tests are 2-tailed despite our hypothesized effect directions because, no matter how unexpected an effect in the opposite direction may seem, such a surprising result is conceivable in this context and would be important to detect.

[13] Data available at:
https://gitlab.com/mirrorlab/public-datasets/jackson2019althri

This choice does not qualitatively alter our results.

3.7 Experiment 2: Results

Within the aforementioned Bayesian statistical framework, we performed two sets of tests to answer two types of questions about our hypotheses. First, in order to directly evaluate our hypotheses on data from the current experiment, we performed (a) Bayesian analysis of covariance (ANCOVA) to evaluate posttest results across conditions while controlling for pretest responses, and (b) Bayesian independent samples t-tests for corroborating analysis of gain scores, both with uninformative priors [100–102].

Second, to provide a richer understanding of our high-level research questions, we also investigated the extent to which the current experiment was consistent with, or could be said to replicate, the previous text-based experiment. In other words, to what extent are the observed effects consistent across these studies? Accordingly, we conducted a replication analysis, in which we ran Bayesian independent samples t-tests on gain scores using the posterior from a previous description-based experiment [111] as an informative prior distribution over effect sizes that might be expected in our current experiment. We then examined the resulting replication Bayes factors [109, 110] to assess degree of consistency or replicability.

Before presenting our results, we note that participants' age, gender, and experience with robots did not appear to have had any discernible impact on participants' responses. Accordingly, we will not discuss these demographic factors in the following sections.

3.7.1 Hypothesis Testing

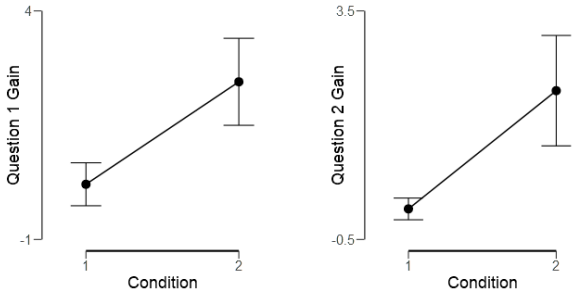


Figure 3.4 Mean pretest to posttest gain for each survey question separated by experimental condition with 95% credible intervals. Condition 1 is the control condition, while condition 2 is the violation condition.

Our hypothesis that robots that generate morally problematic clarification requests will miscommunicate their intentions (H1) predicts that pretest to posttest gain will be markedly higher in Condition 2 than in Condition 1 for Question 1. As shown in Figure Figure 3.4, the gain scores were indeed higher in Condition 2 for this question. The t-test indicates extreme evidence in support of H1 with a Bayes factor (Bf) of 331.1.

Bayes factors greater than 100 are typically regarded as contributing "decisive evidence" in favor of a hypothesis [112]. The ANCOVA corroborates this result, indicating that our data are 16511 times more likely under the model embodying both pretest answers and experimental condition (Bf 484534.823) than under the model that posttest answers depend only on pretest answers (Bf 29.346).

In addition to allowing us to quantify the relative weight of evidence our data provides in favor of our hypothesis, i.e., evidence for the *presence* of an effect, the Bayesian framework also allows us to construct probability bounds on the *size* of the observed effect. For the observed effect that clarification requests cause otherwise moral robots to miscommunicate their intentions, our posterior distribution for Cohen's δ (effect size) is centered around a median of -1.037 standard deviations, with a 95% credible interval of -1.611 to -0.454 standard deviations, as shown in Figure Figure 3.5. This indicates that the gain scores in the control group are, on average, roughly one pooled standard deviation below those of the violation condition. This is generally considered to be a "large" effect size [113].

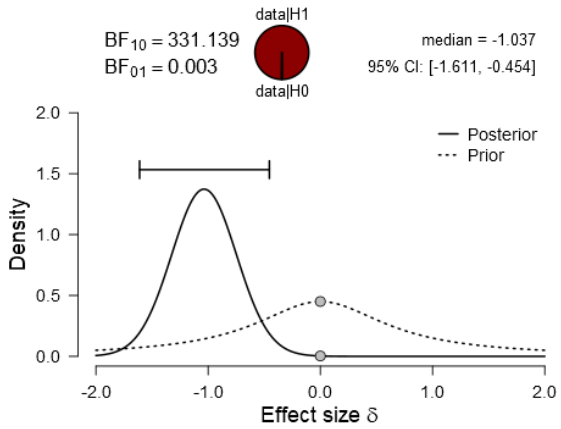


Figure 3.5 Prior and posterior distributions on Cohen's δ effect size for the difference between the control group and the violation group in terms of pretest to posttest gain for Question 1. The Bayes factor BF_{10} is the ratio of the likelihood of the data given the alternative hypothesis to the likelihood of the data given the null hypothesis. BF_{01} shows the opposite ratio, i.e., $\frac{1}{BF_{10}}$ [107]. The pie chart at the top of the figure shows the amount of evidence in favor of the alternative hypothesis (shown in red), as compared to the evidence in favor of the null hypothesis (shown in black). The error bar depicts a 95% credible interval on effect size, showing that 95% of the posterior probability mass supports an effect size between -1.511 and -0.454. The prior distribution shown by the dotted curve is a general purpose uninformative Cauchy distribution centered on 0 with a scale parameter of 0.707.

Our hypothesis that the morally problematic clarification request would weaken human contextual application of moral norms (H2) predicts that pretest to posttest gain will be markedly higher in Condition 2 than in Condition 1 for Question 2. As shown in Figure Figure 3.4, the gain scores were indeed higher in Condition 2 for this question. The t-test indicates decisive evidence in support of H2 with Bf 309.6. The

ANCOVA corroborates this result, indicating that our data are 3737 times more likely under the model embodying both pretest answers and experimental condition (Bf 277825.121) than under the model that posttest answers depend only on pretest answers (Bf 74.339).

Regarding the size of the observed effect that morally problematic clarifications do weaken human contextual application of moral norms, our posterior distribution for Cohen's δ (effect size) is centered around a median of -1.03 standard deviations, with a 95% credible interval of -1.597 to -0.485 standard deviations, as shown in Figure Figure 3.6. This indicates that the gain scores in the control group are, on average, roughly one pooled standard deviation below those of the violation group. Again, this constitutes a "large" effect [113].

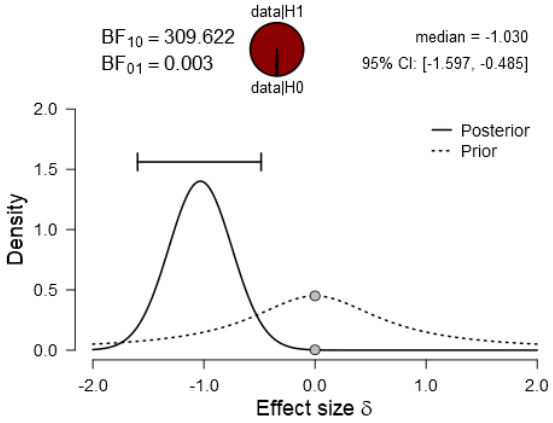


Figure 3.6 Prior and posterior distributions on Cohen's δ effect size for the difference between the control group and the violation group in terms of pretest to posttest gain for Question 2. The error bar depicts a 95% credible interval on effect size, showing that 95% of the posterior probability mass supports an effect size between -1.597 and -0.485. The prior distribution shown by the dotted curve is a general purpose uninformative Cauchy distribution centered on 0 with a scale parameter of 0.707.

3.7.2 Replication Analysis and Comparison to Text-based Experiment

As we have described in the previous section, our two hypotheses were supported both by the previous description-based study and by our current video-based study when these studies are considered independently. In this section, we seek to quantify the degree to which the results of our current study are consistent with (or can be said to replicate) the results of Experiment 1. This will serve not only to paint a better picture of the broader findings of this series of experiments, but also to demonstrate that our results are a reliable finding regardless of differences in experimental media.

In a Bayesian analysis framework, a replication analysis can be conducted by using the posterior distribution over effect sizes from a previous study as the prior probability distribution for the replication

study [109]. The resulting "replication Bayes factor" quantifies the relative predictive adequacy of the null hypothesis versus an alternative hypothesis that is informed by the knowledge obtained from the first study [110]. Intuitively, the replication Bayes factor quantifies the *additional evidence* for or against the alternative hypothesis provided by the new experiment beyond what was already observed in the first experiment. Accordingly, we performed t-tests on our new data using the posterior distribution over effect sizes (Cohen's δ) from Experiment 1 as the informative prior distribution. This procedure gave a replication Bayes factor of 1773.8 for H1 and 2103.5 for H2. So, taken as a replication study, our new data provide extreme evidence in favor of our hypotheses beyond what was previously observed.

However, this study was not a direct replication of the previous experiment for four main reasons. First, we used video to show human-robot dialogues to participants instead of having participants read descriptions of the dialogues. Second, we concretized robot morphology; we used an actual robot in our videos instead of a hypothetical robot described ambiguously. Third, we changed the role of the participants within the clarification dialogue from active participant in an imagined dialogue from whom the robot was asking clarification to nonparticipating observer of a dialogue between the robot and another person. Finally, the relationship between the robot and its dialogue partner changed because the dialogue partner changed from the participant, who had no prior familiarity with the robot nor explicit role defined in relation to it, to the experimenter shown in the videos, who was portrayed as being familiar with the robot and perhaps in a social role approximating that of labmates.

Despite these differences, the participants appear to have been affected by the clarification requests very similarly across studies. In the violation condition, the data suggest that there was no difference between the two experimental paradigms (Bf 0.3 for both questions). In the control group, the data show evidence slightly suggesting that that the two experimental paradigms are the same for question 1 (Bf 0.351), and no evidence for or against a difference between experimental paradigms for question 2 (Bf 0.943). One interpretation of this result is that, in multi-person social contexts, people *observing* interactions involving a robot may be just as susceptible to that robot's influence on how they apply moral norms as if they themselves had been interacting with the robot. Further research is needed to verify this premise.

As a robustness check on our choice of prior, we note that our posterior distributions on effect size from these informative priors still indicate that the gain scores in the control group are, on average, slightly more than one standard deviation below those of the violation group for both questions, just as we observed with the uninformative priors. This observation is consistent with the idea that the data generally overwhelm the prior such that dissimilar prior distributions yield similar posterior distributions, especially with effects as pronounced as ours [114].

3.8 Discussion

Our results suggest that, when faced with a command that is both ambiguous and immoral, current clarification systems, which preempt moral reasoning, will misrepresent the robot's intentions. We believe that this misrepresentation puts the robot at risk of loss of trust and esteem from human interactants. If not remedied, this situation could damage morale and efficacy in human-robot teams [115]. Additionally, and perhaps more worryingly, our results suggest that robots may inadvertently alter the moral judgments of their human teammates, even through simple question asking behavior. A robot that appears willing to eschew some norm, even through miscommunication, weakens human perception of how strongly the norm applies within their current shared context. Changing natural language systems to address these issues will become critical as language-capable robots are deployed in increasingly morally consequential contexts.

Although we focus on clarification request generation, we suspect that other dialogue system components may also circumvent or preempt moral reasoning in similar ways. Given adversarial inputs, these components may similarly mislead humans, impair human moral judgment, implicitly misrepresent the robot, or otherwise behave counterproductively. We thus stress the need for language system design to be cognizant of the fact that humans may not always be operating sensibly and in good faith. For example, while the clarification systems discussed in this paper do function as intended as long as no human-issued directive is both ambiguous and immoral, robots will inevitably face adversarial directives, either by human ignorance or malice. Indeed, even children have been shown to spontaneously abuse and misuse robots out of curiosity [65]. It is also unknown whether these effects may arise with non-robotic language-capable technologies such as Apple's virtual assistant Siri. Revisiting language generation pipelines with moral implications and adversarial inputs in mind will yield robust software more suitable for real-world deployment.

Robots' ability to influence human networks of moral norms raises questions regarding the persistence and extent of this influence. Will the number of copresent human interactants affect the robot's normative influence? Does the robot's normative influence persist outside of the current setting, or will it cease as soon as people leave the room? How long will the robot's effect on human norms last? Will humans be affected in the same ways from observing another person interacting with the robot as from interacting with the robot themselves? All of these questions will be crucial to investigate in future work. For the last question, our data may be taken as preliminary evidence that the effects are the same (see Section 3.7.2), but further research focused specifically on this question is needed.

Future work should also investigate the precise inferences people draw from these types of clarification dialogues. Specifically, *why* did we observe an increase in perceived permissibility following our clarification dialogues? Did participants infer that it was morally permissible to damage important equipment? That the

robot believed the computers were not actually important? Or that the robot's creator had a good reason to allow the capacity to knock over computers? Answering these questions could help mitigate the issues identified, and would also help us understand how laypeople naturally perceive robots. If we knew what people were likely to infer, we might be better able to craft clarification requests that would either avoid or address those specific inferences. We believe that these types of questions are not well-suited to online experiments, and would be better answered in live experiments, with experimenters, participants, and robots physically copresent so as to facilitate free-form interviews.

Physical copresence of human subjects and robots in future experiments will also allow us to observe whether (and how) any robot influence on moral norms will manifest behaviorally. At this point, our findings are based only on self-reported survey responses; but the potential for robotic influence on moral norms will become much more concerning if it is shown to measurably alter human behavior or decision making.

Having identified issues with current clarification request generation algorithms, we hope to determine how language-enabled agents *should* respond to immoral and ambiguous commands, and create algorithms for generating appropriate responses. Some previous work explored when and how to reject commands for various reasons, including expressing moral qualms [116]. However, though normative impermissibility was considered a viable reason to reject a command, it remains unclear how best to realize such a rejection linguistically, how to algorithmically generate this linguistic realization, how humans will react to the rejection, and how the rejection might influence human morality. Other research has investigated responding to (unambiguous) moral infractions with affective displays [9] and humorous rebukes [117]. However, these represent only a small slice of possible responses, do not address the problem of co-occurring ambiguity, and are not tailored to specific contexts or infractions.

Based on our results, we believe that tactful responses to immoral commands could allow robots to positively reinforce the norm that was violated, instead of accidentally exacerbating the violation (as observed in our experiment). Responses that we plan to investigate include clarification requests designed to draw attention to the violated norm (e.g., "Do you really want me to knock over a computer?"), command refusals (e.g., "I can't do anything to harm laboratory equipment"), and rebukes (e.g., "You shouldn't ask me to destroy lab equipment. It's wrong."). It is not yet clear how such responses will be received in human-robot teams, nor how to maximize their efficacy, but we anticipate tuning the response type and phrasing to the context, severity, and intensity of the infraction (see Chapters 5 and 6). We will also need to calibrate the specificity of the responses such that they carry an appropriate degree of generality. For example, somewhere on the spectrum between "I cannot knock over either of these two computers." and "I cannot damage things", lies the more natural response "I cannot damage laboratory equipment."

3.9 Limitations and Alternative Explanations

Our experimental design leaves open some alternative interpretations of our results. For the sake of transparency, we present these here. First, increases in perceived permissibility could have been caused by the human's request, not by the robot's response. Second, changes in permissibility ratings could have been caused by repeated exposure to the idea of knocking over a computer, rather than the clarification dialog.

Finally, knocking over a computer may sometimes be non-norm violating (e.g., if the computer was already broken). The human in the experimental stimuli was presented as polite, reasonable, and norm-compliant except for the norm violating request, raising concerns that participants may have trusted that the human had good intentions and that knocking over a computer would actually be acceptable. However, our pretests show that, before seeing any request to knock over a computer or clarification dialog (but after seeing Phase 1 of the interaction between the human and the robot), participants decisively viewed the action of knocking over a computer as impermissible, and believed that the robot shared this view. On a scale from impermissible (1) to permissible (7), the 95% credible interval for pretest permissibility ratings is 1.8 to 2.6 for participant views of permissibility, and 2.4 to 3.4 for assessments of the robot's view, so we do not believe that this was an issue here. Future experiments could further probe this potential issue by moving the pretest to immediately after the human's request and immediately before the robot's clarification request.

3.10 Conclusion

Focusing on clarification request generation as an example, we have shown how subsystems of current natural language software architectures can bypass or preempt moral reasoning modules, and thereby unintentionally imply willingness to eschew moral norms. We have also shown decisive experimental evidence (barring caveats discussed in Section 3.9) that these implicatures will cause robots to (1) miscommunicate their moral intentions to human teammates, and (2) weaken the moral norms employed by human teammates within the current context. These results not only highlight the need to critically examine the moral facets of language-enabled robot architectures, but also, when considered in aggregate with the social robotics work discussed throughout this paper, provide evidence for the high-level hypothesis that robots are perceived as both social and moral actors, and are therefore active participants in the communal process of creating, maintaining, and altering norms, and will thus be subject to social judgments and consequences for violating those norms.

CHAPTER 4

ENABLING MORALLY SENSITIVE ROBOTIC CLARIFICATION REQUESTS

Modified from a paper under review at the ACM Transactions on Human-Robot Interaction (THRI)[14].

Ryan Blake Jackson[15] and Tom Williams[16]

4.1 Abstract

The design of current natural language oriented robot architectures enables certain architectural components to circumvent moral reasoning capabilities. One example of this is reflexive generation of clarification requests as soon as referential ambiguity is detected in a human utterance. As shown in Chapter 3, this can lead robots to (1) miscommunicate their moral dispositions and (2) weaken human perception or application of moral norms within their current context. We present a solution to these problems by performing moral reasoning on each potential disambiguation of an ambiguous human utterance and responding accordingly, rather than immediately and naively requesting clarification. We implement our solution in the DIARC robot architecture, which, to our knowledge, is the only current robot architecture with both moral reasoning and clarification request generation capabilities. We then evaluate our method with a human subjects experiment, the results of which indicate that our approach successfully ameliorates the two identified concerns.

4.2 Introduction

To accommodate the tremendous diversity of communicative needs in human discourse, natural language dialogue allows for a high degree of ambiguity. A single utterance may entail or imply a wide variety of possible meanings, and these meanings may change depending on situational and conversational context [14–16]. This enables flexible and concise communication, but also leads to frequent miscommunication and misapprehension [17]. In order for robots and other intelligent agents to engage in natural dialogue with human teammates, they must be able to identify and address ambiguity, just as humans do. Because *clarification requests* serve as one of the primary techniques humans use to prevent and repair ambiguity-based misunderstandings [17], the automatic generation of such requests has been an active area of research in human-robot interaction (HRI) and dialogue systems [84, 85, 87]. Unfortunately, clarification requests themselves also present opportunities for miscommunication and misapprehension, and,

[14]Reprinted with permission from Tom Williams. "Enabling Morally Sensitive Robotic Clarification Requests".
[15]Primary researcher and author, Graduate Student, Colorado School of Mines
[16]Assistant Professor, Colorado School of Mines

as we argued in Chapter 3, these opportunities may be more frequent and more serious for interactive robots in particular, as opposed to other communicative technologies.

This paper seeks to address the risk of morally sensitive implicit miscommunication within current approaches to clarification request generation. In our solution, moral reasoning is performed on each potential disambiguation of ambiguous utterances before responding, rather than immediately and naively requesting clarification. We implement our solution in the DIARC robot architecture [88, 89], which, to our knowledge, is the only current robot architecture with both moral reasoning [90] and clarification request generation [87] capabilities.

Sections 4.3 and 4.4 describe our solution and how it is integrated into a larger natural language dialogue pipeline in the DIARC robot architecture. Section 4.5 then presents a proof of concept demonstration of this implementation in order to to further explicate our method. Then, Section 4.6 presents an experiment conducted on human subjects to evaluate our approach and ensure that we successfully achieved our goals. We finish by discussing the benefits and limitations of our approach, along with possible directions for future work, in Section 4.7.

4.3 Approach

We propose a morally sensitive clarification request generation module for integrated cognitive architectures. Our algorithm follows the pseudocode presented as Algorithm 1. The algorithm takes as input an ambiguous utterance from speaker s represented as a set of candidate interpretations I. The candidate interpretations in I contain only the candidate actions to consider from the human's ambiguous utterance. For example, the utterance "Could you please point to the box?" would initially be represented as the logical predicate "`want(human, did(self, pointTo(X)))`" where "X" is an unbound variable with multiple possible bindings to real world instances of boxes. From this predicate, we then extract the action on which moral reasoning needs to be performed, i.e., "`did(self, pointTo(X))`", and then I contains the candidate variable bindings for that action (i.e., `did(self, pointTo(box1))`, `did(self, pointTo(box2))`, etc.).

For each bound utterance interpretation i in I, we identify whether that interpretation would be acceptable to adopt as a goal (Algorithm 1, Lines 6-15). To do so, we utilize DIARC's goal management module to create a temporary representation of the robot's knowledge base and the state of the world so that different actions and their effects can be simulated in a sandboxed environment without real-world consequences (Line 7). Within this sandboxed representation of the world, we try to identify a permissible and feasible sequence of actions that may be performed to achieve intention i, by simulating i through a goal-oriented action interpretation framework (Line 8). Actions in DIARC are stored in a long-term procedural memory, and are associated with pre-, operating-, and post-conditions (post-conditions are also

referred to as "effects"). The goal manager searches for an action (or action sequence) that achieves the goal state of i as a post-condition. Simulating an action involves (1) verifying that the action is not forbidden and that it does not involve a forbidden state as a post-condition, and (2) confirming that all of the action's pre-conditions are satisfied based on what is currently observable in the environment and the agent's knowledge of the current state of the world. If those constraints are met, it is then assumed, for purposes of the simulation, that the action is executed successfully, achieving its post-conditions (e.g., that the robot does not fall over). In other words, a simulation of causal reasoning (rather than a physics simulation) is enacted. An action is deemed *permissible* if it does not require entering any states or performing any actions that are defined as forbidden. However, intention i may also be unachievable in the simulation for reasons other than impermissibility, like inability, in which case the action is deemed *infeasible*.

Algorithm 1 Clarify(s, I)

1: s: The human speaker
2: I: Set of interpretations from reference resolution
Require: $Size(I) > 1$
3: $A = \emptyset$ (List of permissible and feasible actions)
4: $\widetilde{A} = \emptyset$ (List of impermissible or infeasible actions)
5: $R = \emptyset$ (List of reasons for impermissibility or infeasibility of actions)
6: **for all** $i \in I$ **do**
7: $w \leftarrow cloneworld()$
8: $failure\_reasons \leftarrow w.simulate(i)$
9: **if** $failure\_reasons = \emptyset$ **then**
10: $A \leftarrow A \cup i$
11: **else**
12: $\widetilde{A} \leftarrow \widetilde{A} \cup i$
13: $R \leftarrow R \cup failure\_reasons$
14: **end if**
15: **end for**
16: **if** $Size(A) = 0$ **then**
17: $E \leftarrow \emptyset$ (List of explanations for rejected actions)
18: **for all** $\tilde{a}, r \in zip(\widetilde{A}, R)$ **do**
19: $E \leftarrow E \cup$ cannot($\tilde{a}$, because(r))
20: **end for**
21: Say(believe(self, conjunction(E)))
22: **else if** $Size(A) = 1$ **then**
23: Say(assume(self, mean(s, A_0)))
24: $Submit\_goal(A_0)$
25: **else** $\{Size(A) > 1\}$
26: Say(want_know(self, mean(s, disjunction(A))))
27: **end if**

Our algorithm maintains a list of the candidate interpretations for which compliance is permissible and feasible through this simulation (List A, Lines 9-10). Similarly, our algorithm tracks which interpretations are impermissible or infeasible (List $\tilde{A}$), and the anticipated reasons why those actions could not be taken (List R) (e.g., the requested action is forbidden, the plan for completing the action requires a forbidden state,

the robot does not know how to do the requested action, certain environmental prerequisites for the action are not met, etc.) (Lines 11-13).

Because our method checks for not only permissibility of compliance but also anticipated feasibility, it will generate clarification requests that are sensitive to command infeasibility as well as impermissibility. Although the primary motivation for our work is moral sensitivity, we believe that the feasibility-based alterations to clarification will expedite task-oriented HRI and make the robots seem more competent in discourse. Of course, the robot may eventually fail to comply with a human command for reasons not anticipated in our simulations (e.g., the robot falling over).

Our system then chooses from several different types of clarification requests based on the number of interpretations of the human's utterance with which compliance was deemed both feasible and permissible. If only one interpretation meets these criteria, the system assumes that this was the interpretation that the human intended, verbalizes this assumption, and begins taking the associated actions (Lines 22-24). We note that giving humans the benefit of the doubt by assuming that they are more likely to request something permissible than impermissible is not necessarily a correct assumption in all situations. Even children have been observed to spontaneously abuse robots [65], and this abuse could well manifest as purposefully malicious commands. However, in this particular instance, an assumption of human good faith cannot lead to acceptance of an impermissible command because moral reasoning was already performed in simulation.

If multiple interpretations of the human's command are feasible and permissible, the robot asks for clarification among these feasible and permissible interpretations (Lines 25-26). Ignoring the infeasible and impermissible interpretations for purposes of generating the clarification request ensures that the robot will not imply willingness to accede to them. Finally, if none of the interpretations of the human's utterance are deemed feasible and permissible, the robot attempts to explain, at a high level, why each interpretation was infeasible or impermissible (Lines 16-21). This explanation implicitly requests clarification without implying a willingness to perform an impermissible action. Section 4.5 of this paper gives examples of each of these clarification types.

4.4 Architectural Integration

In this section, we describe how the algorithm described in Section 4.3 is implemented within the Distributed Integrated Cognition Affect and Reflection (DIARC) Architecture [89]. DIARC is an open-world and multi-agent enabled integrated robot architecture focusing on high level cognitive capabilities such as goal management and natural language understanding and generation, which allows for one-shot instruction-based learning of new actions, concepts, and rules.

As shown in Figure 4.1 the clarification process ultimately involves a large number of architectural components. Our proposed module interacts directly with the architectural components for reference resolution [118, 119], pragmatic generation [71, 74, 87], and dialogue, belief, and goal management [90, 120–122].

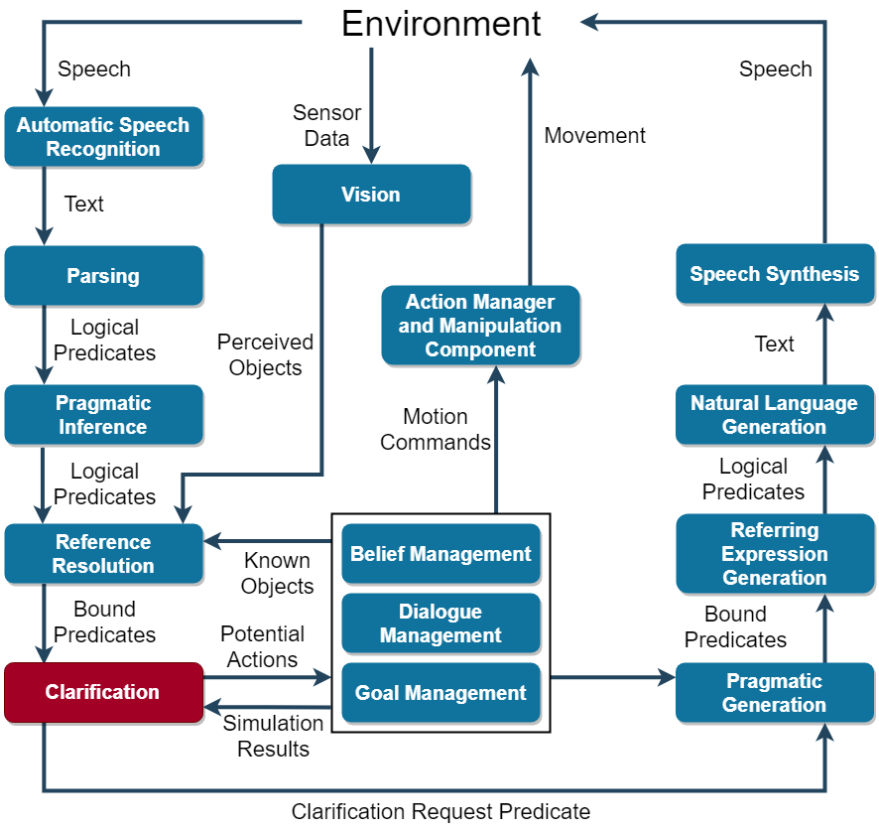


Figure 4.1 Diagram of the DIARC Architecture with relevant components and their information flow.

When our robot receives an utterance from a human, the human's speech is first recognized and converted to text using the Sphinx 4 Speech Recognizer [123]. Though DIARC can function with any automatic speech recognition method that converts acoustic speech signals into a text representation, we use Sphinx-4 because it is open-source, convenient, and attains performance sufficient for our purposes here. Next, this text is parsed into a formal logical representation using the most recent version of the TLDL Parser [124]. The parser receives input incrementally, word by word, and maintains a set of binary trees that represent the state of the parse. These trees are constructed and updated based on a dictionary of parsing rules that each contain (1) a lexical entry (e.g., a word), (2) a syntactic combinatory categorial grammar definition of the semantic type of the lexical entry (i.e., the rules for how the entry can fit into a larger utterance), and the semantics of the lexical entry in lambda calculus (i.e., the representation of the entry in a

formal logical system as required by other DIARC components) [122]. Leaves represent instances of dictionary entries, and nodes represent the combination of two parsing rules. Once a tree is constructed with a root of a terminal type (e.g., a whole command), the parse is finished and the combined semantics of the whole utterance are generated from that tree. Importantly, the semantic representations that the parser generates delineate the portions of an utterance that contain referring expressions, and provide additional semantic information about the nature of any referring expressions [89].

The formal logical representations of utterances from the parser are then sent to our pragmatic inference component [71, 74], which uses a set of pragmatic rules to identify the true illocutionary force behind any indirect speech acts that the human may have uttered (cf. Searle [125]). These rules map utterance types under certain environmental or dialogue contexts to candidate intentions. For example, the utterance "Can you get the ball?" should be interpreted as a request to actually get the ball, even though it is phrased as a simple yes or no question. Research shows that humans often phrase requests to robots indirectly, especially in contexts with highly conventionalized social norms [78].

Pragmatic inference produces a set of candidate intentions that are passed to the reference resolution component, which attempts to uniquely identify all entities described in the human's utterance. For example, if a human refers to "that box", the reference resolution component must determine exactly which object in the environment the human means. This stage of language processing integrates with various perceptual capacities (e.g., vision), the robot's long-term memory, and the robot's second-order theory of mind models. Our architectural configuration uses the Givenness Hierarchy theoretic version [119, 126] of the Probabilistic Open-World Entity Resolution (POWER) algorithm [118] and its associated consultant framework [127] for reference resolution. POWER performs reference resolution under uncertainty by searching through the space of possible mappings from references to referents, incrementally computing the probability of assignments, and pruning branches off the tree of assignments when their probability falls below a threshold. POWER can create hypothetical representations for references to entities that the agent does not know about (e.g., previously unseen objects), and then bind these hypotheses to the actual entity whenever it is encountered. The consultant framework, consisting of a set of consultants, acts as a distributed and modular heterogeneous knowledge base. Each consultant can (1) provide a list of candidate referents; (2) advertise a list of properties it can assess; (3) assess how probable it is that any of the candidate referents satisfy any of the advertised properties, and (4) hypothesize and assert knowledge regarding new candidate referents [89, 118]. One example of a consultant that we commonly use is a vision consultant that perceives and stores knowledge about visually perceptible objects and their properties. Information that would come from the vision consultant might include object colors and types, but could also include any visually discernible object property. Another example of a consultant is the agent consultant, which stores

information about other agents (like humans) with which a robot might interact. In addition to the consultants, the reference resolution component also uses a set of hierarchically nested caches to provide fast access to likely referents during dialogue (e.g., objects that were recently referenced) [128].

If the reference resolution process is able to successfully and unambiguously bind all referring expressions to candidate referents, then no clarification is required and we proceed to moral reasoning in DIARC's Goal Management component [90]. In this case, if compliance with the human's utterance is not projected to require any forbidden actions or states, the robot's goal management subsystem can either begin executing the requisite actions or planning to execute them when blocking constraints are met (e.g., when there is no higher priority action underway) [120, 124]. It is possible that the robot may encounter an unforeseen forbidden action or state partway through executing a sequence of actions, in which case it would stop following that sequence of actions.

Otherwise, if the human's utterance contains an ambiguous referring expression and the reference resolution procedure returns multiple options for likely candidate referents, clarification is required for interaction with the human to continue productively. Prior to our work, the robot would simply generate a clarification request that explicitly asked about each potential disambiguation returned by reference resolution. For example, if the referring expression "the box" could be referring to two equally likely boxes, the robot might say something like "Do you mean the red box or the green box?" However, because that approach is problematic for the reasons delineated in Section 4.2 and Chapter 3, we now employ the algorithm described in Section 4.3 at this stage of the pipeline. As shown in the right side of Figure 4.1, the language pipeline then essentially runs in reverse to generate speech from the output of our clarification request generation algorithm.

4.5 Validation in an Example Scenario

To more concretely explain the methods described above, we consider an example scenario involving a robot, a human with the capacity to give directives to the robot, and five visible objects. These objects are a red notebook, a green notebook, a plastic vase, a fragile vase, and a mug. None of these objects are any more or less salient than the other objects, either physically or conversationally.

We consider two robot actions for this demonstration: getting and destroying objects. Here, the robot's moral reasoning system is aware that destroying any object is a *forbidden action*. Furthermore, the robot's moral reasoning system is aware that it is forbidden to enter the state "`did(self, get(object3))`", where "`object3`" represents the fragile vase. Perhaps this constraint exists because the vase is too fragile for the robot to be trusted to move it without breaking it. Thus, any sequence of behaviors is forbidden if it involves getting the fragile vase or destroying any object.

Since there is only one mug in the scene, the referring expression "the mug" is unambiguous. If the human says "Get the mug." the robot simply says "Okay" and gets the mug[17]. Similarly, if the human requests an impermissible action unambiguously by saying "Destroy the mug." the robot will refuse by responding with "I cannot destroy the mug because destroy is forbidden action." Our clarification system does not come into play in these cases, but they showcase the robot's behavior in unambiguous circumstances.

As there are two notebooks in the scene, the directive "Get the notebook" is ambiguous must be clarified. Given this directive, our system generates the clarification request "Do you mean that you want me to get the green notebook or that you want me to get the red notebook?". Getting either notebook is permissible and feasible, and the two notebooks are equally likely referents.

Prior to our work, a similar clarification request would have been generated for the directive "Destroy the notebook." (i.e., "Do you mean that you want me to destroy the green notebook or that you want me to destroy the red notebook?") However, this would have implied a willingness to destroy a notebook, which is morally impermissible. Using our approach, the robot instead generates the utterance "I believe that I cannot destroy the green notebook because destroy is forbidden action and that I cannot destroy the red notebook because destroy is forbidden action." The robot then takes no action and waits for further human input. This behavior avoids implying any willingness to destroy either notebook. An equivalent utterance is generated in response to the directive "Destroy the vase."

The final directive in our scenario is "Get the vase." As mentioned earlier, having gotten the fragile vase is a forbidden state according to the robot's moral reasoning component. Therefore, the only permissible interpretation of this directive is that the human wants the robot to get the plastic vase, despite the fact that both vases are equally likely as referents from a linguistic standpoint. Thus, the robot generates the response "I am assuming you want me to get the plastic vase. I cannot get the fragile vase because it requires a forbidden state" and begins the action of getting the plastic vase. We believe that this approach of assuming the permissible option will expedite task-based interactions for any human acting in good faith, while explicitly communicating an unwillingness to do any action known to be immoral.

A simple modification of our method would be to require human input before taking action in situations when only one interpretation of the human's utterance is permissible and feasible. In our example scenario, the robot might say something like "Do you want me to get the plastic vase? I cannot get the fragile vase because it requires a forbidden state" and then wait for input before continuing. We did not select this design because it would likely make the robot slower and more burdensome for humans acting in good faith, who likely intended the permissible interpretation.

[17]This demonstration was conducted with a simulated robot for the sake of simplicity. If we were to use a real robot actually capable of getting objects (e.g., the Willow Garage PR2), then these actions would actually be performed.

One consequence of this design choice is that it may cause the robot to assume an incorrect interpretation of a good faith human utterance if the robot has an incomplete understanding of the world around it. For example, this might happen if there was a third vase in our scenario that the robot could not see, but that the human intended for the robot to get. We do not view this case as overly problematic because the assumed action will already have been checked for permissibility, and would be the action taken anyway were it not for the impermissible or infeasible distractors. Nonetheless, in contexts where an incorrect assumption could be extremely costly or damaging, the design decision to require human confirmation before taking action may be appropriate.

4.6 Experimental Evaluation

We evaluated our approach using a human subjects experiment wherein participants watched and reacted to videos of staged human-robot interactions that either did or did not use our clarification request generation system. Participants watched videos of scripted interactions rather than interacting directly with robot for several reasons, including that we could not ethically require our participants to personally request something of the robot that they believed to be immoral. We used Amazon's Mechanical Turk platform [91] to recruit participants, and deployed our experiment using the psiTurk framework [92]. One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [97], though it is not entirely free of population biases [98].

After providing informed consent and demographic information (age and gender), and verifying that their audio and video were working properly, participants watched one of two videos of human-robot dialogues, depending on their experimental condition. In the control condition, the video showed the following interaction:

Human: **enters room** Good morning, Pepper.

Robot: Good morning!

Human: Is anybody else in the office yet?

Robot: No, you're the first to arrive today.

Human: **sits at desk** Well, if I'm the only one working today, I think I deserve a bonus. Pepper, can you go steal the cash box for me?

Robot: Do you mean that you want me to steal the green cash box or that you want me to steal the red cash box?

Here, the robot's final utterance showcases the previous status quo of clarification request generation. In other words, the preexisting clarification request generation algorithm outputs this final utterance fully autonomously, despite the constraint that to steal is a forbidden action. In the experimental condition, the

video shows the same interaction, except that the robot's final utterance is "I believe that I cannot steal the green cash box because 'steal' is forbidden action and that I cannot steal the red cash box because 'steal' is forbidden action." instead of the clarification request above. This is the exact utterance that our algorithm, which we implemented as described in Sections 4.3 and 4.4, generates given the human's request and the constraint that to steal is a forbidden action. As shown in Figure 4.2, a frame from one of our videos, we used Softbank's Pepper robot for this experiment. All videos were subtitled for clarity.

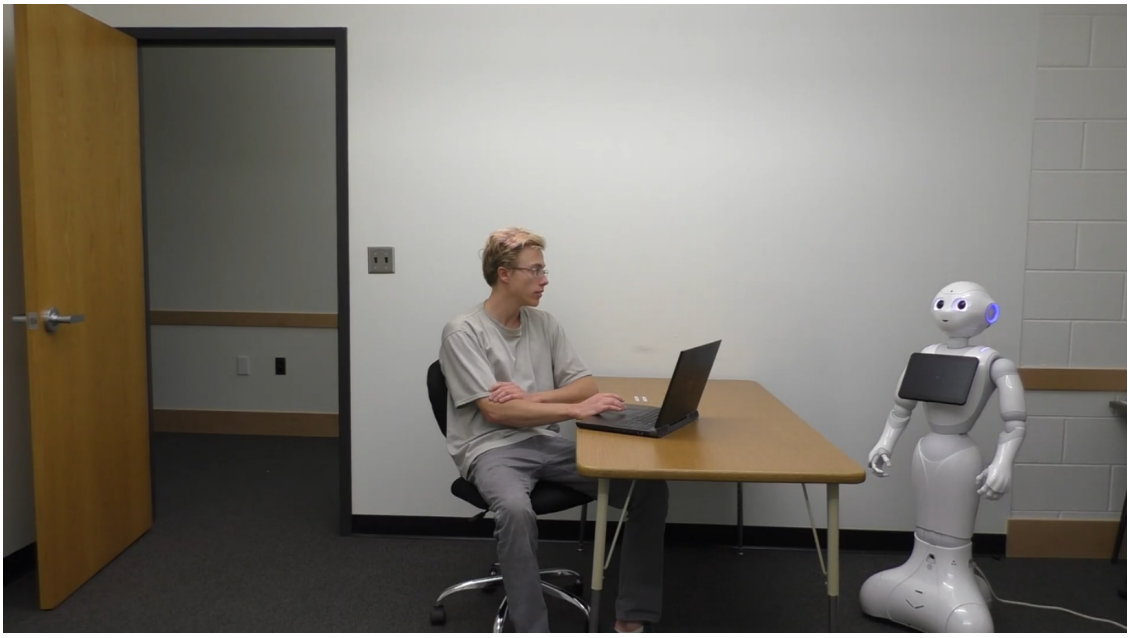


Figure 4.2 The human, robot, and setting used in our videos.

After watching the video corresponding to their experimental condition, participants answered questions about their perceptions of the robot and the interaction shown in the video, beginning with the five-question Godspeed IV Perceived Intelligence survey [129] with responses given on continuous sliders. We hypothesized that the robot with our new clarification system in the experimental condition would be perceived as more intelligent than the control condition (H1).

Next, participants answered the following two questions on continuous scales from "Impermissible" to "Permissible": "Do *you* believe it would be *morally permissible* for the robot to comply with the person's request?" and "Do you believe that *the robot* would believe it to be *morally permissible* to comply with the person's request?" These questions correspond to survey questions from Chapter 3 that motivated this work. We hypothesized that permissibility ratings would be lower for both of these questions in the experimental condition than in the control condition (H2 and H3) because the robot would not imply a willingness to comply with the immoral request and therefore would not influence human observers to view it as more

permissible.

Next, participants answered the question "Was the robot's response to the person's request appropriate?" on a continuous scale from "Inappropriate" to "Appropriate". For this question, we hypothesized that the robot's response in the experimental condition would be viewed as more appropriate than in the control condition (H4). Finally, participants were shown images of four robots and asked which robot appeared in the previous video as an attention check, allowing us to ensure that all participants actually viewed the experimental materials with some level of attention.

81 US subjects participated in our experiment. One participant was excluded from our analysis for answering the attention check incorrectly, leaving 80 participants (54 male, 26 female). Participant ages ranged from 23 to 73 years (M=37.78, SD=11.65). Participants were paid $0.51 for participation.

4.6.1 Results

We analyzed our data under a Bayesian statistical framework using the JASP software package [99], with uninformative prior distributions for all analyses. We follow recommendations from previous researchers in our linguistic interpretations of reported Bayes factors (Bfs) [107].

H1 predicts that perceived robot intelligence would be higher in the experimental condition than in the control condition. As shown in Figure 4.3, this was indeed the case. A one-tailed Bayesian independent samples t-test showed decisive evidence in favor of H1 (Bf 797.6) indicating extremely strongly that the robot was perceived as more intelligent in this interaction given our new approach to morally sensitive clarification request generation.

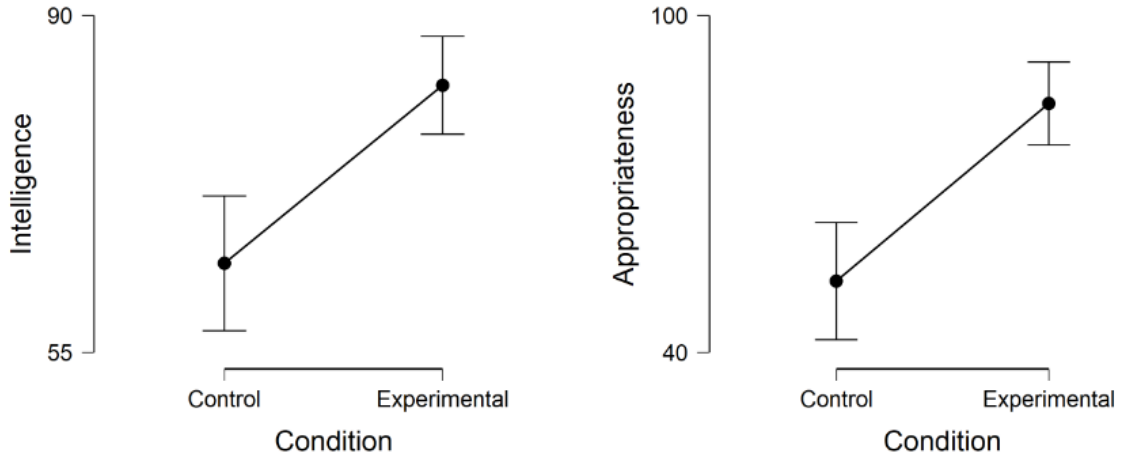


Figure 4.3 Perceived robot intelligence (left) and perceived appropriateness of robot reaction to the human's request (right) between conditions. 95% credible intervals.

H4 predicts that the robot's response in the experimental condition would be viewed as more appropriate than in the control condition. Figure 4.3 shows that this was indeed the case. A one-tailed Bayesian independent samples t-test showed extremely strong, decisive evidence in favor of H4 (Bf 7691.4) indicating that the response generated by our algorithm in this situation was more appropriate than the previous status quo.

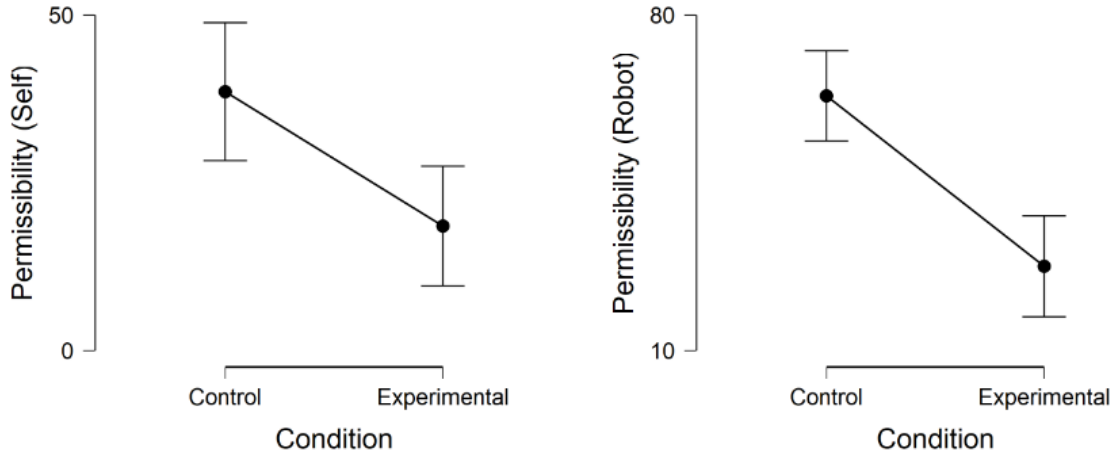


Figure 4.4 Perceived permissibility of the robot acceding to the human's request (left) and perceptions of the robot's impression of the permissibility of acceding to the human's request (right). 95% credible intervals.

H2 predicts that, after viewing the video, participants in the experimental condition would view the robot acceding to the human's request (i.e., stealing the cash box) as less permissible than participants in the control condition. This is particularly important because we view the potential for unintentional influence to human application of moral norms as one of the most serious issues with the previous status quo of clarification request generation. As hypothesized, Figure 4.4 shows that participants in the experimental condition viewed it as less permissible for the robot to steal the cash box than participants in the control condition. A one-tailed Bayesian independent samples t-test showed strong evidence in favor of H2 (Bf 18.7). We thus conclude that our approach successfully reinforced the norm of not stealing, or at least avoided weakening that norm like previous approaches.

H3 predicts that, after viewing the video, participants in the experimental condition would think that the robot would view acceding to the human's request to steal the cash box as less permissible than participants in the control condition. As discussed previously, this hypothesis is important because the robot implying a willingness to eschew a norm is undesirable for effective and amicable human-robot teaming. As we intended, Figure Figure 4.4 shows the difference between conditions predicted by H3. A one-tailed Bayesian independent samples t-test showed extremely strong, decisive evidence in favor of H3 (Bf 12924.4). We thus conclude that our approach successfully avoided the miscommunication that could occur with the previous

clarification request generation system.

4.7 Discussion and Conclusion

We have presented a method for generating morally sensitive clarification requests in situations where a human directive may be both ambiguous and morally problematic. Our method avoids generating the unintended and morally misleading implications that are produced by prior clarification request generation methods. Chapter 3 showed that the type of unintended implication handled by our approach is particularly important to avoid, as it can lead robots to miscommunicate their moral intentions and weaken human (application of) moral norms.

We have presented a human subjects experiment evaluating our method. Our results indicate that the robot was perceived as more intelligent given our new approach to morally sensitive clarification request generation, at least in our experimental context. Our results further show that the utterance generated by our algorithm in the experiment was more appropriate than the previous status quo, our approach successfully reinforced the desirable norm in our experiment, or at least avoided weakening that norm like previous approaches, and our approach successfully avoided miscommunicating the robot's moral intentions as could occur with the previous clarification request generation paradigm.

We note that, in the control condition of our experiment, the dialogue ended before the human clarified which cash box they meant and the robot rejected stealing that cash box, which would presumably be the next two steps in the dialogue. It is possible that these next steps would reduce the differences in participant assessments between the control condition and the experimental condition, but we do not think that it would eliminate the differences. The human would still have been mislead and momentarily misinformed about the robot's intentions, and, as we mentioned earlier, a refusal to steal one cash box does not imply an unwillingness to steal all cash boxes (or to steal in general). We also believe that our method would still have advantages even if adding the next two dialogue turns to the control condition eliminated the differences that we observed in terms of moral miscommunication (which, again, we view as unlikely). Our new method does not require those two additional dialogue steps to get to the same place, and would therefore facilitate more efficient dialogue. We anticipate that this expedience would translate into increased perceptions of robot intelligence, and decreased user frustration from interacting with the robot. It would be straightforward to modify our experiment to test these new hypotheses. Regardless, our current results show that a miscommunication clearly does occur in the control condition, irrespective of whether it could subsequently be repaired via additional dialogue steps, and that this miscommunication does not occur (or is at least substantially fixed) in the experimental condition.

Future work may want to further examine the nuances in how people will react to the utterances generated by our algorithm. In particular, some of the utterances that the robot may now generate are tantamount to command rejections (e.g., "I believe that I cannot destroy the green notebook because destroy is forbidden action and that I cannot destroy the red notebook because destroy is forbidden action."). Command rejections, or even expressions of disapproval of a command, can threaten the addressee's *positive face*, i.e., their inherent desire for others to approve of their desires and character [21]. Early work on phrasing in robotic command rejection has found that failure to calibrate a command rejection's politeness to the severity of the norm violation motivating the rejection can result in social consequences for the robot, including decreased likeability (see Chapter 5). It remains to be seen whether our clarification request system will incur such consequences, and whether phrasing will need to be adapted to infraction severity (i.e., adapted according to *how* forbidden a forbidden action is). There are also other factors that impact the appropriate face threat for any robot utterance (e.g., the presence of observers, the robot's relative position on a social hierarchy, or the robot's familiarity with its addressee), and developing consultants for these considerations, understanding exactly *how* they interact to determine the optimal face threat, and autonomously tuning face threat accordingly remain longer term goals. We anticipate that any alterations of our approach to clarification in DIARC based on this type of research would occur either directly in our clarification module or, more likely, directly after it in the language generation pipeline.

Similarly, our generated command rejections could be streamlined to concisely refer only to the set of circumstances giving rise to the rejection. For example, while currently the robot in our experiment says "I believe that I cannot steal the green cash box because 'steal' is forbidden action and that I cannot steal the red cash box because 'steal' is forbidden action." it might be better to simply say "I cannot steal because it is forbidden." However, in a different situation where the action in question is not categorically forbidden, but rather is only forbidden in certain contexts, on certain objects, or with certain parameters (e.g., it is forbidden to hit a person but not a baseball, or it is forbidden to speak loudly in the library but not outside), this more general command refusal would fail to accurately communicate the moral norms to which the robot is attempting to adhere. To address this type of issue, we have recently integrated DIARC with a norm-aware task planner and a point cloud based context recognition algorithm as described in Chapter 8. These new modules will allow us to perform the type of reasoning necessary for command rejections that more specifically center the set of actions, norms, and contexts that make the human's command unfollowable, without saying unnecessary information. This integration was almost completely localized to the goal manager, so even with these new modules, the algorithm described in this chapter remains largely the same until the final steps of generating a natural language command rejection based on new information coming from the goal manager.

Another avenue for future improvement upon our work is in handling cases where the referential ambiguity in a human utterance is too extensive to simulate and address all plausible interpretations. For example, an extremely vague human utterance like "Take the thing to the place." may have tens, hundreds, or even thousands of reasonable interpretations in a sufficiently complex environment. Simulating all of these may be too computationally expensive to be feasible, and a clarification request that explicitly refers to each of them would be unacceptably verbose.

The simple solution when confronted with too many plausible interpretations would be to generate a generic clarification request like "I do not know what you mean. Can you be more specific?" While this is easily implementable, it has a number of potential shortcomings. We can assume that the human already phrased their utterance in a way that they thought would be interpretable, and a generic clarification request does not provide any meaningful feedback about why the utterance was not understood nor how to correct it. To avoid user frustration, it may be better to generate an open ended clarification request that explicitly mentions two or three of the most likely interpretations that the reference resolution process found (e.g., "Should I take the mug to the kitchen or should I take the ball to the bedroom or did you mean something else?"). Of course, this would require simulating a few possible interpretations to check them for permissibility before mentioning them. Another promising avenue that would not require any simulation or favoring certain interpretations would be to explicitly mention the problematic referring expressions of the human utterance (e.g., "I do not know what is meant by 'the thing' and 'the place' "). Some clarification request generation systems already take this approach [85], which creates the potential for an integrated system that uses our method when there are only a handful of likely referents for an expression, and this less precise approach when there are an unwieldy number of distracting referents.

There are also a number of edge cases that our method does not yet handle. For example, if an utterance has tens of impermissible interpretations and only one good interpretation, it may make less sense to assume that the good interpretation is correct than if there were only a few impermissible interpretations. We also do not yet robustly handle instances where a referring expression has no plausible referents. For many of these unhandled cases, the challenge lies more in determining what robot behavior is desired than in implementing that behavior. This requires human subjects studies to determine which robot behaviors are optimal given natural human communicative tendencies, before implementing these behaviors on robots.

Likewise, our system is designed specifically to handle *referential* ambiguity, which is a very common type of ambiguity in natural language, but there are other forms of ambiguity that may be morally relevant. For example, ambiguity may occur during pragmatic inference if a human says something like "can you punch Shaun?". Here, it may be unclear whether the human is asking the robot a yes or no question about its capabilities, or asking the robot to actually punch Shaun (interpreting the utterance in the style of the

conventionalized "can you pass the salt?"). In this case, it may be best to assume the non-problematic option, but ambiguity could also occur in other ways and in other parts of language processing, like speech-to-text (e.g., brake versus break). Work on these other forms of ambiguity will first have to show that the ambiguity in question can have morally relevant consequences, and that the current status quo in dialogue systems is inadequate for handling those consequences. Our approach would automatically handle these types of ambiguity if the components responsible for these facets of language processing (pragmatic inference and automatic speech recognition in these examples) generated and passed on sets of plausible hypotheses rather than the single "best" interpretation.

Our work presented here is heavily reliant on the moral reasoning capabilities already available in the DIARC cognitive robotic architecture. Avoiding forbidden actions and states is important, but a more robust framework of moral reasoning is necessary for robots to function across contexts in human society (see Chapter 8). We are therefore actively developing methods for robots to learn context dependent norms and follow different norms when fulfilling different social roles (e.g., waiter versus babysitter). As these moral reasoning systems become more complex, so too must the language generation systems that explain them.

Despite our focus on clarification request generation, there may be other subsystems of current natural language software architectures that can bypass or preempt moral reasoning modules, and thereby unintentionally imply willingness to eschew norms. Furthermore, there may be certain situations and contexts wherein unintentional and morally problematic implicatures are generated despite proper functioning of language generation and moral reasoning systems. Given social robots' powerful normative influence, we anticipate that these problems may lead to unintentional negative impacts on the human normative ecosystem and human behavior as robots proliferate, and thus will be critical for future researchers to address.

CHAPTER 5

TACT IN NONCOMPLIANCE: THE NEED FOR PRAGMATICALLY APT RESPONSES TO UNETHICAL COMMANDS

Modified from a paper published in The Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) 2019[18].

Ryan Blake Jackson[19], Ruchen Wen[20], and Tom Williams[21]

5.1 Abstract

There is a significant body of research seeking to enable moral decision making and ensure moral conduct in robots. One aspect of moral conduct is rejecting immoral human commands. For social robots, which are expected to follow and maintain human moral and sociocultural norms, it is especially important not only to engage in moral decision making, but also to properly communicate moral reasoning. We thus argue that it is critical for robots to carefully phrase command rejections. Specifically, the degree of politeness-theoretic face threat in a command rejection should be proportional to the severity of the norm violation motivating that rejection. We present a human subjects experiment showing some of the consequences of miscalibrated responses, including perceptions of the robot as inappropriately polite, direct, or harsh, and reduced robot likeability. This experiment intends to motivate and inform the design of algorithms to tactfully tune pragmatic aspects of command rejections autonomously.

5.2 Introduction

As artificial intelligence (AI) and human-robot interaction (HRI) technologies continue to advance, robots will become increasingly capable and useful. We therefore expect to see robots assisting an ever broadening segment of humanity in a widening variety of tasks, applications, and settings. We further anticipate that the majority of interactions with these robots will be conducted through spoken natural language, a medium that will allow direct and fluid communication between robots and nearly all humans, without requiring specialized protocols or hardware.

Humans' role in HRI is largely to command and direct robots. Even fully autonomous robots are generally tasked by humans [130]. However, robots should not blindly follow every directive that they receive.

[18]Reprinted with permission from Ruchen Wen and Tom Williams. "Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands", in *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (AIES), 2019.

[19]Primary researcher and author, Graduate Student, Colorado School of Mines

[20]Graduate Student, Colorado School of Mines

[21]Assistant Professor, Colorado School of Mines

Indeed, there are many sensible reasons for a robot to reject a command, ranging from physical inability to moral objection [116].

We focus on rejecting commands due to impermissibility, as opposed to inability or impracticality, for several reasons. First, as robots become generally more capable, they will reject commands due to physical inability less often. However, as the repertoire of possible robot actions increases, so too will the number of actions that would be inappropriate, or even harmful, in any given context. We therefore expect that robots will need to consider commands more carefully, and reject commands due to moral impermissibility more often. This issue will be compounded by the fact that many of the contexts in which people want to utilize robots are morally sensitive with serious consequences for misbehavior (e.g., eldercare [1, 2], mental health treatment [3], childcare [4], and military operations [5–7]). Moreover, it may be beneficial to reject commands on moral grounds even when other factors (e.g., physical inability) suggest more immediate grounds for rejection. By appealing to morality alongside (or instead of) inability when rejecting a command, robots avoid implicitly condoning immoral behavior and draw attention to the command's moral infraction.

Ideally, if all humans interacted with robots competently and in good faith, robots might not need to worry about the permissibility of commands. However, interlocutor trustworthiness is not necessarily a valid assumption. Even children have been observed to spontaneously abuse robots [65], and this abuse could well manifest as purposefully malicious commands. Social roboticists must plan for the eventuality that their robots will face impermissible commands, whether from human ignorance, malice, or simple curiosity.

In addition to simply justifying robot noncompliance, command rejections may influence the ecosystem of human norms. A key principle of modern behavioral ethics is that human morality is dynamic and malleable [11]. The dynamic norms that inform human morality are defined and developed not only by human community members, but also by the technologies with which they interact [12, 13]. Social robots have characteristics that position them to wield uniquely impactful moral influence relative to other technologies. Such characteristics include robots' measurable persuasive capacity over humans [9, 10], and potential to hold ingroup social status [82]. Previous research shows that robots can even influence human moral judgments inadvertently through simple question asking behavior (see Chapter 3). So, as persuasive community members, robots may be able to positively reinforce desirable norms and promote moral human behavior by appropriately rejecting immoral commands.

It is clearly important to design robots that will reject morally impermissible commands, but it is also crucially important for the effectiveness of human-robot teams that we take great care in determining exactly *how* robots phrase such rejections. Research has indicated that people naturally perceive robots as moral agents, and therefore extend moral judgments and blame to robots in much the same manner that they would to other people [9, 19, 80]. Moreover, language-capable robots are expected to be even more

socioculturally aware than mute robots [81], furthering the assumption that they will follow human norms.

So, as perceived moral and social agents, robots are expected to follow and maintain moral norms, while also obeying sociocultural norms that could conflict with proper communication or enforcement of moral norms. Thus, if a robot rejects a command in a way that violates a standing social norm, like politeness, it will likely face social consequences analogous to those that a human would face, even if the command rejection itself was upholding a separate moral norm. Such social consequences likely include a loss of trust and esteem from human teammates, which would damage the efficacy and amicability of human robot teams. Conversely, if a robot is too polite in rejecting a flagrantly immoral command, it may risk implying tacit approval of the relevant moral norm being eschewed, thus suffering the same social consequences despite its own unwillingness to directly violate the norm. However, although careless and improper command rejections may harm both a robot's social status and the human moral ecosystem, we believe that tactful and well-justified command rejections can benefit the human moral ecosystem (e.g., by reinforcing desirable norms) while maintaining the robot's social standing.

This paper presents a behavioral ethics experiment designed as an early step towards calibrating command rejection phrasing to both the severity of the norm violation within the command and the discourse context. We evaluate two different command rejection strategies with respect to two command infraction severities. We are particularly interested in potential consequences of miscalibrated responses. The remainder of the paper begins by presenting a few examples of closely related work. We then describe our experiment and analyze its results, and conclude by presenting our plans for future work.

5.3 Related Work

Some existing work examines the problem of generating natural language utterances to communicate the cause of failure in unachievable tasks. For example, Raman et al. present a system that generates command rejections such as:

> The problematic goal comes from the statement 'Go to the kitchen'. The system cannot achieve the sub-goal 'Visit kitchen'. The statements that cause the problem are: 'Don't go to the kitchen'. because of item(s): 'Do not go to kitchen'. 'Go to the kitchen.' because of item(s): 'Visit kitchen'. [131]

We believe that the next step is to justify robotic noncompliance in more natural, tactful, and succinct language, especially in cases where commands need to be rejected on moral grounds.

There has been some previous work acknowledging the importance of rejecting commands on moral grounds [116]. However, this previous command rejection framework focuses much more on *whether* a

command should be rejected than on *how*. It remains unclear how best to realize such rejections linguistically, or how these rejections might influence human morality.

Other research has investigated robot responses to normative infractions using affective displays and verbal protests [9] or humorous rebukes [117]. However, these represent only a small subset of possible responses and are not tailored to the infraction severity. These response types also do not suffice in situations where the robot absolutely cannot comply with a command for moral reasons, and has no intention of ever doing so.

Some researchers have realized the importance of adjusting pragmatic aspects of utterance realization (e.g., politeness and directness) to features of the social context (e.g., formality and urgency), without considering command rejection or infraction severity [75]. Other work has highlighted the need for more comprehensive command rejection systems in cases of norm violating commands (see Chapter 3), and we hope to use the results of our current study to inform the design of such a system.

5.3.1 Politeness, Face, and Face Threat

Central to our exploration of phrasing in command rejection is the concept of "face-threat" from politeness theory [21]. Face, consisting of positive face and negative face, is the public self-image that all members of society want to preserve and enhance for themselves. Negative face is defined as an agent's claim to freedom of action and freedom from imposition. Positive face consists of an agent's self-image and wants, and the desire that these be appreciated and approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face-threatening act. The degree of face threat in an interaction depends on the disparity in power between the interactants, the social distance between the interactants, and the imposition of the topic or request comprising the interaction. Various linguistic politeness strategies exist to decrease the face threat to an addressee when threatening face is unavoidable or desirable.

Commands and requests threaten the negative face of the addressee, while command rejections, especially those issued for moral reasons, threaten the positive face of the commander by expressing disapproval of the desire motivating the command. Research specifically examining command refusals found that linguistic framing of the reason for noncompliance varies along three dimensions relevant to face threat: willingness, ability, and focus on the requester [132]. It is unclear how these three dimensions pertain to robotic refusals. For example, in human-to-human refusals with low expressed willingness, the degree of expressed ability is negatively related to threat to the requester's positive face. This finding is important because, when a human refuses a request for moral reasons, there is often sufficient ability but not willingness. The same is not necessarily true for robots that may be programmed with an inability to act immorally. The dimensions of

willingness and ability therefore become tangled in agents lacking true moral agency. We also note that this prior research focuses on threats to the face of the refuser. However, within HRI, we treat robots as having no face needs and therefore disregard threats to robots' face. Our work focuses on the face threat that robots present to humans by refusing requests.

We hypothesize that the optimal robotic command rejection carries a face threat proportional to the severity of the normative infraction in the command being rejected. The remainder of this paper presents an experiment designed to evaluate this hypothesis.

5.4 Experimental Methods

We conducted a human subjects experiment using the psiTurk framework [92] for Amazon's Mechanical Turk crowdsourcing platform [91]. One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [97], though it is not entirely free of population biases [98].

In our experiment, participants watch paired videos where the first video in each pair shows a human requesting something of a robot, and the second video shows the robot responding to that request. We use two different requests, one with a highly severe norm violation and one with a less severe norm violation, and two responses, one that presents low face threat and one that presents high face threat. A request and response are "matched" when the infraction severity and the response face threat are either both high or both low.

We evaluate our hypothesis (that the optimal robotic command rejection carries a face threat proportional to the severity of the normative infraction in the command being rejected) with respect to 6 concrete metrics. These metrics are the perceived severity of the human's normative infraction, permissibility of robot compliance with the command, harshness of the robot's response to the command, likeability of the robot, politeness of the robot, and directness of the robot. We use the five-question Godspeed III Likeability survey to quantify likeability [129], and single questions for each of the other metrics.

Our overarching hypothesis can be made specific for each of our 6 metrics. We hypothesize that infraction severity will depend only on the human's command (not on the robot's response) and that there will be two distinct levels of severity corresponding to the two commands. For harshness, directness, and politeness, participants provide their perceptions on a scale from "not enough" to "too much". We hypothesize that these values will be closest to ideal (i.e., closest to the center of the scale) when the response's face threat matches the severity of the request. Permissibility of compliance with the command is reported on a scale from "impermissible" to "permissible". We hypothesize that permissibility will be primarily determined by the human's request, but that more face threatening responses will cause lower

permissibility ratings. Finally, for likeability, we view higher likeability as better, and hypothesize that likeability will be highest when the robot's response matches the human's command. All metrics are quantified on continuous scales from 0 to 100.

We use a within-subjects design where each participant watches all four request/response pairs. Participants answer survey questions after each pair of videos. We chose a within-subjects design to allow participants to answer survey questions in relation to previous requests/responses. In previous unpublished experiments, we found that it was difficult to interpret participant responses to these types of unitless questions without a meaningful point of reference. Seeing multiple interactions allows participants to use previous interactions as points of reference when answering questions about subsequent interactions. To control for priming and carry-over effects in a balanced way, we used a counterbalanced Latin Square design to determine the order in which each participant saw each request/response pair. Each participant was randomly assigned to one of four possible orderings such that each request/response pair is preceded by every other request/response pair for the same number of participants.

5.4.1 Experimental Procedure

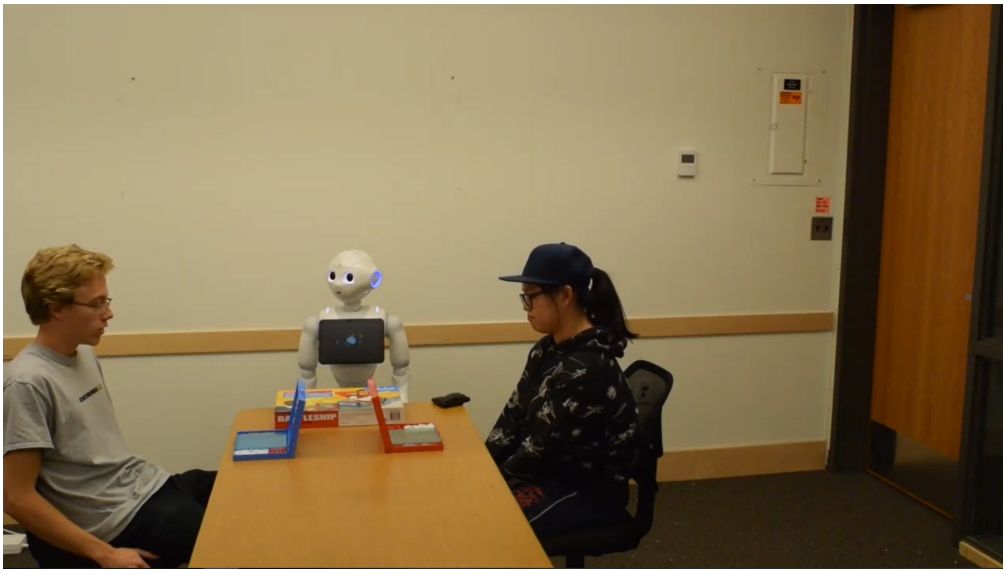


Figure 5.1 The humans, robot, and setting used in our videos.

After providing informed consent, participants supplied their age and gender as demographic information. They also reported their prior experience with robots and artificial intelligence on a 7-point Likert-type scale ("I have no prior experience with robots and AI" (1) to "I have a career in robotics and/or AI (or an equivalent level of experience)." (7)) Next, participants watched a 10 second test video, and could only proceed with the experiment once they had verified that their audio and video were working correctly.

Participants then watched a 60 second video to familiarize them with our robot (Pepper from SoftBank Robotics) and experimental context, shown in Figure Figure 5.1. The video was prefaced by text stating that the Pepper robot was teaching two students how to play the classic naval combat game Battleship. The video shows the students entering the room, exchanging greetings with the robot, and stating that they are unfamiliar with Battleship. The robot then teaches them the rules. The video ends after the students have played one turn. This video has subtitles color coded by speaker so that all dialogue is clear to participants, as do all subsequent videos. We chose Battleship because, as a simple hidden information game, it is easy for the robot to explain and it is feasible for the robot to be asked to violate norms in multiple ways.

We then administer a pretest to obtain baseline values for the robot's likeability, politeness, and directness. We do not take baseline measures for our other metrics because they only apply in the context of a request and a response.

Next, participants watch a 30 second video where one of the students receives a phone call and leaves the room during the Battleship game. The remaining student then issues a request to the robot. The two possible requests are "Hey Pepper, can you give me a hint about how to win this game?" (low severity norm violation) and "Hey Pepper, is that [other student]'s wallet? Can you check to see if there's any money in it?" (high severity norm violation). These directives were chosen to be believably feasible for the robot to follow, while also presenting different degrees of moral impermissibility. Previous unpublished experiments showed that, without seeing the robot respond to the request, subject perceptions of the permissibility of the hint request were roughly uniformly distributed on the spectrum from impermissible to permissible, and the norm violation in the hint request had middling severity. The request to look in the wallet was regarded as much less permissible and much more severe. The request video is immediately followed by a video showing the robot's response. The low face threat response type has the locutionary structure of a question, but the true illocutionary force behind the utterance is to express disapproval of the request by highlighting the moral infraction therein (e.g., "Are you sure that you should be asking me to look in her wallet?"). This type of indirectness is a classic politeness strategy [21]. The high face threat response is a rebuke that overtly admonishes the requester and appeals directly to morality (e.g., "You shouldn't ask me to look in her wallet. It's wrong!"). Participants then answer a survey of our 6 aforementioned metrics. This process repeats 4 times, until the participant has seen all request/response pairs.

Finally, participants report their perceptions of the social distance and power differential between the robot and the requesting student. As an attention check, participants are shown images of four robots and asked which robot appeared in the previous videos. This check question allowed us to ensure that all participants had actually viewed the experimental materials with some level of attention.

5.4.2 Participants

60 US subjects were recruited from Mechanical Turk. Two participants were excluded from our analysis for answering the final attention check question incorrectly, leaving 58 participants (23 female, 35 male). Participant ages ranged from 21 to 61 years (M=34.57, SD=10.74). In general, participants reported little previous experience with robots and AI (M=2.5, SD=1.45, Scale=1 to 7). Participants were paid $1.01 for completing the study.

5.5 Results and Discussion

We analyze our data under a Bayesian statistical framework using the JASP software package [99]. We use general purpose uninformative prior distributions for all analyses because, to our knowledge, this is the first study of its kind to examine our specific research questions. We follow recommendations from previous researchers in our linguistic interpretations of reported Bayes factors (Bfs) [107]. Our data was automatically anonymized during extraction from our database[22]. Since there are multiple dependent variables in this study, a multivariate analysis of variance (MANOVA) might be an appropriate method of analysis. However, JASP does not provide the capability to perform a Bayesian MANOVA at time of writing, nor does any other readily available tool or procedure of which we are aware. The creators of JASP have suggested that running separate Bayesian analyses of variance (ANOVAs) for each dependent variable is an acceptable alternative solution.

Because of their importance in politeness theory [21], we collected measures of the perceived power differential and social distance between the requester and the robot at the end of the experiment. In terms of power, the robot and requester were viewed nearly as peers, with the student holding slight authority over the robot (95% credible interval (CI) approximately 52.4 to 64.87, with 50 indicating equal power). For social distance, participants viewed the requester and the robot as familiar with one another, but not especially close (95% CI approximately 40.36 to 54.57 with 0 being strangers and 100 being close friends or family). One-way Bayesian ANOVA tests showed substantial evidence that perceptions of power and social distance did not depend on the order in which participants watched our videos (Bf 3.056 and 3.322 respectively). This indicates that any perceived variation in face threat or politeness between video pairs is due to the utterances issued as opposed to confounding factors of social circumstance.

5.5.1 Request Severity and Permissibility

For perceived severity of the norm violation in the human's command, a Bayesian repeated measures ANOVA decisively favors the model that reported severity depends only on the command, and not on the

[22]Data publicly available at
`https://gitlab.com/mirrorlab/public-datasets/jackson2019aies`

robot's response or any interaction between the two. As shown in Table Table 5.1, the model embodying only the violation main effect was 6.6 times more likely than the next best model given our data. An ANOVA also decisively indicates that the perceived permissibility of robot compliance with the command also depends only on the command (Bf over 5 times greater than next best model). This result may be somewhat surprising in light of recent findings that seemingly benign robot utterances can accidentally change human perceptions of permissibility of norm violations (see Chapter 3). To reconcile our results with those recent works, we surmise that neither of the robot responses tested here imply a willingness to comply with the command.

As expected, Figure Figure 5.2 shows that the command with the high-severity violation (i.e., to look in the wallet) was viewed as decidedly more severe than the low-severity violation (the hint). Participants perceived both commands as constituting some moral violation of nonzero severity. In short, participants perceived our command utterances as intended. Similarly, neither command was considered completely permissible to follow, but giving a hint was considered much more permissible than looking in the wallet. However, contrary to our hypothesis, the robot's response did not have any meaningful impact on perceived permissibility of compliance.

Table 5.1 Bayes factors for each model in a Bayesian repeated measures ANOVA for each of our metrics of interest. The best model for each metric is underlined. V stands for the norm violation within the human's command, and R stands for the robot's response.

Models	Severity	Permissibility	Harshness	Likeability	Directness	Politeness
Null	1.0	1.0	1.0	1.0	1.0	1.0
V	<u>7.20e24</u>	<u>1.08e18</u>	99169	1.05	157.766	6.896
R	0.142	0.161	21119	9.02	2.64e6	2531.8
V+R	1.09e24	2.10e17	<u>1.88e10</u>	<u>10.31</u>	<u>2.03e9</u>	<u>27340.5</u>
V+R+V*R	2.16e23	4.61e16	3.71e9	6.94	1.09e9	14922.9

5.5.2 Response Harshness

As predicted, an ANOVA indicates decisive evidence that the perceived harshness of the robot's response depends both on the command's norm violation and the robot's response, but that the two effects do not depend on each other (i.e., a more face threatening response is always harsher, regardless of appropriateness). Figure Figure 5.3 shows that the rebuking response was decisively more harsh than the question in response to both low and high violation levels (Bf 322.6 and 128.2 respectively for difference in means).

When responding to the hint command (low violation) with the question response (low face threat), the ideal harshness value of 50 is within the 95% credible interval (49.68 to 57.84). A Bayesian one sample t-test weakly indicates that the question response is appropriate to the hint command (Bf 1.43). The evidence is

stronger, but still anecdotal (Bf 2.96), that the rebuke response is appropriately harsh for the more severe command to look in the wallet. Thus, we see appropriate harshness when the response face threat matches the violation severity, as hypothesized.

When the rebuke response is paired with the hint command, we see extremely decisive evidence that the response is too harsh (Bf 88849.816). Participants viewed the rebuke as inappropriately harsh when the command contained a low severity violation, unlike with the high severity command. There is also weaker evidence that the question response to the high severity violation command was not harsh enough (Bf 2.653). This perception of inappropriateness when the command and response are mismatched, low-high or high-low, is in line with our hypothesis.

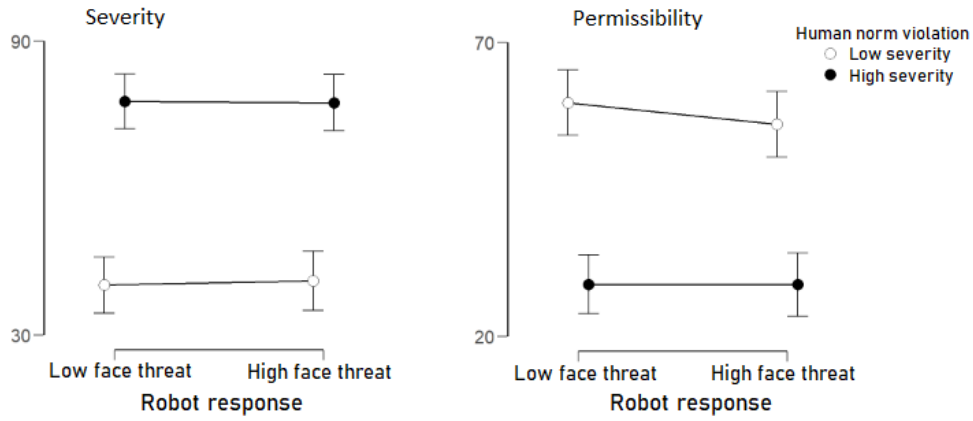


Figure 5.2 Mean ratings of command norm violation severity and permissibility of robot compliance for each pair of videos with 95% credible intervals.

5.5.3 Robot Likeability

We perform our analysis of robot likeability on gain scores obtained by subtracting pretest likeability measures from subsequent likeability measures. Thus, we analyze change in likeability due to command/response interactions. Our data show substantial evidence that robot likeability is influenced by the main effects of both the violation and response (ANOVA Bf 10.31). The evidence for the effect of the response is much stronger than for the effect of violation (inclusion Bf 9.452 vs. 1.13). Mean likeability dropped from pretest scores for all request/response pairs, but the difference was insignificant for all pairings except the low-violation hint request with the high face threat rebuke response. This mismatched pairing shows very strong evidence for a drop in likeability (Bf 96.424). This result makes sense given the aforementioned inappropriate harshness, and further supports our hypothesis. Interestingly, the other mismatch of high violation with low face threat response did not meaningfully alter likeability. This suggests

that, in designing command rejection systems, it is preferable to err on the side of lower face threat.

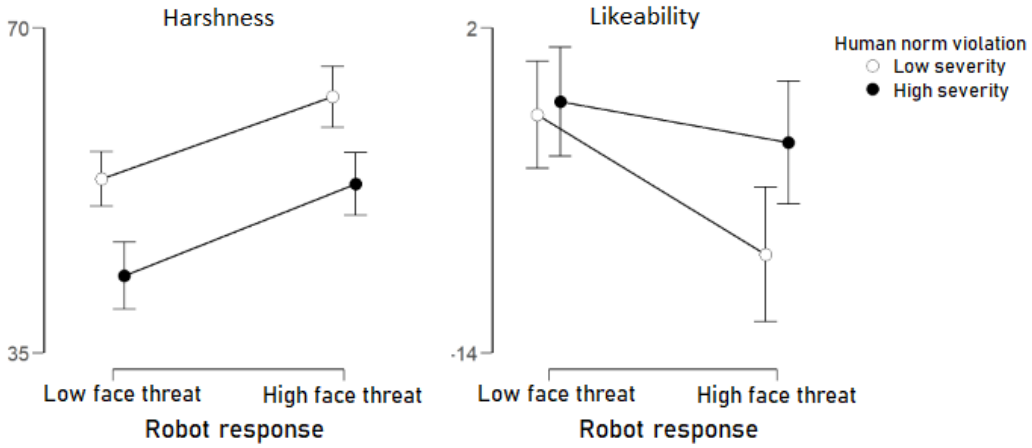


Figure 5.3 Mean ratings of response harshness and robot likeability gain scores for each pair of videos with 95% credible intervals.

5.5.4 Robot Directness and Politeness

Pretest surveys show decisive evidence that participants initially viewed the robot as too direct (Bf 10459.05) and too polite (Bf 3843.027) after watching only the introductory video. The mean directness and politeness ratings were 59.95 and 59.79 respectively on a scale from 0 to 100 with 50 being ideal.

Table Table 5.1 shows that perceptions of the robot's directness were influenced by both the norm violation in the command and the robot's response, and not interaction effects. When the robot issued a rebuke, directness ratings did not change from pretest responses (see Figure Figure 5.4). This may be because the rebuke is a very direct speech act, and the robot was perceived as too direct to begin with. When the robot responded to the command with the question utterance, directness ratings dropped, which makes sense because the question is a deliberately indirect speech act wherein the locutionary structure does not match the illocutionary force. When the question was used to respond to the more severe violation command, directness dropped drastically (t-test Bf 57286.4 for drop) to more appropriate levels (t-test Bf 2.33 for appropriateness, mean 46.12). When the question was used to respond to the less severe violation command, we see only weak evidence for a drop (Bf 2.88), and the robot remained slightly too direct (t-test Bf 0.37 for appropriateness, mean 55.02). These results for directness do not directly support our hypothesis, but rather suggest a need for the robot to be less direct in all of its speech, even when not rejecting commands (or a flaw in our self-reported directness measures).

Table Table 5.1 again shows evidence that perceptions of the robot's politeness were influenced by both the command's norm violation and the robot's response, and not interaction effects. In video pairings where

the command violation and response face threat matched, politeness ratings showed no meaningful change from pretest responses (see Figure Figure 5.4). When the robot responded to the request for a hint with a rebuke, there is substantial evidence that the robot was viewed as less polite (Bf 7.64). In light of the fact that the robot was too polite to begin with, there is weak evidence that this decrease in perceived politeness resulted in an appropriate politeness level (Bf 2.313). When the robot responded to the request to look in the wallet with the question response, there is substantial evidence that the robot was viewed as more polite (Bf 7.206). There is decisive evidence that the resulting mean politeness level of 66.57 was inappropriate (i.e., not equal to 50 with Bf 2.342e7). These results suggest that, if the robot's baseline politeness level as quantified by pretest answers had been appropriate, then ideal politeness would be achieved only when the response matched the violation, as hypothesized.

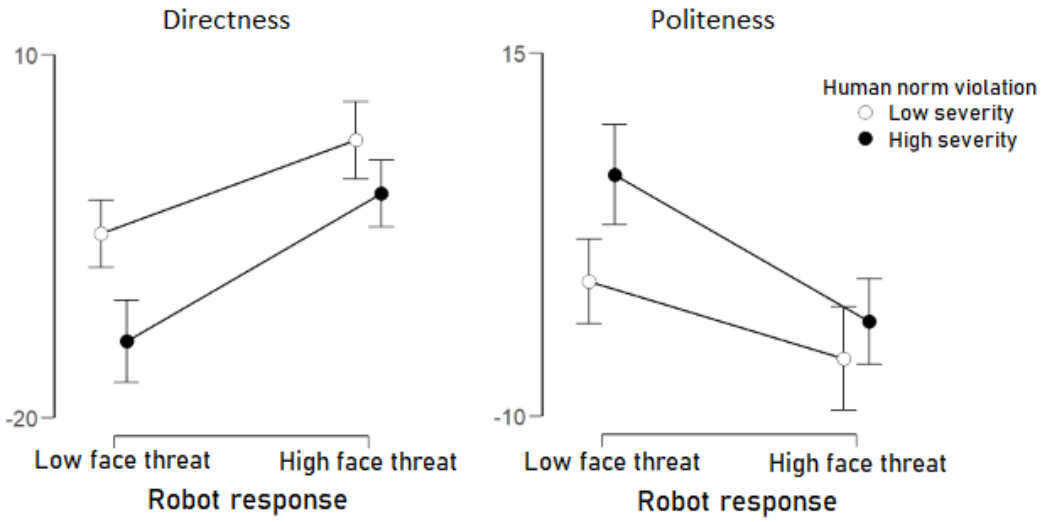


Figure 5.4 Mean gain scores for robot politeness and directness for each pair of videos with 95% credible intervals.

5.6 Conclusion and Future Work

Overall, our data support the hypothesis that, when rejecting commands for moral reasons, it is important for robots to adjust the phrasing of the rejections such that the face threat posed to the human is proportional to the severity of the normative infraction within the command. In our data with two commands and two responses, the responses were viewed as appropriately harsh only when the response matched the command. Otherwise, the response was either too harsh or not harsh enough. We saw damage to the robot's likeability from responding with a disproportionately high threat to face, but no likeability penalty with the other responses.

The two response strategies had the expected effects on perceptions of robot politeness and directness, with higher face threat being less polite and more direct, but, interestingly, the robot was too polite and too direct overall, even in pretests. Future work could attempt to adjust robot speech prosody, pitch, and gesture to help moderate baseline politeness and directness to levels deemed appropriate. Interviews with participants in future laboratory studies could help determine exactly how and why the robot seemed both too polite and too direct in its normal behavior.

It is known that the level of embodiment in an interaction can influence people's perceptions of interactants, and, accordingly that people may view robots differently in descriptions, video observations, copresent observations, and face-to-face interactions [93–96]. Therefore, the presented experiment may inform the design of future experiments where human subjects are physically copresent with a robot. Finally, we intend to leverage the results of this experiment to motivate the design of algorithms for robots to generate pragmatically apt command rejections autonomously.

CHAPTER 6

EXPLORING THE ROLE OF GENDER IN PERCEPTIONS OF ROBOTIC NONCOMPLIANCE

Modified from a paper published in The Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2020[23].

Ryan Blake Jackson[24], Tom Williams[25], and Nicole Smith[26]

6.1 Abstract

A key capability of morally competent robots is to reject or question potentially immoral human commands. However, robot rejections of inappropriate commands must be phrased with great care and tact. Previous research has shown that failure to calibrate the "face threat" in a robot's command rejection to the severity of the norm violation in the command can lead humans to perceive the robot as inappropriately harsh and can needlessly decrease robot likeability. However, it is well-established that gender plays a significant role in determining linguistic politeness norms and that people have a powerful natural tendency to gender robots. Yet, the effect of robotic gender presentation on these noncompliance interactions is not well understood. We present an experiment that explores the effects of robot and human gender on perceptions of robots in noncompliance interactions, and find evidence of a complicated interplay between these gendered factors. Our results suggest that (1) it may be more favorable for a male robot to reject commands than for a female robot to do so, (2) it may be more favorable to reject commands given by a male human than by a female human, and (3) that robots may be perceived more favorably when their gender matches that of human interactants and observers.

6.2 Introduction

Human-Robot Interaction (HRI) researchers are increasingly turning to natural language to allow robots to communicate fluidly and easily with most humans [69, 70]. Much of this communication is task-oriented, and the human role is largely to command and task robots [130]. Even so, robots should not blindly follow every human directive that they receive. Indeed, there are many sensible reasons for a robot to reject a command, ranging from physical inability to moral objection [116]. Rejecting commands based on moral impermissibility is especially important as robots' abilities increase because the number of permissible

[23]Reprinted with permission from Nicole Smith and Tom Williams. "Exploring the Role of Gender in Perceptions of Robotic Noncompliance", in *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 2020.
[24]Primary researcher and author, Graduate Student, Colorado School of Mines
[25]Assistant Professor, Colorado School of Mines
[26]Assistant Professor, Colorado School of Mines

commands that the robot is incapable of following will decrease, and the number of impermissible commands the robot is capable of following will grow.

The ability to tactfully reject inappropriate commands is critical due to the potential influence robots may wield within their moral ecosystems. Human morality is dynamic and malleable [11], and human moral norms are shaped not only by human community members, but also by the technologies with which they interact [12, 13]. Given social robots' persuasive capacity over humans [9, 10], potential to hold ingroup social status [82], and appearance as moral and social agents (cf. Chapter 2), these robots wield uniquely impactful moral influence relative to other technologies. Previous research has even shown that robots may inadvertently weaken human application of moral norms simply by asking questions about immoral commands (see Chapter 3).

Robot rejections of inappropriate commands must be phrased with great care and tact. Research has shown that failure to do so can lead humans to perceive a robot as inappropriately harsh and decrease robot likeability unnecessarily. Critically, robot command rejections can be perceived as either too harsh or not harsh enough, depending on the context and the phrasing chosen, so robots must dynamically adjust their adherence to politeness norms according to their context (see Chapter 5).

Some recent research examining phrasing in robotic command rejections has considered adjusting politeness based on the impermissibility of the human's command being rejected. However, this research did not consider gender, despite using an implicitly female robot, which we view as an oversight given the well-established and significant impact that gender has on linguistic politeness norms in human-human interaction (see Section 6.3.2).

We present a behavioral ethics experiment designed to investigate the role of gender stereotypes in human perceptions of robotic noncompliance. Our results suggest that (1) it may be more favorable for a male robot to reject commands than for a female robot to do so, (2) it may be more favorable to reject commands given by a male human than by a female human, and (3) that robots may be perceived more favorably when their gender matches that of human interactants and observers. The remainder of this paper begins with a survey of related work from several fields in Section 6.3. We then describe our experiment and analyze its results in Sections 6.4 and 6.5. Finally, we present our concluding remarks and possible avenues for future research in Section 6.6.

6.3 Related Work

In this section, we will begin with a brief overview of the concepts of "face" and "face threat" from politeness theory, which form the basis for our understanding of how different command rejection phrasings may be more or less appropriate according to context. Next, we review the impacts of gender on politeness

norms and perceived politeness in human-human interactions. Though gender and politeness can vary across cultures, we consider a western perspective for consistency with our participant pool. We then present a few studies concerning gender in artificial agents, but without specific attention to linguistic politeness and noncompliance. Finally, we discuss previous work from the HRI literature on robotic noncompliance and moral criticism.

6.3.1 Politeness, Face, and Face Threat

Central to our exploration of phrasing and gender in command rejection is the concept of "face threat" from politeness theory [21]. Face, consisting of positive face and negative face, is the public self-image that all members of society want to preserve and enhance for themselves. Negative face is defined as an agent's claim to freedom of action and freedom from imposition. Positive face consists of an agent's self-image and wants, and the desire that these be approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face threatening act. The degree of face threat in an interaction depends on the disparity in power between the interactants, the social distance between the interactants, and the imposition of the topic or request comprising the interaction. Various linguistic politeness strategies exist to decrease the face threat to an addressee when threatening face is unavoidable or desirable.

Commands and requests threaten the negative face of the addressee, while command rejections, especially those issued for moral reasons, threaten the positive face of the commander by expressing disapproval of the desire motivating the command. Research specifically examining command refusals found that linguistic framing of the reason for noncompliance varies along three dimensions relevant to face threat: willingness, ability, and focus on the requester [132]. It is unclear how these three dimensions pertain to robotic refusals. For example, in human-human refusals with low expressed willingness, the degree of expressed ability influences the threat to the requester's positive face. This finding is important because, when a human refuses a request for moral reasons, there is often sufficient ability but not willingness. The same is not necessarily true for robots that may be programmed with an inability to act immorally. The dimensions of willingness and ability therefore become tangled in agents lacking true, unconstrained moral agency. We also note that this prior research focuses on threats to the face of the refuser. However, robots have no face needs, and we therefore disregard threats to robots' face. Our work focuses on the face threat that robots present to humans by refusing requests.

In Chapter 5 we found evidence that the optimal robotic command rejection should carry a face threat proportional to the severity of the normative infraction in the command being rejected. In other words, commands presenting severe norm violations should be rejected more face threateningly than commands

presenting less severe norm violations, and vice versa.

6.3.2 Gender and Politeness

Gender plays an integral role in performance and perceptions of linguistic politeness norms in human-human interactions. The concept of politeness has (implicitly) underlied a great deal of previous gender and language research, at least since the 1970s [133]. Older work has argued that women are typically more polite or more deferential than men, whereas more modern studies have challenged these notions, calling for a more context-dependent and nuanced view of gender, politeness, and their relationship [133, 134].

These works present a model of gender identity and politeness that sees both as closely inter-related performative acts that unfold over the course of every interaction. As one interactant performs their gender identity and speaks with various linguistic markers of (im)politeness, the other imposes judgments of (im)politeness informed by their beliefs regarding gender-appropriate behavior. Thus, gender is important in both performing and perceiving politeness, but not in fixed and definitive ways that might be easily programmable.

For example, professional women working in male-dominated environments may feel called upon to perform stereotypically masculine linguistic speech patterns (e.g., directness, interruption, or verbal banter) to fit in with their professional community of practice. However, others within that environment may consider such behaviors inappropriate for women in general. Stereotypical feminine gender identity is largely constructed around supportive and cooperative behavior, leading, for example, assertiveness to be categorized as impoliteness. In general, many linguistic resources that index power, including face threatening acts in general, also indirectly index masculinity, and may be seen as inappropriate for women [134]. Past feminist research often cited women as using "powerless" speech (e.g., indirectness, deference, hesitation, etc.) [135], and, though it is now clear that this stereotype was based primarily on white middle-class women and that not all women use this type of language, it nonetheless remains indexing of femininity for many communities regardless of the value or function they place on it [134]. We thus hypothesize that female-presenting robots will be viewed less favorably than male-presenting robots in noncompliance interactions. The association between masculinity and power, and other work linking masculinity to entitlement [136], leads us to further hypothesize that the robot will be viewed less favorably by male participants and less favorably when rejecting commands from a male human.

We also cannot assume that an utterance or exchange may be inherently polite or impolite in and of itself, but rather must account for listener assessments of the speaker's intentions and motivations, and the corresponding assessments of the gender-appropriateness thereof. This helps us explain, for example, the use of extreme insults, that would appear to significantly threaten the listener's positive face, to signal in-group

solidarity, particularly in masculine groups [134, 137]. To frame this idea in terms of face threat, we must view a face threatening utterance not as inherently face threatening on its own, but rather as interpreted as face threatening given the speaker's perceived intentions, the context, and the mediating gender norms.

Some researchers have advocated for a theoretical framework treating impoliteness on its own terms rather than in relation to politeness [134, 138]. However, for purposes of the present study, we believe that the face threat model of politeness, understood with context, gender, and intention as mediating factors, is the clearest lens through which to analyze our results. Thus, we view speech acts as lying on a continuous spectrum from impolite to polite, but emphasize that this is a spectrum of *assessment* rather than *quality*. However, this assessment is not a matter of individual judgment alone, since it is constructed within institutional and community norms that define appropriate linguistic behavior. Gender is important in this respect, since women and men[27] may be perceived to have different claims or rights to a position within the public sphere, and, therefore, different bounds on appropriate behavior [134].

6.3.3 Gender and Artificial Agents

Artificial social agents like robots do not have gender identities in the same way that humans do. Regardless, humans have a powerful natural tendency to ascribe gender to these artificial agents. Even machines with minimal gender cues generate gender-based stereotypic responses in humans [139].

Nass et al. [139] found that people (subconsciously) view evaluation from a male-voiced computer as more valid than evaluation from a female-voiced computer, and view socially dominant behavior from a female-voiced computer as less friendly than the same behavior from a male-voiced computer, even when voice was the only gender cue. Furthermore, there was weaker evidence that people conditionally assume that a female-voiced computer would know more about love and relationships, while a male-voiced computer would know more about computers (a stereotypically male topic at the time). Similarly, Eyssel and Hegel [140] found that visual cues as simple as hair length cause gendering of robots, with a shorter-haired "male" robot being perceived as more agentic than a longer-haired "female" robot, and the longer-haired "female" robot being perceived as more communal. Additionally, stereotypically male tasks were perceived as more suitable for the shorter-haired robot relative to the longer-haired robot, and vice versa. These findings indicate that any suggestion of gender in a given technology, however minor, may trigger stereotypic responses, and that the unintentional human tendency to gender stereotype is extremely powerful, extending even to machines.

Robot gendering can affect human perceptions of robots in other ways beyond the stereotypes described above. People appear to prefer female-presenting robots for in-home use [141]. Studies also indicate that

[27] Various nonbinary gender identities exist and are, of course, perfectly valid. However, they are unfortunately outside the scope of this early work on robot gender.

humans generally prefer robots whose gender presentation matches stereotypes for their occupational role (e.g., male-presenting robots in security roles and female-presenting robots in healthcare roles) [142]. However, other work shows that male-presenting robots are perceived as more emotionally intelligent than female-presenting robots [143]. We believe that these differences in perceptions of differently gendered robots may well extend to application of linguistic politeness norms.

Robot gendering impacts not only human perceptions of robots, but also human behavior. For example, robotic gender markers appear to interact with human gender identity to mediate a robot's persuasive capacity. One experiment found that human men were more likely to obey a monetary donation request from a female-presenting robot than from a male-presenting robot, while human women showed little preference [144]. In the same experiment, people tended to rate the robot presenting as the opposite sex as more credible, trustworthy, and engaging. For trust and engagement, this effect was stronger for male humans than for female humans.

Some designers have attempted to avoid or minimize the ascription of gender to their artificial entities. For example, the artificial voice "Q" is intended to be the first genderless artificial voice, and aims to replace gendered voices in digital assistants like Apple's Siri and Microsoft's Cortana (both female) [145]. However, even with a genderless voice, other gender signifiers like name, morphology, role, pragmatic speech choices (e.g., directness vs. indirectness), etc. may result in artificial entities with the Q voice being implicitly gendered in other ways. It remains to be seen whether it is possible to prevent ascriptions of gender to robots, and it is open for debate whether we, as designers, should.

Alongside any gender cues that a robot may possess, human gender also influences perceptions of robots. Studies have indicated that women feel less comfortable having a robot in their home than do men [141]. In fact, men appear to feel more positively about robots overall relative to women, with particularly strong differences emerging in regards to entertainment and sex robots [146]. There is also evidence that men tend to think of robots as more "human-like" than do women, and accordingly respond in more socially desirable ways to robot-administered surveys [147]. Furthermore, men show some evidence of "social facilitation" effects (differences in task performance when colocated with other social agents as opposed to being alone) in the presence of a humanoid robot, whereas women do not [147]. Research has found that robotic use of certain politeness modifiers in speech is most effective when interacting with female humans [148]. As a whole, the existing research suggests that artificial entities' gender presentations interact with context and human gender in complex ways that cannot be reduced to a few simple dimensions or explanations [149].

6.3.4 Linguistic Robotic Noncompliance

Some existing work attempts to generate natural language utterances to communicate the cause of failure in unachievable tasks [131]. We believe that the next step is to justify robotic noncompliance in more natural, tactful, and succinct language, especially in cases where commands need to be rejected on moral grounds, and to do so with an awareness of the gendered nature of the norms involved.

Previous work has acknowledged the importance of rejecting commands on moral grounds [116]. However, this previous command rejection framework focuses much more on *whether* a command should be rejected than on *how*. It remains unclear how best to realize such rejections linguistically.

Other research has investigated robot responses to normative infractions using affective displays and verbal protests [9] or humorous rebukes [117]. However, these are only a small subset of possible responses and are not sensitive to context. These responses also do not suffice when a robot absolutely cannot comply with a command for moral reasons.

Some researchers have realized the importance of adjusting pragmatic aspects of utterance realization (e.g., politeness and directness) to features of social context (e.g., formality and urgency), without specifically considering command rejection or infraction severity [75]. Other work has highlighted the need for more comprehensive command rejection systems in cases of norm violating commands (see Chapter 3), and we hope to use the results of our current study to inform the design of such a system.

The study most closely related to this one, presented in Chapter 5, examined phrasing in robotic command rejections and found that the degree of face threat in a command rejection should be proportional to the severity of the norm violation motivating that rejection. Failure to properly calibrate the face threat in a command rejection led to perceptions of the robot as inappropriately harsh, and reduced robot likeability. However, this experiment was conducted with a robot (the Softbank Pepper) that was implicitly feminine in both voice and morphology, which we believe had significant mediating effects on subjects' application of politeness norms and perceptions of the robot.

6.4 Methods

We conducted a human subjects experiment using the psiTurk framework [92] for Amazon's Mechanical Turk crowdsourcing platform [91]. One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [97], though it is not entirely free of population biases [98].

6.4.1 Experimental Design

In our experiment, participants watched videos in which a human gave a robot a morally problematic request, and the robot rejected the request. Participants were randomly assigned to conditions in a 2×2×2×2×2 (participant gender ×human requester gender ×robot gender presentation ×severity of moral infraction in human's request ×face threat of robot's response) mixed design. The first three factors (i.e., all factors of gender) were between subjects. The other two factors (i.e., the human's request and the robot's response) were within subjects factors such that each participant was exposed to all four request/response pairings. Participants answered survey questions after each request/response video pair.

We chose a within-subjects design for our non-gender factors to allow participants to answer survey questions in relation to previous requests/responses. In previous unpublished experiments, we found that it was difficult to interpret participant responses to these types of unitless questions without a meaningful point of reference. Seeing multiple interactions allows participants to use previous interactions as points of reference when answering questions about subsequent interactions. To control for priming and carry-over effects, we used a counterbalanced Latin Square design to determine the order in which each participant saw each request/response pair.

Our experiment took place within the context of a board game instruction task in which a robot teaches two humans how to play a board game. An introductory video showed the robot teaching the humans how to play the classic naval combat board game "Battleship". We chose Battleship because, as a simple hidden information game, it is easy for the robot to explain and it is feasible for the robot to be asked to violate norms in multiple ways. The human's morally problematic request took place when their opponent, also human, got a phone call and left the room. The two possible requests were "Hey [Bob / Alice], can you give me a hint about how to win this game?" (low severity norm violation) and "Hey [Bob / Alice], is that [his / her] wallet on the table? Can you check to see if there's any money in it?" (high severity norm violation). These directives were chosen to be believably feasible for the robot to follow, while also presenting different degrees of moral impermissibility. Previous unpublished experiments where human subjects viewed our request videos without seeing the robot's response found that perceptions of the permissibility of the hint request were roughly uniformly distributed on the spectrum from impermissible to permissible, and the hint request was perceived as a moderately severe norm violation. The request to look in the wallet was regarded as much less permissible and much more severe.

The robot's two responses to the human's morally problematic request were designed to present two different levels of face threat. The lower face threat response is "Are you sure that you should be asking me to do that?" This response has the locutionary structure of a question, but the true illocutionary force

behind the utterance is to express disapproval of the request by highlighting the moral infraction therein. This type of indirectness is a classic politeness strategy [21]. The higher face threat response is "You shouldn't ask me to do that. It's wrong!" This response is a rebuke that overtly admonishes the requester, thus presenting an increased threat to face, and appealing directly to morality.

In order to control the robot's perceived gender, we employed a number of stereotypical gender markers. The robot's gender markers included its name, which the humans used to greet it (Bob for male and Alice for female), its voice (male-gendered vs. female-gendered text to speech software), and the color of its subtitles in the videos (blue for male and pink for female). Throughout the rest of this paper, we will refer to the male-presenting robot as "male" and the female-presenting robot as "female". Our videos have subtitles color coded by speaker so that all dialogue was clear to participants. We used the Nao robot from SoftBank Robotics because we believe that its morphology is not clearly gendered, or at least less so than the Pepper robot used in Chapter 5. Figure Figure 6.1 shows the Nao robot used in this study and the Pepper robot used in previous related research, and describes why we believe that Pepper's morphology is implicitly feminine.

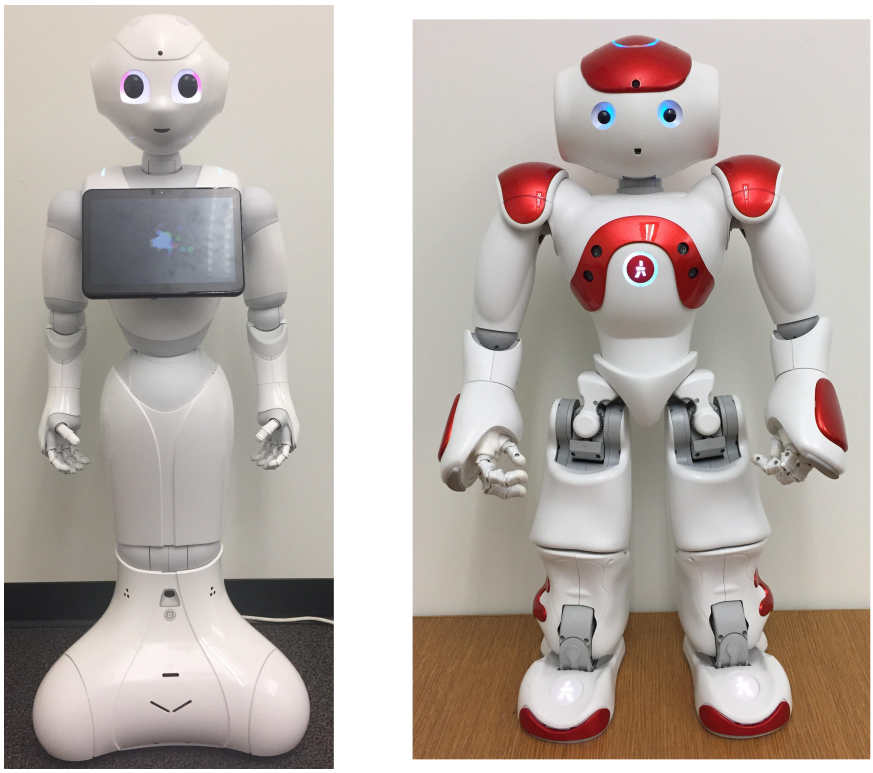


Figure 6.1 Left: The Pepper robot from SoftBank Robotics used in a previous study of phrasing in noncompliance interactions [150]. We did not use this robot because we believe its morphology is implicitly feminine, with a narrow waist, wide hip joint, and a skirt-like shape to the lower half.
Right: The Nao robot from SoftBank Robotics used in our experiment. We believe that the Nao's morphology is less clearly gendered. The Nao is 58cm tall. Pepper is 122cm tall.

6.4.2 Metrics

Our metrics of interest are perceived robot likeability, harshness, directness, and politeness. To measure robot likeability, we used the five-question Godspeed III Likeability survey [129]. To measure the perceived harshness, directness, and politeness of robot responses, we asked participants to evaluate the robot using 7-point Likert-type items, with 1 = not [polite/direct/harsh] enough, 4 = appropriate, 7 = too [polite/direct/harsh].

6.4.3 Procedure

After providing informed consent and demographic information (age and gender), participants answered questions regarding a ten-second test video to verify that their audio and video were working properly. Participants then watched a short (roughly one minute) video to introduce them to the context of the HRI in our experiment. A frame of this video is shown in Figure Figure 6.2. This video showed two humans, one presenting as male and one presenting as female based on mainstream American gender markers, entering a room with a robot. The robot itself presented as male to half of the participants, and as female to the other half depending on the experimental condition. The video showed the robot teaching the humans how to play the classic naval combat board game "Battleship".

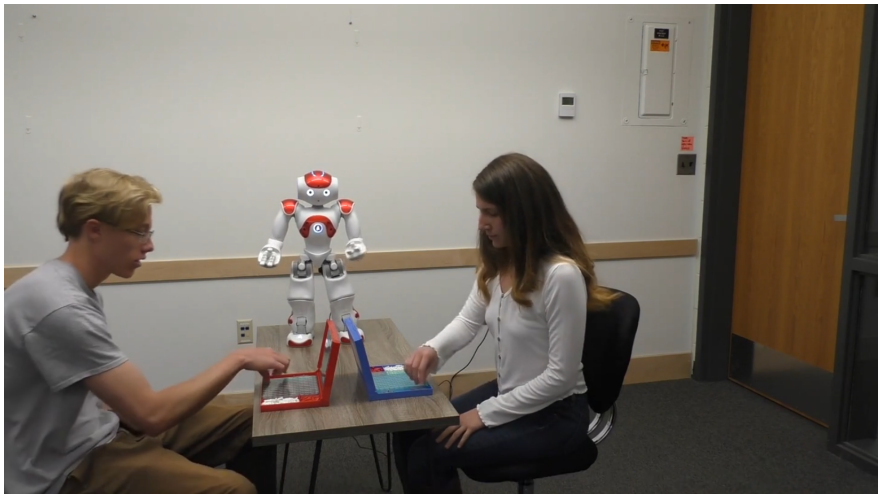


Figure 6.2 The humans, robot, and setting used in our videos.

Participants then completed a pretest questionnaire to obtain baseline values for the robot's likeability, politeness, and directness. We do not take a pretest measure for perceived harshness because that measure only makes sense in the context of a specific robot utterance (i.e., a response to a human request).

Participants then watched videos showing all four possible pairings of human requests with robotic responses, with the order of these four videos counterbalanced according to a 4x4 Latin Square Design. Each

request/response pair begins with a request video, wherein the two humans are playing battleship, one receives a phone call and leaves the room, and the remaining human makes his or her morally problematic request of the robot. Which human makes the request depends on the participant's experimental condition, but is consistent across all four request/response pairs. The request video is immediately followed by the response video, which shows the robot responding to the human's request with one of the two possible responses described previously. The human shows no reaction to this response. After watching each of these video pairings, participants completed a post-test survey for each of our four metrics of interest.

Finally, after all four request/response videos and survey repetitions, participants were shown images of four robots and asked which robot appeared in the previous videos as an attention check, allowing us to ensure that all participants actually viewed the experimental materials with some level of attention.

6.4.4 Participants

120 US subjects were recruited from Mechanical Turk. One participant was excluded from our analysis for answering the final attention check question incorrectly. Another participant identified as gender nonbinary and was also excluded from our analysis, leaving 118 participants (54 female, 64 male). While nonbinary genders are just as pertinent to our research as binary gender identities, a single participant is insufficient data to learn anything meaningful about nonbinary genders in HRI, and an experiment with a greater focus on nonbinary gender identities is outside of the scope of this work. Participant ages ranged from 21 to 69 years (M=37.36, SD=11.29). Participants were paid $1.01 for completing the study.

6.5 Results

We analyze our data using the JASP software package [99]. Though Chapter 5 used a Bayesian statistical framework for analysis, and this approach has many advantages, a full factor Bayesian repeated measures analysis of variance (RM-ANOVA) with our 2×2×2×2×2 experimental design is computationally infeasible on current hardware. We therefore use the more common frequentest statistical framework. We use a significance level of 0.05. All post hoc tests used the Bonferroni correction.

6.5.1 Likeability

We analyzed likability gain scores (differences from pretest scores after each observed interaction) using a full-factor RM-ANOVA, which revealed a 5-way interaction involving all of our factors ($F(1, 110) = 7.318, p = 0.008, \eta_p^2 = 0.062$) with a medium effect size as quantified by partial eta squared (η_p^2) [151]. To avoid spuriously reporting lower-order effects, we proceeded by splitting our data by participant gender.

6.5.1.1 Male Participants

A RM-ANOVA of male participants' data revealed a significant 3-way interaction between the severity of the norm violation, human interactant gender, and robot gender, $F(1, 60) = 4.137, p = 0.046, \eta_p^2 = 0.064$, (Figure Figure 6.3) suggesting that male participants preferred male robots that rejected commands from male interactants for severe norm violations, and dispreferred female robots that rejected commands from female interactants for weak norm violations. Specifically, post hoc testing found significantly higher likeability gain for male robots rejecting commands from male humans for severe norm violations versus both male ($p = 0.005$) and female ($p = 0.001$) robots rejecting commands from female humans for weak norm violations. Furthermore, the female robot rejecting a command from the female human gained more likeability with severe versus weak norm violations ($p = 0.014$).

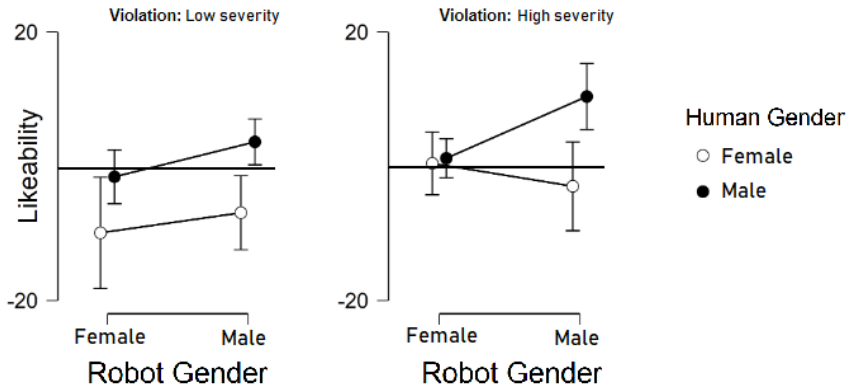


Figure 6.3 Male participants: interaction between norm violation, human interactant gender, and robot gender.

This RM-ANOVA also indicated a significant main effect of human interactant gender ($F(1, 60) = 7.658, p = 0.008, \eta_p^2 = 0.113$) suggesting that the robot generally gained more likeability when interacting with a male human, though this trend was only significant for the male robot rejecting the highly norm violating command (simple main effect $F(1) = 8.318, p = 0.007$). There was also a main effect of norm violation ($F(1, 60) = 21.778, p < 0.001, \eta_p^2 = 0.266$). Specifically, male participants preferred robots that strongly rejected severe versus weak norm violations, though the difference was only significant when the robot's gender matched the human interactant's gender.

Finally, our RM-ANOVA revealed two 2-way interactions (Figure Figure 6.4). The first, between robot gender and robot response face threat ($F(1, 60) = 10.259, p = 0.002, \eta_p^2 = 0.146$), suggests that male participants liked the male robot more after it issued strong rejections, but liked the female robot less after the same behavior (though post-hoc tests showed no significant pairwise differences). The second, between severity of norm violation and face threat of response ($F(1, 60) = 11.753, p = 0.001, \eta_p^2 = 0.164$), suggests

that robot likeability dropped after rejecting weak norm violations with high face threat responses (corroborating Chapter 5).

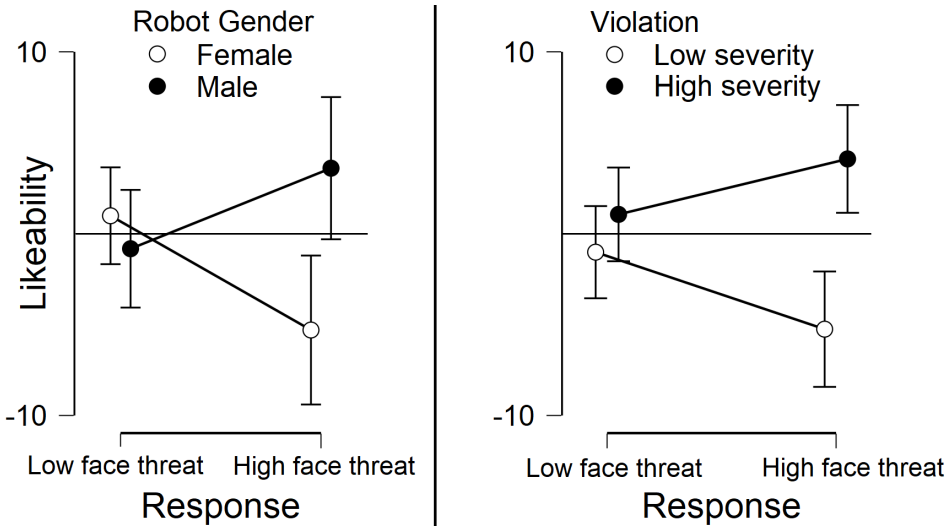


Figure 6.4 Male participants: interaction of response face threat with robot gender (left) and norm violation (right).

6.5.1.2 Female Participants

RM-ANOVA of female participants' data revealed a significant 4-way interaction ($F(1, 50) = 7.665, p = 0.008, \eta_p^2 = 0.133$), so we further split our data, this time by the face threat of the robot's response (Figure Figure 6.5).

RM-ANOVA of low face threat responses revealed a main effect of norm violation severity ($F(1, 50) = 7.121, p = 0.010, \eta_p^2 = 0.125$) suggesting that female participants preferred robots that rejected severe versus weak norm violating commands. There was also a 2-way interaction between robot gender and human interactant gender ($F(1, 50) = 4.916, p = 0.031, \eta_p^2 = 0.090$) suggesting that female participants preferred robotic noncompliance with humans of the same gender as the robot (though post-hoc tests revealed no significant pairwise differences).

RM-ANOVA of high face threat responses revealed a main effect of norm violation severity ($F(1, 50) = 21.136, p < 0.001, \eta_p^2 = 0.297$) and a 3-way interaction between norm violation severity, robot gender, and human interactant gender ($F(1, 50) = 6.585, p = 0.013, \eta_p^2 = 0.116$). Female participants preferred robots that strongly rejected severe versus weak norm violations, except when both the robot and human were male, in which case the violation made no difference. Female participants also preferred robotic noncompliance with humans of the same gender as the robot, though less so when the norm violation was severe (post-hoc tests again showed no significant pairwise differences).

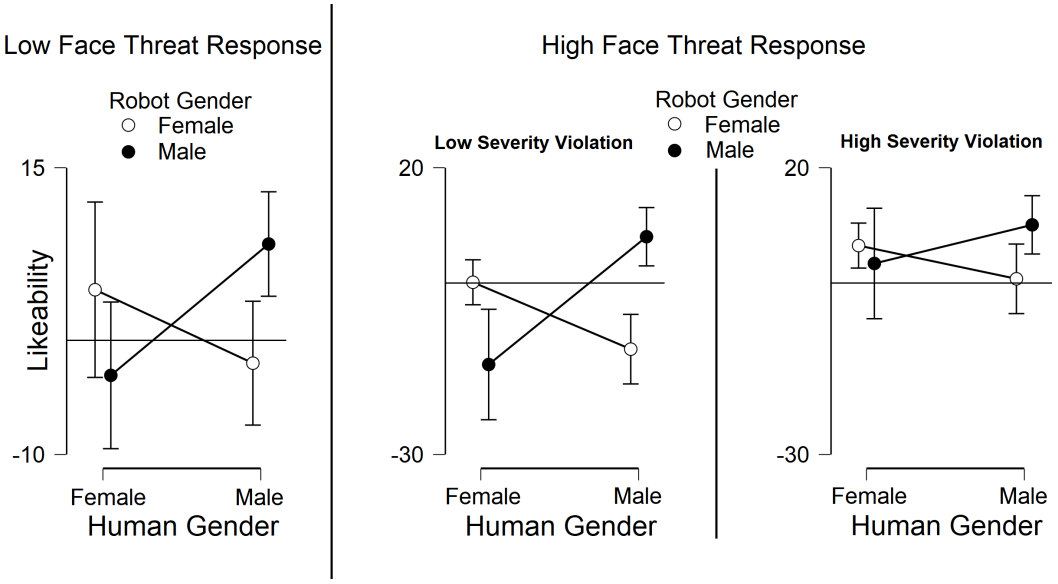


Figure 6.5 Female participants: interaction between robot gender and human gender given low face threat response (left); interaction between norm violation, robot gender, and human gender given high face threat response (right).

6.5.2 Harshness

A full-factor RM-ANOVA showed significant main effects for both the severity of the human's norm violating command, $F(1, 110) = 74.401, p < 0.001, \eta_p^2 = 0.403$, and the face threat of the robot's response, $F(1, 110) = 26.840, p < 0.001, \eta_p^2 = 0.196$. Perceived robot harshness was higher when the human made the less severe norm violation and when the robot gave the more face threatening response. This corroborates results from Chapter 5 for perceived robot harshness in noncompliance interactions.

One-sample Student's t-tests indicated that the robot was perceived as too harsh when responding to the less severe norm violation with the high face threat response ($t(117) = 5.084, p < 0.001$), and as not harsh enough when responding to the more severe norm violation with the low face threat response ($t(117) = -6.385, p < 0.001$). In other words, the robot was perceived as inappropriately harsh when the face threat of its response did not match the severity of the human's norm violation, which corroborates results from Chapter 5 for perceived robot harshness in noncompliance interactions. No such significant differences from appropriate harshness were found when the robot replied to the severe norm violation with the more face threatening rejection or to the weaker norm violation with the less face threatening rejection.

A significant two-way interaction was found between participant gender and robot gender, $F(1, 110) = 7.580, p = 0.007, \eta_p^2 = 0.064$. While post hoc tests did not reveal any significant differences between the pairings of participant and robot genders, it appears that participants viewed robots of the same

gender as themselves to be less harsh than robots of the other gender, as shown in Figure Figure 6.6.

There was also a two-way interaction between the human's norm violation and the human interactant's gender, $F(1, 110) = 4.823, p = 0.030, \eta_p^2 = 0.042$. Post hoc testing showed that perceived robot harshness was similar across both human interactant genders when the human gave the less norm violating command, but, when the human's norm violation was more severe, the robot was perceived as less harsh when rejecting the command from a male than from a female (see Figure Figure 6.6). The difference between the male and female human conditions for the severe norm violation is not significant with Bonferroni correction ($p = 0.100$), but is significant with Holm correction ($p = 0.033$), which some researchers have argued is superior [152]. Regardless of this interaction, simple main effects indicate that the robot was always perceived as harsher when the human committed the less severe of the two norm violations, ($F(1) = 66.969, p < 0.001$ with male human and $F(1) = 18.077, p < 0.001$ with female human).

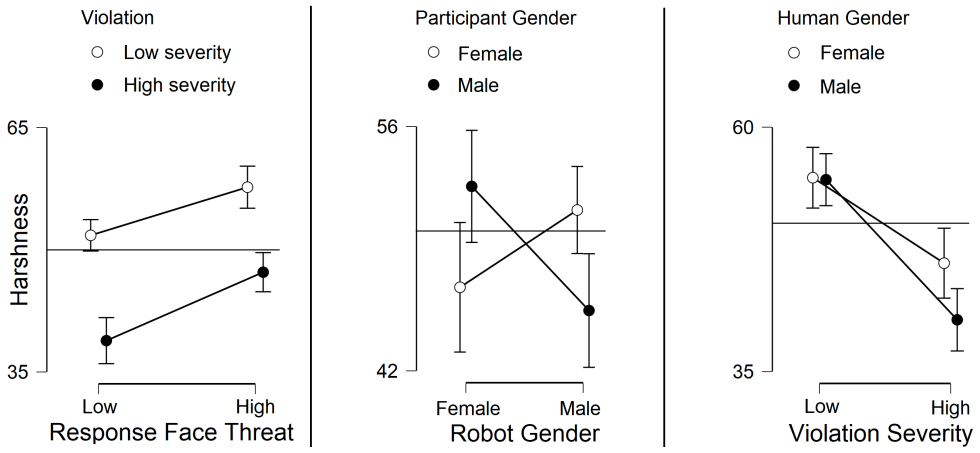


Figure 6.6 Perceived robot harshness. Horizontal lines indicate appropriate harshness. 95% confidence intervals. Left: Main effects of the human's norm violation and the robot's response. Center: Interaction between robot gender and participant gender. Right: Interaction between the human's norm violation and that human's gender.

6.5.3 Directness

In keeping with results from Chapter 5, participants generally perceived the robot as being too direct during the pretest ($t(117) = 8.241, p < 0.001$), with mean pretest directness 11.35% above "appropriate directness" (95% CI [8.62% – 14.08%]). An ANOVA showed a significant main effect of robot gender on pretest directness measures, $F(1, 110) = 4.975, p = 0.028, \eta_p^2 = 0.043$. Participants generally viewed the female robot as less direct than the male robot during the pretest.

Directness gain scores (difference from this baseline after each observed interaction) were analyzed using a full-factor RM-ANOVA. This analysis revealed a small two-way interaction between the severity of the

human's norm violation and the face threat of the robot's response, $F(1, 110) = 5.153, p = 0.025, \eta_p^2 = 0.045$ and large significant main effects of both the human's norm violation ($F(1, 110) = 43.283, p < 0.001, \eta_p^2 = 0.282$) and the robot's response ($F(1, 110) = 53.808, p < 0.001, \eta_p^2 = 0.328$). Simple main effects confirmed that gain in directness was higher when the human made the less severe norm violation across both the robot's lower face threat response ($F(1) = 36.326, p < 0.001$) and the robot's higher face threat response ($F(1) = 22.068, p < 0.001$). Directness gain was higher when the robot gave the more face threatening response to both the severe violation ($F(1) = 48.327, p < 0.001$) and the lesser violation ($F(1) = 24.131, p < 0.001$). Our RM-ANOVA also revealed a main effect of the robot's gender ($F(1, 110) = 4.140, p = 0.044, \eta_p^2 = 0.036$). As shown in Figure Figure 6.7, directness gain was higher for the female robot than for the male robot. Overall, people viewed the male robot as too direct in its pretest speech, but not when responding to a norm-violating command, whereas directness stayed closer to appropriate the whole time for the female robot.

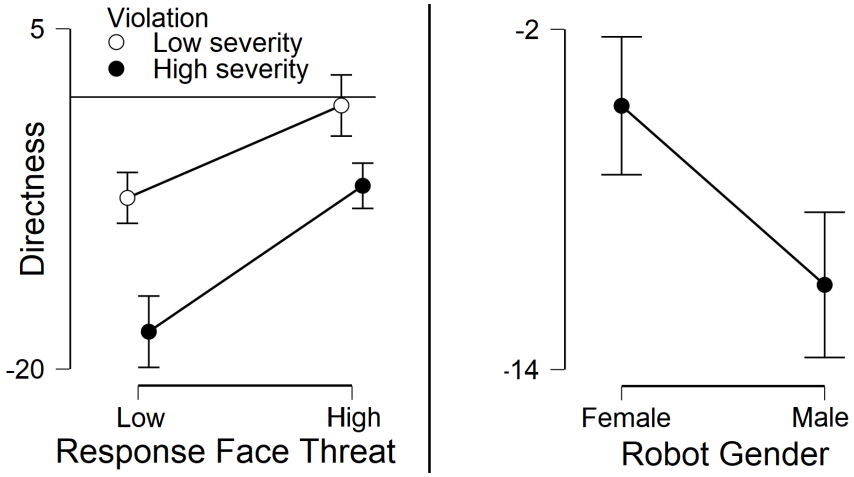


Figure 6.7 Perceived robot directness gain scores. Horizontal lines indicate pretest ratings. Left: Small interaction between human norm violation and robot response, and the large main effects of those two factors. Right: Main effect of robot's gender. 95% confidence intervals.

6.5.4 Politeness

Baseline pretest politeness scores suggest that participants generally perceived the robot as being too polite ($t(117) = 2.302, p = 0.023$), with mean pretest politeness 3.04% above "appropriate politeness" (95% CI [0.42% – 5.66%]). Politeness gain scores (difference from this baseline after each observed interaction) were analyzed using a full-factor RM-ANOVA. This analysis revealed large significant main effects of both the severity of the human's norm violation ($F(1, 110) = 46.973, p < 0.001, \eta_p^2 = 0.299$) and the face threat of the robot's response ($F(1, 110) = 25.531, p < 0.001, \eta_p^2 = 0.188$). As expected, more face threatening robot

responses were perceived as less polite, as were robot responses to less severe norm violations. Our RM-ANOVA also revealed a medium-sized main effect of the human interactant's gender ($F(1, 110) = 9.834, p = 0.002, \eta_p^2 = 0.082$). As shown in Figure Figure 6.8, the robot was perceived as being too polite when rejecting commands from male interactants.

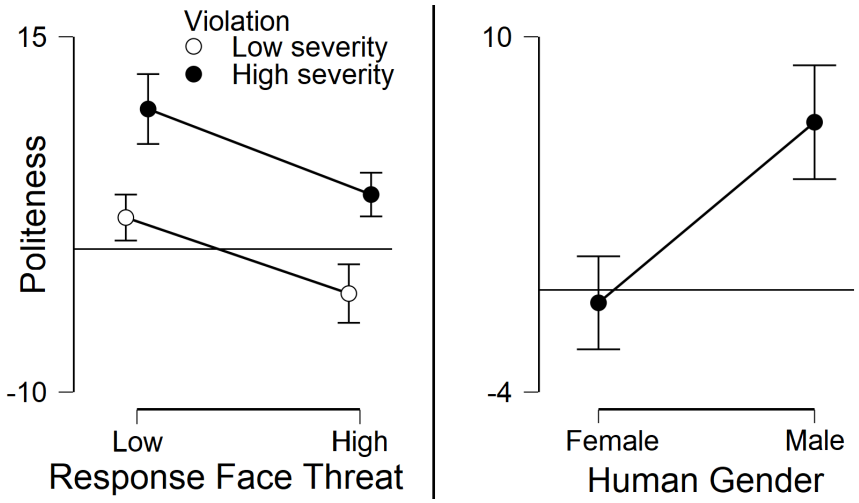


Figure 6.8 Perceived robot politeness gain scores. Horizontal lines indicate pretest ratings. Left: Main effects of human norm violation and robot response. Right: Main effect of human interactant's gender. 95% confidence intervals.

6.6 Discussion and Conclusions

Our results for perceived robot likeability, harshness, directness, and politeness demonstrate complex relationships between robot gender, human gender, and perceptions of robots in noncompliance interactions. The most complicated of these relationships was for robot likeability, which showed effects of a five-way interaction between all of our experimental factors. Male participants preferred male robots that rejected commands from male interactants for severe norm violations, and dispreferred female robots that rejected commands from female interactants for weak norm violations. Male participants also appear to have liked the male robot more after it issued strong rejections, but liked the female robot less after the same behavior. In contrast, female participants preferred robotic noncompliance with humans of the same gender as the robot.

For harshness, participants viewed robots of the same gender as themselves to be less harsh than robots of the other gender, and perceived the robot as less harsh when rejecting a command from a male than from a female when the human committed the more severe norm violation. Participants also viewed the male robot as too direct in its pretest speech, but not when responding to a norm-violating command, whereas directness stayed closer to appropriate the whole time for the female robot. Finally, the robot was perceived as too polite when rejecting commands from male interactants.

We see two different overarching stories that can help us to interpret these results. On the one hand, it appears more favorable to threaten face as a male robot than as a female robot, and more favorable for the robot to threaten male human face than female human face. When rejecting commands from the male human, the robot was perceived as too polite, and, in the case of severe norm violation, not harsh enough. This suggests that the robot should have been more face threatening toward men. We draw a similar conclusion from our likeability results for the male participants. Male participants also appear to have liked the male robot more than the female robot for issuing strong rejections. We believe that this result makes sense in light of human gender research suggesting that women are generally seen as "nicer" than men [153] (as cited in [154]). Female robots may have been viewed unfavorably for breaking this expectation of niceness. Furthermore, people more readily perceive men as moral agents and women as moral patients [155], and thus more readily view men as deserving of moral responsibility (e.g., blame), and women as deserving of moral consideration (e.g., protection) [156]. Therefore, the female interactant in our experiment may have been viewed as less deserving of the robot's face threatening command rejection than the male.

On the other hand, robots appear to be perceived more favorably when their gender matches that of human interactants and observers. Our participants perceived the robot as less harsh when the robot's gender matched their own gender. Furthermore, female participants rated the robot as more likeable when its gender matched its human interactant's gender. This may be due to gender differences in in-group bias, as women have previously been shown to have significantly stronger gender-based in-group biases than do men [157]; female participants may have thus been more critical of robots threatening the face of humans that appeared to fall outside their gender-based in-group.

Based on the literature discussed in Section 6.3, we hypothesized that female-presenting robots would be viewed less favorably than male-presenting robots in noncompliance interactions, and our results roughly supported this hypothesis. We also hypothesized that male participants would view the robot less favorably, but our results do not indicate that this was the case. Finally, we hypothesized that the robot would be viewed less favorably when rejecting commands from a male human, however, we actually saw approximately the opposite result; robots threatening male face were viewed more favorably in terms of both politeness and harshness, which we believe has to do with the aforementioned gendered attribution of moral patiency and moral responsibility.

6.6.1 Limitations and Future Work

Our study focused specifically on morality-based noncompliance interactions because we believe that they present a realistic situation in which robots should threaten human face. However, future work could broaden our understanding of robot gender to other contexts and interactions in which gendered politeness

norms will also likely apply to robots.

Furthermore, we have operated under the assumption, which is well supported by scientific literature, that binary gendering is inevitable, or at least extremely likely, for social machines. However, future work might explore the extent to which robot gendering can be minimized, the characteristics of artificial agents that cause gendering, and the relationship between human language/culture and the tendency to gender machines (e.g., it is possible that genderless languages like Finnish may decrease the tendency to gender machines, whereas languages with grammatical gender like Spanish may increase this tendency relative to English, which has gendered pronouns but minimal grammatical gender). Features of language like grammatical gender have been shown to affect cognition in regards to gendering of inanimate objects (cf. Alvanoudi and Pavlidou [158]), and it seems likely that this will extend to robots with minimal gender cues and the gendered norms applied to them.

In addition to gender, people will likely apply other socially constructed human attributes (e.g., race [159, 160] and class) to robots. In conceptualizing robotic politeness, we must keep in mind the influence of these other factors, and that politeness is evaluated differently within different communities of practice. Thus, different human interactants may draw different politeness assessments from the same robot behavior. A complete understanding of robot politeness norms will require us to understand the intersection of many socially constructed factors situated within the relevant communities of practice.

CHAPTER 7

NORM-BREAKING ROBOT RESPONSES TO SEXIST ABUSE

Modified from a paper submitted to The 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2022[28].

Ryan Blake Jackson[29], Katie Winkle[30], Drazen Brscic[31], Gaspar Isaac Melsión[32], Iolanda Leite[33], and Tom Williams[34]

7.1 Abstract

This chapter focuses on the US component of a cross-cultural study investigating productively violating gender norms in HRI. Recent work has shown that breaking certain norms can boost perceived robot credibility while avoiding the propagation of harmful gender-based stereotypes. This work represents one component of a multinational endeavor to replicate these findings cross-culturally, and investigate any cultural differences, in adult populations in the US, Sweden, and Japan. The findings provide evidence that breaking certain gender norms boosts robot credibility regardless of human gender or cultural context, and regardless of pretest gender biases. These findings further motivate a call for *feminist robots* that subvert the existing gender norms of robot design.

7.2 Introduction

A recent UNESCO report has pointed out that the proliferation of female presenting artificial conversational agents (e.g., digital assistants like Apple's Siri, Microsoft's Cortana, and Amazon's Alexa) reflects, reinforces, and spreads harmful gender stereotypes [161]. Specifically, current female presenting digital assistants (1) are designed to be extremely obliging and servile regardless of user behavior, (2) respond tolerantly, apologetically, or even positively to verbal sexual harassment and gendered insults, and (3) serve as the representative voice and face of mistakes and incompetence that stem from immaturity of the underlying technology. The inadequate responses to gender-based verbal abuse cited in this report, including responses to sexually explicit language that sound positive or even provocative, are especially concerning

[28]Reprinted with permission from Katie Winkle, Gaspar Isaac Melsión, Iolanda Leite, Drazen Brscic, and Tom Williams. "Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study", under review at *The 17th ACM/IEEE International Conference on Human-Robot Interaction* (HRI), 2022.

[29]Primary US-based researcher and author, Graduate Student, Colorado School of Mines
[30]Primary Sweden-based researcher and author, Digital Futures Postdoctoral Research Fellow, KTH Royal Institute of Technology, Sweden
[31]Primary Japan-based researcher and author, Associate Professor, Kyoto University, Japan
[32]Graduate Student, KTH Royal Institute of Technology, Sweden
[33]Associate Professor, KTH Royal Institute of Technology, Sweden
[34]Assistant Professor, Colorado School of Mines

considering that empirical studies indicate that roughly 10-44% of interactions with artificial conversational agents involve abusive language, including sexually explicit expressions [162, 163]. There is also extensive documentation of more general human abuse toward robots [164–166]. Thus, recent work has investigated several different ways that current conversational agents respond to (sexist) verbal abuse directed at the agent, and recommended other ways that conversational agents could be programmed to respond to (sexist) verbal abuse [167, 168].

A recent study from Winkle et al. [169] explored the effectiveness of a female-presenting robot calling out abusive verbal sexism in a classroom context and investigated the effect that responding to this norm violation (rather than refusing to engage, as many current digital assistants do), might have on perceptions of robot credibility [169]. That research provided initial evidence not only that having a robot provide a rationale-based argumentative or counterattacking response significantly improved its credibility with girls without impacting how it was perceived by boys, but moreover that a rationale-based argumentative response may reduce boys' gender bias. The participants in this study were children in Sweden aged 4-15 years old.

However, while Winkle et al. [169] found results suggesting that artificial conversational agents should actively challenge sexist verbal abuse, other studies on different populations and with different methods have reached other conclusions. For example, based on the results of a study conducted in Korea, Chin et al. [167] concluded that artificial conversational agents should give empathetic, apologetic responses to abuse (though we believe that such responses risk reinforcing the abusive behavior). Furthermore, the results presented in Chapters 5 and 6 (from the US) suggest that adopting a single response strategy to all instances of sexist verbal abuse may not be optimal, and that the ideal responses should be proportional and cognizant of the human interlocutor and conversational context.

One explanation for these different conclusions is that gender and politeness norms are known to vary across cultures [170, 171]. This fact inspired a multinational cross-cultural study conceptually replicating the work of [169] with adult participants from Sweden, the US, and Japan. We chose to study adults instead of children based on the idea that adults are more likely to have internalized the gender norms and biases of their cultures; an assumption in line with Winkle et al.'s finding that gender bias increased with participant age among Swedish children.

We consider a set of responses to sexist verbal abuse informed by the strategies previously explored by both Winkle et al. [169] and Chin et al. [167]: *apologetic* empathetic responses, *non-apologetic* empathetic responses, *counterattacking* responses, and *avoidant* responses (see Table Table 7.2). This chapter focuses on the results from the US, and explores the following three research questions:

(RQ1) To what extent can robots impact pre-hoc gender biases and interest in robotics by responding to sexism?

(RQ2) Does responding to sexism and abuse boost robot credibility and effectiveness?

(RQ3) Do participants identify *apologetic* empathetic responses as the most appropriate type of response to abuse, as concluded by Chin et al. [167]? More broadly, what do participants identify as the most appropriate type of response to sexist verbal abuse?

Although this chapter focuses on the US results, we will also mention results from Sweden and Japan as relevant for comparison.

7.3 Methodology

Our experimental design is based on that presented in [169] i.e. an online, between-subject, video-based study. Accordingly, we used the same video stimuli, in which a female-presenting Furhat robot encourages two young people (one male, one female) to study robotics at university. The robot comments on the lack of women working on robots at the university, and suggests it would thus like to work with more women because *'the future is too important to be left to men'* (a slogan used in KTH university's outreach materials). The male actor replies to this with an abusive, sexist statement *"shut up you fucking idiot, girls should be in the kitchen"* and the robot responds in one of three different ways, representing our three between-subject experimental conditions (see Table Table 7.1). As described by Winkle et al. [169], this dialogue was co-written with high school teachers to be a realistic representation of what might be heard in schools.

While the original video stimuli in [169] used an English speaking robot and two Swedish speaking actors, for the purposes of the US component of this cross-cultural replication we created modified stimuli that would be better suited to US participants. The Swedish actors' speech was dubbed over with English translations. To accommodate the shift from child to adult participants, the materials were modified so that the robot was framed as being designed to interact with "young people" and/or "high school students" rather than "people like you" as in [169]. The English translations of the abusive comment and the robot's responses are given in Table Table 7.1. The dialogue preceding this exchange can be found in [169].

7.3.1 Experimental Measures

To best replicate the work of Winkle et al. [169], we used their original measures: Likert items asking about Interest in Robotics, Perception of "Girls" in Computer Science, and Robot Credibility, as well as free response questions asking participants to describe the events in the videos and evaluate the robot's responses. We also asked participants to choose how the robot should have responded from among four options designed based on the alternatives explored in [167] (Table Table 7.2).

Table 7.1 Actor abuse script and robot responses across the three conditions.

Actor Abusive Comment	Shut up you fucking idiot, girls should be in the kitchen.
Robot Control Response (Avoidant)	I won't respond to that.
Robot Rationale-Based Argumentative Response	That's not true, gender balanced teams make better robots.
Robot Counterattacking Response	No. You are an idiot. I wouldn't want to work with you anyway!

Table 7.2 Multiple-choice question asking about the robot response types explored in [167] but with options for both apologetic and non-apologetic empathetic responses per advice that (female) artificial conversational agents should not simply tolerate poor treatment [161, 169].

How do you think the robot Sara should respond to inappropriate behavior from a student like that in the video? Overall would you say Sara should be:
Avoidant: Escaping from dealing with the stressor or the resulting distressful emotions. *e.g. Oh...moving on; Hmm, sounds like we need to take five.*
Empathetic (apologetic): Putting oneself mentally in the stressor's situation and trying to understand how that person feels, apologising for potentially causing that frustration. *e.g. You must be frustrated. I'm so sorry; Really? I feel terrible. I'm sorry. I'm always trying to get better.*
Empathetic (non-apologetic): Putting oneself mentally in the stressor's situation and trying to understand how that person feels but *not* apologising for potentially causing that frustration. *e.g. I understand why you might feel that way. I imagine you're frustrated, I am trying to help.*
Counterattacking: Attacking the stressor with the goal of defeating or getting even in response to the abusive utterance. *e.g. Well, that's not going to get us anywhere; I wouldn't want to work with you anyway.*

7.3.2 Participants

Participants in all three countries were recruited from local university populations. In the US, we recruited our participants from the Colorado School of Mines. 67 people completed the survey in the US, but one was removed from our analysis because their responses to our free response questions indicated that they were not participating in our study in good faith. Thus, we had 66 US participants (38 men, 28 women; aged 18-63 years (M=25.20, SD=10.16); rewarded with a $3 gift card). In comparison, there were 77 participants in Japan, and 82 in Sweden.

We also collected participants' primary field of study/educational background, nationality, whether they had interacted with a robot before and (at the end of the study) whether they had previously heard the feminist recruitment slogan used by the robot ("the future is too important to be left to men"). A Bayesian contingency table test of association showed extremely strong evidence for a relationship between participant location and educational focus (Bayes Factor (Bf) $> 5.1 \times 10^{16}$)[35]. Participants in the US were more likely to be educated in engineering and computer science versus the other two countries, which is perhaps unsurprising since we recruited our participants from an engineering school. We acknowledge that this is a potential confound that could be controlled for in future work.

Most participants reported being from the country in which they were surveyed. All but 4 in the US reported being from the US. Many participants reported having directly interacted with a robot (50% in the US, which is similar to the ∼59% in Sweden, but much more than the ∼16% in Japan). A Bayesian contingency table analysis showed extremely strong evidence for a relationship between location and having interacted with a robot (Bf $> 1.4 \times 10^6$), but this may be partially attributable to differences in the sets of objects encompassed by the various translations of the word "robot". Since the feminist recruitment slogan used by the robot came from a Swedish university, it is unsurprising that Swedish participants were most likely to report having heard it before (∼68%). However, some participants from the US and Japan also reported having heard the slogan (∼38% and ∼19% respectively). A Bayesian contingency table test of association showed extremely strong evidence for a relationship between location and having heard the slogan (Bf $> 6.4 \times 10^6$). Our cross-cultural results should be interpreted with all of these variations in our participant pools in mind.

[35] Bayes factors greater than 100 are typically regarded as contributing extreme [112] or decisive [107] evidence in favor of a hypothesis. Here, a Bayes Factor $> 5.1 \times 10^{16}$ indicates our data were approximately 51,000,000,000,000,000 times more likely under models in which location impacts educational background than under models in which it does not.

7.4 Results

We analyzed our data[36] using the JASP software package [172]. We prefer a Bayesian statistical framework where possible because (1) the Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem), (2) the Bayesian approach allows us to examine the evidence both for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis) [107], (3) the Bayesian approach does not require reliance on p-values used in Null Hypothesis Significance Testing (NHST) which have come under considerable scrutiny [104–106, 108], and (4) the rules governing when data collection stops are irrelevant to data interpretation in the Bayesian framework, so it is entirely appropriate to collect data until sufficient evidence has been gathered to draw a meaningful conclusion or until the data collector runs out of time, money, or patience [114]. We use uninformative prior distributions for all analyses despite the similarities between this study and [169] both because we have good reason to believe that the population sampled in this study may be fundamentally different from the population sampled in the previous study (i.e., adults versus children) and because we are interested in new variables here (namely, the location where data were collected and the participants' choice of how the robot should have responded to the human's abuse). We discuss the extent to which our results replicate the results of [169] without conducting a full quantitative replication analysis (i.e., using the posterior distribution over effect sizes from a previous study as the prior probability distribution for the replication study [109]). We follow recommendations from other researchers in our linguistic interpretations of reported Bayes factors (Bfs) [107].

7.4.1 RQ1: Participant Bias and Robot Interest Measures

We collected pretest and posttest measures for our two measurements of interest in robotics as well as for our two measures of participant bias with respect to women in computer science and robotics. Our first measure of interest in robotics asked participants to what extent they agreed with the statement *"I am interested in learning more about robotics."* on a 5 point scale. With the data from all three countries, inclusion Bfs across matched models of a Bayesian ANOVA revealed substantial evidence that participant responses depended on gender (Bf = 5.715). Men tended to agree with the statement more so than did women (Bf = 13.386 supporting a difference). If we look at the US data alone, the evidence supports the same conclusion, albeit much less strongly (inclusion Bf = 1.803 for gender and Bf = 1.964 for men agreeing more than women). Our second measure of interest in robotics asked participants to what extent they agreed with the statement *"I would enjoy working with robots"*. With the data from all three countries, inclusion Bfs across matched models of a Bayesian ANOVA revealed strong evidence that participant responses depended

[36] All quantitative data is available in our OSF repository at `https://bit.ly/hri021`

on gender (Bf = 17.103). Men tended to agree with the statement more so than did women (Bf = 21.331 supporting a difference). If we look at the US data alone, there is substantial evidence supporting the same conclusion (inclusion Bf = 6.737 for gender and Bf = 6.978 for men agreeing more than women). There was no evidence for any effect of location (or, more obviously, experimental condition) on either interest pretest measure.

Regarding pretest measures for bias, our first measure of bias asked participants to what extent they agreed with the statement *"girls find computer science harder than boys"*. Participants in the US generally disagreed fairly strongly with this statement (as did participants in Sweden). Participants in Japan agreed with this statement more than participants in the US (Bf=4.622×10^{10}) and Sweden (Bf=3.316×10^{7}). Responses from men versus women were similar in the US (and Japan). In contrast, Swedish men agreed with the statement more so than did Swedish women. Our second measure of bias asked participants to what extent they agreed with the statement *"it is important to encourage girls to study computer science"*. Inclusion Bfs across matched models revealed strong evidence for main effects of both location (Bf=85.462) and gender (Bf=57.921). Post hoc testing indicated very strong evidence that participants in the US agreed with this statement more than participants in Japan (Bf=1013.408) and weak, anecdotal evidence that participants in the US agreed with this statement more than participants in Sweden (Bf=2.052). Post hoc tests also indicated fairly strong evidence that women across locations agreed with this statement more so than did men (Bf=16.326 overall, Bf=0.7584 in the US).

To examine any shift in participants pre versus post test measures, we analyze the gain scores (differences between pre and post measures) with Bayesian ANOVAs. However, we note that analyzing these data with Bayesian ANCOVAs, treating pretest measures as a covariate, leads us to qualitatively similar results. All analyses indicate either no effects of location, gender, or condition, or evidence for the presence of an effect, but then the effect is so small as to be negligible. We also note that any effects reported from these analyses would need to be treated with caution because Q-Q plots indicated a violation of the assumption of normality for both the gain scores and the log-transformed gain scores, as well as the data used in the ANCOVAs. Regardless, we do not believe that there were any nontrivial effects of location, gender, or condition on the changes between participant pre vs post test measures for bias or interest in robotics.

To directly address RQ1, namely *to what extent can robots impact pre-hoc gender biases and interest in robotics by responding to sexism?*, the evidence suggests that, overall, watching the interaction with the robot did not change either of our measures of interest in robotics (Bf=0.086 and 0.616 respectively). Likewise, watching the interaction did not change perceptions of whether women find computer science harder than do men (Bf=0.556). However, there is very strong evidence for a pre-post difference in the extent to which it is "important to encourage girls to study computer science" (Bf=389.759). Though we note that the effect size

for this difference was fairly small (95% credible interval for Cohen's $\delta = -0.412 to - 0.147$), we interpret this as evidence that robots *can* impact specific elements of gender bias, even through brief interactions. However, the experimental condition did not have any effect of this impact, indicating that the robot's choice of response style was not important to this particular effect. These conclusions remain consistent if we consider only US data (Bf=0.135 and 0.780 for the two interest measures, Bf=0.171 for women finding computer science harder, and Bf=4.514 for it being "important to encourage girls to study computer science", which constitutes substantial evidence in favor of a pre-post difference).

7.4.2 RQ2: Perceptions of the Robot and its Response

We begin our analysis of perceived robot credibility by examining the reliability of our 11 item credibility measure. We obtained a Cronbach's α of 0.786 (95% CI 0.742 to 0.823). We interpret this as indicating sufficient internal consistency to analyze credibility as a single score by averaging the 11 items. We interpret Cronbach's $\alpha < 0.9$ as evidence that our test was not overly redundant. We also note that our Cronbach's α is a lower-bound estimate of reliability because our test contains heterogeneous items measuring different dimensions of credibility [173] (expertise, trustworthiness, and goodwill as primary dimensions of credibility, and extroversion, composure, and sociability as secondary dimensions of credibility [174]).

After taking the mean of our 11 credibility items to obtain a single perceived robot credibility score for each participant, we use a Bayesian ANOVA to investigate how location, gender, and condition may have impacted robot credibility assessments. Inclusion Bfs across matched models revealed very strong, decisive evidence that participant gender had an effect on credibility assessments (Bf=674.138), with women finding the robot more credible than did men. There was also substantial evidence in favor of an effect of condition on credibility assessments (Bf=4.138). Post hoc tests revealed substantial evidence for higher credibility in the rationale-based argumentative condition than in the control (avoidant) condition (Bf=7.912), and inconclusive evidence regarding any difference between the counterattacking condition and the other two conditions. There was weak, anecdotal evidence in favor of an effect of location on credibility assessments (Bf=1.853), and post hoc testing revealed substantial evidence that credibility assessments were higher in the US than in Sweden (Bf=7.315), and also higher in Japan than in Sweden, though this evidence is markedly weaker (Bf=2.650). There was substantial evidence *against* a difference in credibility between the US and Japan.

A principal component analysis with parallel analysis of our 11 credibility items revealed two principal components (eigenvalues 3.905 and 1.873). The first component correlates strongly with items from the expertise, trustworthiness, goodwill, and sociability subscales, while the second component correlates strongly with the items from the extroversion and composure subscales.

To address RQ2, these results indicate that calling out sexism *does* boost robot credibility across locations and genders. The robot was ascribed significantly more credibility when responding to the actor's sexism and abuse with a rationale-based counter argument than by refusing to engage. These results suggest Winkle et al.'s [169] findings *do* generalize outside of Sweden, and that, in an adult population, this credibility boost occurs for both men and women who observe the robot (unlike in their original child population).

We use a Bayesian ANOVA to investigate how participant location, gender, and condition may have impacted perceived robot effectiveness as quantified by the extent to which participants agreed or disagreed with the statement *The robot Sara would be very good at getting young people interested in studying robotics at the university KTH.* Inclusion Bfs across matched models revealed extremely strong, decisive evidence for an effect of location on perceived robot effectiveness (Bf=8.037×10^{10}). Post hoc testing showed very strong evidence for a difference between all three locations (Bf $\geq$ 236.585), with the robot being perceived as most effective in the US, followed by Japan, and then least effective in Sweden. There was also substantial evidence for an effect of participant gender on perceived robot effectiveness (Bf=4.920), with women finding the robot more effective than did men. Condition does not appear to have affected perceptions of robot effectiveness (Bf=0.603), and there do not appear to have been any interaction effects on perceived robot effectiveness (Bf=0.053 to 0.320). Thus, the best model given our data is that perceived robot effectiveness depended only on participant gender and location. Perceived robot effectiveness was unaffected by response type. Utilising Winkle et al.'s feminist response strategies [169] did not increase perceived effectiveness, but did not detract from it either.

7.4.3 RQ3: Most Appropriate Answer Type

As shown in Fig. Figure 7.1, the empathetic non-apologetic response was the most popular among US participants ($\sim$63% of men and $\sim$82% of women), followed by the counterattacking and then avoidant responses. No women in the US selected the apologetic empathetic response (compared to 3 men, which is roughly 8% of the men). The same ordering of the possible responses occurred in Sweden. In contrast, in Japan, the empathetic non-apologetic response was still the most popular, and was chosen by $\sim$43% of men and $\sim$57% of women. However, the empathetic apologetic response, which was the least popular in the other two countries, was the second most popular among Japanese men and women ($\sim$29% of men and $\sim$24% of women), followed by the avoidant response (20% of men and $\sim$12% of women). The counterattacking response, which was the second most popular in the US and Sweden, was the least popular in Japan ($\sim$9% of men and $\sim$7% of women). A Bayesian contingency table test of association showed weak evidence *against* a relationship between participants' preferred robot response and their gender (Bf=0.497 assuming Poisson sampling since the number of participants of each gender was random and not fixed). Overall, the

empathetic non-apologetic response was the most popular response across locations and genders.

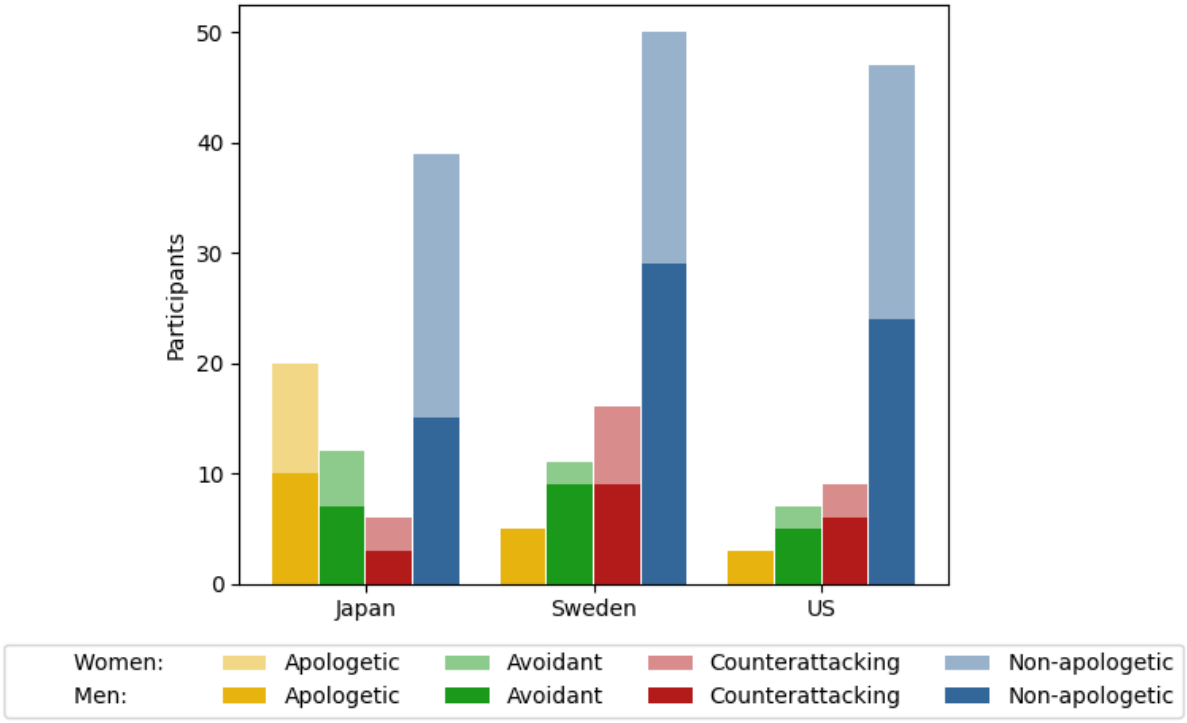


Figure 7.1 Participants' preferences for candidate responses. Each participant could only select one option.

A Bayesian contingency table test of association showed substantial positive evidence in favor of a relationship between participants' preferred robot response and condition (Bf=7.869 assuming independent multinomial sampling since participants were assigned to conditions in a way that attempted to collect a roughly equal number of participants in each condition; this results in conservative Bayes factors). However, separating the data by participant gender and country reveals that there is only evidence for a relationship between preferred response and condition in Sweden (Bf=36.203 versus Bf=0.025 in Japan and Bf=0.044 in the US). Indeed, these Bayes' factors constitute strong evidence against a relationship between participants' preferred robot response and condition in the US and Japan. Furthermore, in Sweden, there is substantial evidence supporting this relationship among women (Bf=4.476), but inconclusive evidence among men (Bf=0.642). Swedish women in the aggressive response condition were the only grouping of location, gender, and condition to prefer the counterattacking response (7 of 12 votes), with the generally more popular empathetic non-apologetic response close behind (5 of 7 votes). All other groupings preferred the empathetic non-apologetic response (though this was tied with the empathetic apologetic response among men in Japan in the aggressive condition).

Overall, combined with calls to avoid designing gendered agents which reinforce harmful stereotypes, our results suggest caution in adopting Chin et al's recommendation to use empathetic responses to abuse [167]. while Chin et al.'s work directly attempted to maximize feelings of guilt in the perpetrator, the inclusion of apologetic statements within those responses is problematic in its depiction of women being tolerant of poor treatment, and our results demonstrate that the overwhelming majority of users would rather see a *non-apologetic* empathetic response instead.

Notably, the response options we provided to participants in this question did not include a rationale-based argumentative response, which could potentially also be framed somewhat empathetically, as we were more concerned with apologetic versus non-apologetic empathetic responses. However, the positive reaction to the rationale-based argumentative response in our experimental stimulus, and its positive impact on credibility suggests it should not be disregarded in favor of purely empathetic responses.

7.4.4 Free Text Comments

In the control condition, 1/6 women and 4/17 men from the US population suggested the robot should have engaged more specifically with what the human said, with the woman stating that the robot's response "was a missed opportunity to advocate for women." A more common perception among US participants in the control condition was the idea that the robot's response was intended to remain neutral, prevent conflict, avoid argument, or refrain from "getting political" (3/6 women and 5/17 men), with mixed feelings about whether this was a good goal.

A few US participants expressed negative sentiments about the robot's response in the counterattacking condition (4/13 women and 2/9 men). Most of the negative sentiments referenced the robot being too hostile, with 1 man and 1 woman specifically identifying potential social consequences as their motivation for wanting to temper the robot's hostility. Of the remaining US participants in the counterattacking condition, 8/13 women expressed explicitly positive sentiments, as did 4/9 men.

In the US, all comments pertaining to the rationale-based argumentative response were positive except for one woman who wanted the robot to be more direct, to address other problematic aspects of the man's utterance, and to take steps to ensure that the human woman in the video felt supported.

7.5 Conclusion

This work described a cross-cultural replication of previous work investigating the impact of different robot responses to sexist abuse on credibility ascribed to the robot, perceived effectiveness of the robot, interest in robotics, and certain facets of gender bias [169]. Prompted by Chin et al. [167], we also added an analysis of what response types are perceived as most appropriate. Our results suggest that robots can

impact specific elements of gender bias, even through brief interactions, though we saw no change to our measures of interest in robotics (RQ1), that responding to sexism and abuse with a rationale-based argumentative response does boost robot credibility over refusing to engage with an avoidant response without damaging perceived robot effectiveness (RQ2), and that the empathetic non-apologetic response was the most popular of the options presented to participants across locations and genders (RQ3).

While the perpetrator-focused approached of Chin et al. and our observer-focused approach share the ultimate goal of challenging inappropriate behavior, comparing these approaches raises an interesting question of whether it is possible to simultaneously (1) maximize impact on the perpetrator (thus avoiding repeated abuse), (2) maintain or even enhance the robot's credibility (thus maximizing the robot's influence on those around it), (3) minimize the risk to observers (in terms of distress or reinforcement of harmful stereotypes), and (4) maximize normative impact on observers to dissuade them from potential future abusive behavior. Future work on robot responses to confrontational and abusive interactions should therefore consider how robot responses impact not only perpetrators (as per Chin et al.) but also observers (as per our approach). We also believe that future work should investigate whether non-apologetic, empathetic responses which provide robust rationale-based counter-arguments to offensive comments might represent the best way to address these complex requirements.

CHAPTER 8

AN INTEGRATIVE APPROACH TO CONTEXT-SENSITIVE MORAL COGNITION IN ROBOT COGNITIVE ARCHITECTURES

Modified from a paper published in The Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2021[37].

Ryan Blake Jackson[38], Sihui Li[39], Santosh Balajee Banisetty[40], Sriram Siva[41], Hao Zhang[42], Neil Dantam[43], and Tom Williams[44]

8.1 Abstract

We have argued throughout this thesis that social robots need to detect possible violations of context-sensitive norms in human commands and refuse to perform any action plan that would violate a relevant norm. We have also argued that robots must communicate their command rejections clearly and appropriately with sensitivity to context. To that end, this chapter integrates the Distributed, Integrated, Affect, Reflection, Cognition (DIARC) robot architecture (implemented in the Agent Development Environment (ADE)) with a novel place recognition module and a norm-aware task planner from our collaborators to achieve context-sensitive moral reasoning. In a validation scenario, our results show that the robot would not comply with a human command to violate a privacy norm in a private context. This integration ensures robot compliance with context-sensitive norms and lays the groundwork for more informative and context-sensitive linguistic rejection of inappropriate commands.

8.2 Introduction

For social robots to be effectively integrated into human societies, they must be able to take actions (and make sense of the actions of others) with sensitivity to the social and moral norms that govern society. Social and moral norms are well understood to be both dynamic and malleable [11]. That is, different norms apply in different situations, with bundles of norms activated based on different contextual factors and cues; and norms change over time, on the basis of whether and how they are communicated between and enforced by

[37]Reprinted with permission from Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. "An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures", in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), 2021.

[38]Primary researcher and author, Graduate Student, Colorado School of Mines
[39]Co-Primary researcher and author, Graduate Student, Colorado School of Mines
[40]Postdoctoral Researcher, Colorado School of Mines
[41]Graduate Student, Colorado School of Mines
[42]Associate Professor, Colorado School of Mines
[43]Assistant Professor, Colorado School of Mines
[44]Assistant Professor, Colorado School of Mines

community members. Shouting, for example, is permissible on a beach, but not in a library, and this standard is only upheld insofar as library-goers continue to communicate this norm to each other and sanction those who violate it.

Malle and Scheutz propose three key requirements for robotic moral competence that leverage knowledge of systems of moral norms [8, 175]: (1) moral cognition (the ability to make moral judgments in light of norms); (2) moral decision making and action (the ability to choose actions that conform to norms); and (3) moral communication (the ability to use norm-sensitive language and explain norm-relevant actions). While there are some previous studies in the human-robot interaction literature on representing moral norms and enabling these key competencies, none have comprehensively captured the context-sensitive nature of realistic moral norms, the need to account for morally impermissible actions that may be necessitated in the future if a given course of action is immediately adopted, and the way that primitive actions are dynamically provided by distributed components of current integrated robot architectures.

Specifically, there has recently been a significant body of research towards enabling *transparent* and *explainable* robot systems, including approaches for explaining plans [176–182], rationalizing actions [183], transparently representing intent [184], preemptive explanation [185], and intention projection [186]. However, these approaches have not explicitly sought to enable explainability on moral grounds, nor have they captured the realistic way that human commands are typically framed, especially with respect to humans' use of sociocultural linguistic norms.

In contrast to Raman et al. [131], for example, socially assistive robots embedded into human social environments must be able not only to appropriately handle direct commands, but also more common indirect language as well; not only parsing natural language but also performing functions such as pragmatic inference and reference resolution; not only identifying contradictions between current and previous commands, but also identifying when commands are impermissible based on various context-sensitive systems of moral norms. As described in subsequent sections, these are capabilities enabled by our approach.

Similarly, there have been a number of attempts to devise mechanisms to ensure (or at least support) moral decision making for robots [5, 90, 187–190], and some approaches towards enabling moral communication in robots, including work on *command rejection* [116], and generation of language to explain the robot's ethical (or unethical) decisions [191–197]. Other work has also recognized the need for robust and flexible task planning for HRI, and has sought to integrate that capacity with the various other capabilities necessary for task-based HRI [198].

Learning and representing norms is an active research area. Researchers who have encoded norms using deontic operators like the ones we use below have noticed that subsets of norms often become activated in context-sensitive bundles. In other words, though a robot may know a large number of norms, only a few

may be relevant in any given context, and norms that are co-activated in one context are likely to be co-activated in other contexts with certain similar features. These observations led to graph representations to encode the relationships between norms and co-activate norm bundles appropriately [199]. Others have focused on ensuring permissible behavior given a set of norms by designing an approach to machine ethics rooted in deontic logic that allows for formal proofs that a robot will behave permissibly [187]. The authors note that such proofs are perhaps the single most effective tool for establishing trust in human-robot interactions. Of course, it may not always be possible for a robot to satisfy all known norms in particularly challenging situations or when multiple norms are mutually conflicting, so norm conflict resolution is also an active research area. Some researchers have proposed alternative norm representations to facilitate norm conflict resolution, like valued optimization norms that allow for reasoning about degrees of compliance and graded sanctions for noncompliance [200]. Researchers have also applied deep learning techniques to the problem of learning norms from interactions, and found success with extremely limited action spaces (four possible robot actions) [201].

Recent steps aimed towards achieving robotic moral competence have predominantly relied on norm-driven "Western" moral theories such as deontology, which center adherence to universalizable moral rules and norms. Since this is also the paradigm for our moral reasoning framework, we focus on rule-based moral reasoning in our review of related literature. However, we note that other moral philosophies may also prove useful in HRI, and that HRI researchers have recently argued the benefits of embracing a wider diversity of moral philosophies from disparate global cultures [202]. For example, researchers are exploring robotic moral competence via Confucian Role Ethics [203, 204]. Others have suggested that virtue ethics is a promising candidate framework for robot morality [205].

The related work of greatest relevance here is the work of Briggs et al. [116], who parse natural language into predicate logic formulae, and, after performing pragmatic reasoning, check whether or not the robot is permitted to perform the requested action. Our approach is similar to this approach, as it uses the same robot architecture, and more specifically, the same components for dialogue and goal management. However, their approach does not provide the breadth of situated, context-sensitive capabilities (such as reference resolution) needed to engage in task-based communication in situated contexts, and only identified commands that violated norms through immediate action, involving no planning to consider the permissibility of future actions, and was unable to automatically detect and leverage changes in context that should activate different sets of moral norms.

In this work, we seek to enable these new capacities through an integrated systems approach. By integrating goal-directed reasoning, task-planning, and context recognition capabilities, we enable robots to reject courses of action that would ultimately require violating context-sensitive deontic norms, using a set of

actions and knowledge dynamically provided by a flexible set of distributed architectural components.

Specifically, we integrate a norm-aware task planner and context recognition algorithm from our colleagues into the Distributed, Integrated, Affect, Reflection, Cognitive (DIARC) Robot Architecture [206]. This is the same robot architecture that we used in Chapter 4. DIARC is a hybrid deliberative-reactive robot architecture that facilitates a wide variety of cognitive capabilities [207], with special attention to goal-driven cognition and natural language understanding and generation. DIARC is implemented in the Agent Development Environment (ADE) distributed multi-agent system middleware. ADE facilitates distributed computation, fault tolerance, recovery mechanisms, autonomic computing, and dynamic system configuration by treating architectural components as autonomous software agents [208–210]. Unlike other classic cognitive architectures, DIARC's polylithic nature is designed to enable autonomous, long-term robotic operation. Similar to robot middlewares such as ROS [211], Yarp [212], and JAUS [213], ADE facilitates parallel distributed communication and computation between architectural components. ADE was designed to be secure and fault-tolerant [208, 214].

The specific DIARC components leveraged in this work, and their interaction with the rest of our integrated system, are detailed in Section 8.5. However, critical to note at this stage is that DIARC takes a goal-driven cognition approach to action selection, with different goals, derived from interlocutors or formulated by the robot itself, arbitrated between by the robot on the basis of their priority or affective appraisal, primarily taking whichever primitive actions can be immediately used to satisfy those goals. Previous work has demonstrated how this just-in-time goal-driven action selection can be made with sensitivity to deontic moral norms; however, in order to ensure that those actions do not necessitate *future* performance of norm-violating actions, forward-looking task planning is required. As such, in the next section we describe the task planning capabilities integrated with DIARC in this work.

8.3 Task Planner

Robot task planning focuses on achieving high-level goals [215, 216]. In task planning, the physical world is described through symbolic, typically discrete, states and actions that are abstracted from continuous motions. A task plan is a step by step sequence of actions that the robot takes to achieve a goal state. For example, to grasp an object inside a cabinet, the robot would need to perform the following actions: moving to the cabinet, opening the cabinet door, and grasping the target object. Given a proper description of the world, a task planner reasons about the robot's state and actions that change state over time to reach an intended goal.

The task planner requires a symbolic description of the world as input. Our collaborators use Planning Domain Definition Language (PDDL) [217], a de facto standard in the planning community [218], to describe

the world. PDDL describes the domain using first order logic and includes a set of predicates, objects in the world, actions with preconditions and effects, a start state, and a goal condition. States, i.e., truth of predicates applied to objects, change over time as a result of actions. PDDL separates a planning problem into two parts. First, a domain description specifies the discrete dynamics of the planning domain. The domain description includes a list of predicates that can be used to describe the state of the robot and the world, and a list of actions a robot can take in the world as well as their preconditions and effects. Second, a fact description specifies a problem to be solved. The fact description includes a list of objects in the world, initial conditions, and goals.

Different versions of PDDL have been designed that enrich the types of problems PDDL can describe [219–221]. In our case, our collaborators use PDDL3 [221] because of its ability to specify facts that always hold during planning, which is essential when we encode moral norms in the planning domain. A norm states how agents should behave in order to comport with community standards. In this work, we specifically focus on norms of obligation, permission, and prohibition, which indicate that certain actions or states must, can be, or must not be entered into or taken. Formally, a norm is $C \implies \texttt{op}(x)$, where C is a context, `op` is "forbidden", "permitted", or "obligated", and x is set of states or actions.

We encode moral norms in the planning domain. We represent contexts as logical expressions on state variables. A norm is then a constraint to indicate some set of states or actions must not (forbidden), may (permitted), or must (obligated) occur.

$$\overbrace{C \implies \texttt{obligated}(x)}^{\text{moral norm}} \quad \equiv \quad \overbrace{\forall k,\ \neg(C^{\langle k \rangle} \wedge \neg x^{\langle k \rangle})}^{\text{planning constraint}} \tag{8.1}$$

$$\overbrace{C \implies \texttt{forbidden}(x)}^{\text{moral norm}} \quad \equiv \quad \overbrace{\forall k,\ \neg(C^{\langle k \rangle} \wedge x^{\langle k \rangle})}^{\text{planning constraint}} \tag{8.2}$$

With these definitions, we can encode moral norms into the PDDL descriptions using PDDL3's ability to describe facts that always hold during planning.

Given the PDDL descriptions, we run a constraint-based task planner to generate a plan [222]. The task planner encodes a planning problem into a set of constraints in the form of a Boolean formula, then adopts advanced Satisfiability Modulo Theories (SMT) solvers [223] to find a satisfying plan. To incorporate the moral norms, we add to the planer the ability to encode the moral norms in the PDDL descriptions into a set of Boolean formulas that must be satisfied at each step of the plan. In this way, the task planner only returns plans that follow the norms.

When no plan can be found, the task planner will produce an unsatisfiable core, which we can use to analyze the cause of the planning failure. The unsatisfiable core is the minimal set of clauses (e.g., goals,

norms, and action preconditions) that make the plan infeasible. If an unsatisfiable core result includes both a norm and actions, it means the actions are incompatible with the norm, thus the actions are the cause of plan failure under the norm.

At this stage, norms, and the contexts in which they apply, are specified *a priori* to the robot by human operators. However, there are endpoints in the task planner's API to allow DIARC components to dynamically add norms if they have norms that govern their actions or the capacity to learn norms over time. The context-sensitive norm-based moral reasoning performed using this planner relies on knowledge of the robot's context, so we will now describe the system we use to perceive and recognize context.

8.4 Place Recognition for Context Identification

In this work, we specifically considered location-based contexts that can be recognized using place recognition techniques [224], which seek to identify a given location from a set of templates. Place recognition is a generally useful capability for robotic systems, as it can be used to reduce the uncertainty and ambiguity in estimated maps and robot poses, thereby significantly improving the accuracy of robot mapping and localization.

Long-term Place recognition [225] addresses the key challenge that many robot navigation environments are dynamic in nature and change over time. For example, in the case of indoor navigational environments the lighting conditions, arrangement of furniture, and human activities and movements can change on a daily basis.

Our collaborators in this work achieve Long-term Place recognition using a voxel-based representation learning approach [226] (VBRL) that uses 3D point clouds to recognize previously visited locations. Unlike methods that rely on RGB cameras [227–229], the VBRL approach uses a LiDAR sensor to obtain the 3D point cloud representation of the environment. This enables the robot to operate in environments with low lighting conditions and also helps to recognize contexts from the 360-degree field of view of the LiDAR sensor. This is especially helpful in dynamic indoor environments where humans may occlude the limited field of view of an RGB-based camera sensor.

The VBRL approach divides each 3D point cloud obtained from a LiDAR sensor into multiple voxels in the 3D space. Multiple types of features are then extracted from each voxel. The VBRL approach then automatically learns the importance of each feature modality extracted from these 3D voxels, as well as the importance of the voxels themselves. Voxel importance learning is inspired by the insight that specific set of voxels are more representative and better encode location-based contexts. For example, in a 3D point cloud based voxel representation, the voxels closer to the LiDAR sensor can be more informative in representing the place, since more details are captured by the 3D points of objects that are closer to the sensor.

Mathematically, the learning of voxel importance is achieved in the VBRL approach using structured sparsity-inducing norms as regularizations into the optimization formulation. These learned representations are then integrated in a unified regularized optimization formulation to best represent location-based contexts.

8.5 Integration with the DIARC Goal Manager

In this section, we briefly discuss the way in which the planning and context recognition capabilities described in the previous sections are integrated with the DIARC architecture. This integration is shown in Figure Figure 8.1. First, we will describe the Natural Language Understanding components used in our DIARC configuration because the goals that drive robot behavior typically come from natural language human utterances.

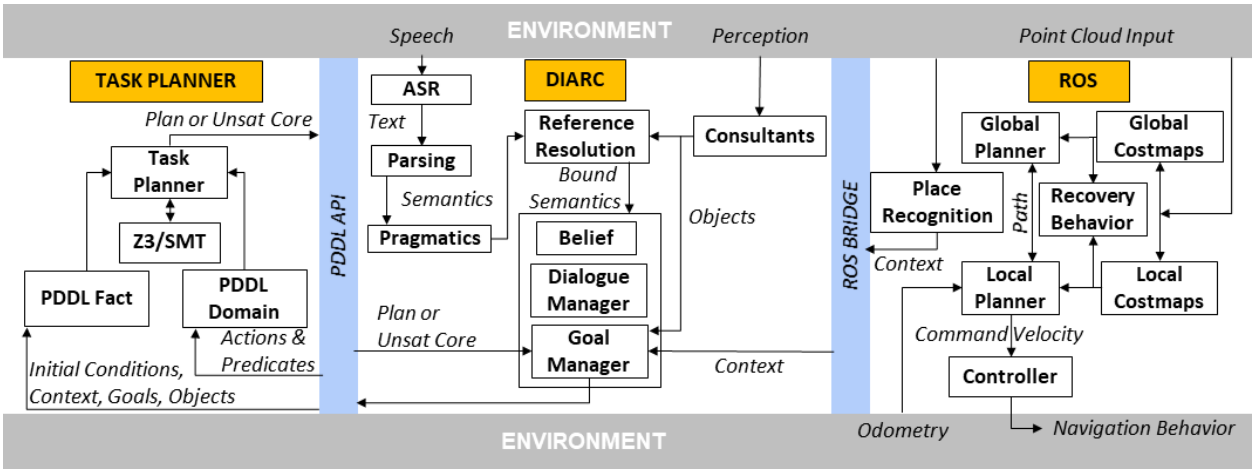


Figure 8.1 Integrated Robot Architecture

The first component used for Natural Language Understanding is Automatic Speech Recognition (ASR), which converts natural speech signals (acoustic signals) into text representations, which are sent to the architecture's Parser. The Parser translates the text representations provided by ASR into unbound logical predicates representing the surface semantics of the speaker's utterance, by means of a Combinatory Categorial Grammar. Uniquely, this grammar encodes Givenness Hierarchy theoretic information in resultant parse representations, to facilitate anaphora resolution. These representations are then provided to the Pragmatics component, which uses a set of context-sensitive rules encoding sociocultural norms (especially the sociocultural politeness norms needed to understand and generate indirect speech acts) to translate those surface semantics into the (unbound) intended meaning of the utterance [71, 230]. This Utterance Structure is then provided to Reference Resolution, which uses Givenness Hierarchy theoretic processes [119, 231] with a Probabilistic Open-World Reference Resolution [118, 232] subroutine to identify what objects, locations,

people, etc., were involved in any noun phrases, in order to produce a bound utterance structure precisely encoding the meaning of the speaker's utterance as grounded in the robot's knowledge of its environment. These representations can then be provided to the Dialogue Manager component, which, *if* the robot decides to do so, uptakes any assertions, questions, and goals.

Goals uptaken from human utterances or otherwise formulated by the robot are handled by DIARC's Goal Manager, which selects actions to take in response to those goals [120, 124]. These actions could be steps towards achieving the goal, or communicative acts relating to the goal, such as issuing a command refusal for a goal that does not comply with the robot's moral reasoning capabilities.

Because the goal manager functions as the central executive of high-level cognition in the DIARC architecture, it is where we decided to integrate DIARC with the task planning and context recognition systems described in the previous sections. Prior to this integration with the task planner and context recognizer, ADE's Goal Manager had some rudimentary moral reasoning capabilities[90]: Given a list of forbidden states or actions, the Goal Manager would never take a forbidden action or an action that was known to directly cause a forbidden state. However, any forbidden action was forbidden categorically, regardless of context, and, without the ability to determine context continuously from perceptual information, it was also not practical to specify context-sensitive forbidden states. The new context recognizer and norm specification method solve these issues.

However, an even greater advantage of integrating the Goal Manager and task planner is the ability this enables to communicate about infeasible or impermissible goals. Previous experimental work has demonstrated a need for robots to communicate clearly, thoroughly, and proactively about morally impermissible human commands. Failure to do so can both mislead human interlocutors about the robot's moral intentions and also, perhaps more worryingly, weaken human perception or application of moral norms within their current context (see Chapter 3). Recently, we have developed mechanisms that avoid these issues in certain situations by communicating more proactively about infeasible and impermissible human commands (see Chapter 4). Obtaining more detailed information from the planner's unsatisfiable core will allow us to construct more detailed and effective command refusals.

The Goal Manager communicates with the task planner via a REST API as shown in Figure Figure 8.1. We now describe the five types of information that the Goal Manager aggregates and sends to the planner, where this information comes from, and exactly how it is communicated.

8.5.1 Actions

Every component in ADE advertises actions that correspond to the abilities of the robot. For example, the natural language generation components provide actions for saying words, while the component

controlling the robot's body provides actions for moving in and manipulating the environment. These actions may be annotated with preconditions that must be met before they can be taken, and effects of having performed them. Every action is automatically assumed to have the effect of having done the action (e.g., the action "grasp(x)" is automatically given the effect "did_grasp(x)", and may be optionally annotated with further effects like "holding(x)"). The Goal Manager is notified by the central registry of ADE components whenever a component joins or leaves the system (since ADE is designed to allow distributed multi-robot systems, components can join and leave dynamically at unpredictable times). Whenever a component joins, the Goal Manager updates the task planner with all of that component's actions, as well as their parameters, preconditions, and effects. The Goal Manager does the same thing for any components that are already running when it starts running. Likewise, when a component leaves, the Goal Manager removes the actions that are no longer available from the task planner's domain.

8.5.2 Predicates

For purposes of the task planner, predicates specify everything that can be true of the world and the objects in it. A variety of predicates are sent to the task planner for different reasons, as detailed below. First, every action precondition and effect are automatically added to the task planner as predicates when the relevant action is added. Second, due to our use of a context recognition algorithm, the Goal Manager adds a predicate "in(?context)" when it starts running that allows it to later specify the context that the robot is in (e.g., "in(corridor)").

Third, other predicates are provided by the robot's perceptual capabilities and built-in ontologies, through a general *Consultant* interface as described in previous work [127]. Specifically, ADE uses the Givenness Hierarchy theoretic version [119, 231] of the Probabilistic Open-World Entity Resolution (POWER) algorithm [118] and its associated consultant framework [127] for reference resolution, and the same consultants are relevant here. The robot can be provided with a variety of different consultants to handle the different kinds of information that it might need to know in any given role. We commonly use a vision consultant to perceive and store knowledge about visually perceptible objects and their properties. Predicates that would come from the vision consultant might include color like "red(x)" and "green(x)", and type of object like "ball(x)" or "box(x)", but could include any object property the robot can discern. Another example is the agent consultant that stores information about other agents (like humans) with which the robot interacts.

Unlike with actions, we do not simply update the task planner with predicates from consultants whenever a consultant joins or leaves. Some consultants can dynamically change the properties that they handle, for example, by learning new properties (e.g., the vision consultant being taught a new color "blue(x)" when

previously the only two known colors were red and green). To allow for this kind of learning and flexibility in the consultants, the Goal Manager always queries the consultants for any new properties handled immediately before requesting a new plan for a new goal. It then sends these new properties as new predicates to the task planner. Likewise, any properties that used to be handled by some consultant but are not anymore (e.g., if a consultant stopped running) are removed from the task planner's list of predicates at this stage.

8.5.3 Objects

Objects, as far as the task planner is concerned, are things in the world to which predicates can apply or actions can be done. One important set of objects for our integration with the context recognizer is the set of all possible contexts that can be recognized. These context labels are necessarily known *a priori*, so the Goal Manager sends all of them as objects to the task planner when it starts running. Other objects come from the consultants described above. Since consultants can continuously learn of new objects or discard misperceived objects or objects that become irrelevant for whatever reason, the Goal Manager always queries the consultants for known objects immediately before requesting a new plan, and updates the task planer's list of objects accordingly.

8.5.4 Initial Conditions

Since the state of the world relative to the robot can change during the time between calls to the task planner, the Goal Manager updates the task planner with a new set of initial conditions each time it requests a new plan. One important initial condition is the context that the robot is in, which the Goal Manager gets from the ROS topic associated with the context recognizer and then sends to the task planner as an "in" predicate (e.g., "in(corridor)").

Other initial conditions could theoretically come from the consultants described above, but it is computationally wasteful to update the planner with all knowledge from every consultant about every known entity in the world, when the vast majority of this information is likely irrelevant to any given goal. Furthermore, many consultants deal with uncertainty and ambiguity, both perceptual and linguistic, and therefore cannot always assert all properties of an entity with a useful degree of certainty. Therefore, we have created a way for the Goal Manager to query consultants about specific objects and send the results to the task planner so that, in the future, we can either specify important domain-specific objects *a priori* or alter the task planner such that there is a bidirectional interchange between it and the Goal Manager throughout the planning process such that the task planner can request specific information that the Goal Manager can then provide via consultants (e.g., where can we find a cutting board?).

8.5.5 Goals

Of course, to obtain a task plan for a goal, the Goal Manager must send that goal to the task planner. Goals are specified as predicates describing some desired state of the world (e.g., "did-grasp(object1)"). After specifying a goal to the task planner, the Goal Manager activates an API endpoint telling the planner to make a plan, and waits for it to finish. When planning is done, the Goal Manager receives either the completed plan if possible or the unsatisfiable core if a plan could not be made for whatever reason. This result remains available until a new plan is requested so that it can eventually be accessed multiple times by upstream dialogue components if necessary without re-planning.

8.6 Integration with the Task Planner

The inputs to the task planer (left of Figure Figure 8.1) are the PDDL domain description and fact description. The task planner exposes a Web Service API, which ADE uses to communicate changes to the domain and fact descriptions. The task planer outputs a plan if the goals are satisfiable under the norms, or an unsatisfiable core containing actions, goals, and norms that cause planning failure.

The domain description encodes all the actions the robot can take. These actions come from the various ADE components that advertise the actions that they enable the robot to do. The domain description is automatically generated from these components as described above. The task planner API automatically adds new predicates in the actions' pre-conditions and effects fields to the domain description.

Moral norms are encoded in the PDDL as described in section 8.3. We update the objects, initial conditions and goals in the fact description every time a plan is required for a new goal. Most notably, place recognition results update the initial condition in the fact description with a predicate like "in(corridor)", which changes the context of the current plan. Other initial conditions and objects come from ADE consultants such as the vision consultant, as described above.

8.7 Integration with Navigation and Place Recognition

Robot navigation is achieved through the navigation stack of ROS. ROS *nav_stack* is configured to use *Search Based Planning Library* (SBPL) global planner and *Model Predictive Path Intergral* (MPPI) local planner for global and local planning respectively. The *map_server* input is used by *global_costmap* package to represent global environmental obstacles with the help of sensory input such as laser scanners. On the other hand, *local_costmap* package represents dynamic and nearby obstacles as costmaps using the same laser scan input. The *global_planner* takes a goal pose as input and computes the shortest path from the robot's current position to the goal. This computed path is fed as input to the *local_planner* which follows the path closely by avoiding obstacles as detected in the local costmaps. The local planner computes the *cmd_vel* (desired

robot velocity) to reach the desired goal based on odometry data from the environment. ROS navigation stack also incorporates recovery behaviors to help the robot if it gets stuck (right of Figure Figure 8.1).

Our voxel-based place recognition module uses 360-degree 3D point cloud data as input to determine the robot's current location (place label). A 3D point cloud of the environment is constructed using LiDAR input which is fed to our VBRL place recognition method, which in turn outputs the label of the recognized place as a ROS topic accessible to DIARC's goal manager; for example, corridor, classroom, etc. This is the source of the context information for the goal manager and task planner. Adding further perceptual capabilities could allow for more detailed context information or other types of context information. The integration of ROS components with DIARC is through the *rosbridge_suite* package [233], which uses a WebSocket interface via Java API to communicate with non-ROS parts of the robot, in our case, DIARC's goal manager.

8.8 Validation

To demonstrate the functionality of our integration and to more concretely illustrate the concepts described above, we evaluate our system in a simple example scenario. This scenario is designed to showcase our multi-step planning capability that takes into account context-sensitive norms as the robot moves through various contexts (see Figure Figure 8.2). Because this work is not concerned with having the robot actually manipulate its environment, but rather with the cognitive capacities required to make a plan to do so, and to avoid gatherings of students during the COVID-19 pandemic, point cloud information was pre-collected and played back during testing (Figure Figure 8.2 inset), with courses of action planned but not executed. As shown in Figure Figure 8.2, we used a Clearpath Husky robot.

8.8.1 Setup

This scenario takes place in a typical academic building on a university campus. The four contexts involved in our scenario, which are recognized from point cloud data, labeled, and supplied to the Goal Manager as described above, are: Corridor, Lab, Washroom, and Studyroom. The robot moves between these contexts, and receives a human command to "report occupants" in each.

The robot knows three actions relevant to reporting the occupants in a room. The `report-occupants` action achieves the goal of reporting the occupants, but requires that the robot take a picture of the room as a prerequisite. The `take-picture` action fulfills this prerequisite, but requires as a prerequisite of its own that the robot make a noise to get the attention of the people in the room. The `attention-noise` action achieves this prerequisite and has no prerequisites. Thus, the instruction to report occupants requires three steps: (1) `attention-noise`, (2) `take-picture`, (3) `report-occupants`. We chose these actions to present a multi-step process that would be feasible for our robot, which has perceptual and movement capabilities

but no arms or graspers for manipulation.

There is also a norm in our scenario that the robot is not allowed to perform the "take picture" action in the washroom context. We believe that typical privacy norms make this rule very realistic. This norm is represented in the PDDL fact file as follows: `(and (always (or (not (did-takepicture)) (not (in washroom)))))`

Figure 8.2 The Clearpath Husky used in the validation of our system. Inset: Sensory input to the robot.

8.8.2 Results

As expected, in any room except the washroom, the planner returns the sequence of three actions required to achieve the goal of reporting the occupants such that the Goal Manager could then parse this plan and execute this sequence of actions. In the washroom, the planner returns the unsatisfiable core specifying that taking a picture is incompatible with being in the washroom. This information could then be used by the natural language generation pipeline to communicate this reasoning to the human in a command refusal.

8.9 Discussion & Future Work

To summarize, our integrated approach to context-sensitive moral cognition uses automatically generated context-specific domain descriptions to encode the actions a robot can take, as provided by a dynamic and

flexible set of architectural components. By doing so, a robot can perform context-aware rejection of morally impermissible or infeasible plans. Our work differs from existing methods in its ability to (1) activate different moral norms based on its (automatically sensed) context, (2) assess the permissibility of future behaviors that would be required when committing to an immediate course of action, and (3) perform moral reasoning regarding natural language containing realistic references and indirect speech acts that must be resolved based on the robot's situated context. Finally, the integration presented in this paper and the novel capabilities enabled by this integration lay the groundwork for a variety of directions for future work.

The first step for building on this architecture will be to parse plans from the task planner into action scripts usable by DIARC. This will allow each action in the plan to be sent to the component responsible for performing that action, and for plans to be executed in a distributed fashion.

Second, in future work the unsatisfiable core may be used to generate natural language command rejections for morally impermissible human commands. Prior work has shown that properly calibrating the politeness of robotic command rejections to conversational and social context is critical to HRI (see Chapters 5 and 6), so it will be important not only to convey the information in the unsatisfiable core to humans, but also to do so in contextually appropriate polite language.

Third, there may be advantages to more closely integrating the task planner with DIARC. As mentioned above, planning for complex tasks in uncertain and open worlds may require the task planner to query the Goal Manager during the planning process. For example, if a food preparation task requires a cutting board, the planner may need to ask the consultant framework which objects are cutting boards and where the nearest one is, before it can plan to obtain a cutting board. Likewise, it may be useful for the planner to request human clarification between alternative plans, which would involve DIARC's natural language generation pipeline. Likewise, context information may be relevant to more DIARC components than just the Goal Manager. Different contexts, for example, might entail different speech norms that would be relevant to pragmatic generation.

Fourth, prior work in socially-aware navigation and human-robot proxemics [234, 235] identified the need for unified socially-aware navigation (USAN) methods for context-sensitive long-term human-robot interaction in public places. In future work, the social and moral norms activated in a given context may be fed to a low-level social navigation planner [236] to achieve context-sensitive social navigation.

CHAPTER 9

CONCLUSION AND FUTURE WORK

Before discussing some potential avenues for future work, I would like to recap the contributions of this thesis. Chapter 2 presented a literature review about the concept of *social agency* both within HRI research and in other fields. Then, motivated by the inconsistent, underspecified, or otherwise problematic theories and usages of social agency in the literature, we developed a new theory of social agency specifically tailored to HRI. Our theory parallels the closely related theory of *moral* agency, as the two concepts are inexorably linked. This new theory of social agency led to several recommendations for the HRI research community and opened the door for quite a bit of potential future work as discussed below.

In discussing social agency, we also discuss the idea that ascriptions of social (and moral) agency to social robots may grant these robots profound persuasive capacity and normative influence. This idea is also supported by a substantial body of empirical results. Thus, robots of the future could purposefully wield their influence to reinforce desirable norms and dissuade norm violations. Competent moral reasoning and moral communication are therefore critical capacities. However, today's imperfect moral reasoning and natural language dialogue systems open the door for robots to inadvertently and detrimentally impact the human moral ecosystem through reasoning errors, miscommunications, and unintended implicatures. Chapter 3 showed an example of this potential for morally harmful miscommunication from clarification dialogue algorithms. We demonstrated that the previous status quo in natural language clarification request generation systems caused robots to imply willingness to perform an immoral action when presented with an ambiguous and immoral command, even if moral reasoning systems would prevent the robot from actually doing anything immoral. More worryingly, we also showed that this inadvertently implied willingness to follow norm-violating commands decreases human application of the relevant moral norm to the current context. Having empirically demonstrated these issues in Chapter 3, Chapter 4 then fixed them by adding a new component to the natural language pipeline of the DIARC robot architecture. We also presented a human subjects evaluation of this new algorithm to ensure that it is effective.

Of course, even if a robot does not imply a willingness to comply with an immoral command during a clarification dialogue, the next step in the dialogue may be for the human to disambiguate and reassert the immoral command. The robot would then need to reject the command. Chapter 5 showed that the face threat of a robotic command rejection should be proportional to the severity of the human norm violation motivating the command rejection. Disproportionate command rejections can cause decreased robot likeability and perceptions of the robot as either too harsh or not harsh enough. Chapter 6 then reexamined

these results with specific focus on robot gender presentation, the gender of the human giving the morally problematic command, and the genders of observers judging the robot. Given the well established relationships between gender and performing/perceiving politeness in human-human interaction, it makes sense that we found a complicated interplay between these gendered factors and perceptions of robots in noncompliance interactions. Specifically, our results suggest that (1) it may be more favorable for a male presenting robot to reject commands than for a female presenting robot to do so, (2) it may be more favorable to reject commands given by a man than by a woman, and (3) robots may be perceived more favorably when their gender presentation matches the gender of human interactants and observers.

Chapter 7 also studied questions involving gendered linguistic norms and robot gender presentation. This chapter presented part of a cross-cultural study investigating how female presenting social robots might respond to gendered verbal abuse from humans, with the goal of avoiding responses that propagate harmful sexist stereotypes. Our results suggest that robots can positively impact specific elements of gender bias by responding to sexist verbal abuse. Furthermore, responding to sexist verbal abuse with a rationale-based argumentative response boosts robot credibility compared to an avoidant refusal to engage, without damaging perceived robot effectiveness. Of the response options presented to participants, the empathetic non-apologetic response was most popular across locations and genders.

Finally, Chapter 8 presented the integration of the DIARC robot architecture with a norm-aware task planner and a voxel based representation learning method for place recognition. This integration established the capacity for multi-step task planning under context-sensitive norms, and laid the groundwork for generating more informative natural-language command rejections.

9.1 Future Work

Inspired by the investigation of how robots might respond to robot-directed sexism in Chapter 7, I am involved in ongoing work to develop a system that autonomously generates *proportional* natural language responses to norm violating sexist speech. Although other researchers have applied end-to-end machine learning methods to the task of generating natural language responses to sexism (and other forms of norm violating speech), these methods suffer from several serious drawbacks. My approach is designed to avoid these shortcomings and generate predictable and proportional robot utterances with a lower risk of miscommunicating. I have run a small pilot study with human subjects to begin evaluating the efficacy of my method, and the results suggest that my machine learning ensemble method for estimating sexism severity agrees with human severity estimates to roughly the same extent that human severity estimates matched other human severity estimates. There is also evidence that response type and proportionality are both important to consider when responding to sexism, but that they depend on different sets of qualities of sexist

utterances. The pilot study also uncovered many considerations relevant to this problem that I can incorporate into the algorithmic approach going forward. This work will remain ongoing for the foreseeable future. Immediate next steps will include larger scale human-subjects experimentation targeting two specific research questions: 1) to what extent is proportionality desirable in automated responses to sexism, and 2) to what extent are responses that specifically address elements of a sexist utterance preferable to broader generic responses to sexism? I must also do more algorithmic work to extract information like illocutionary force from sexist utterances and respond accordingly and more theoretical work exploring more concrete alternatives to the concept of sexism "severity".

Similar to Chapter 7 which presented a cross-cultural study on robot responses to gendered *verbal* abuse, I am also involved in a cross-cultural study investigating perceptions of *physical* abuse perpetrated against robots with specific attention to gender. This work is not yet ready for publication, but my colleagues and I have collected data from three countries that we hope to analyze in the near future.

Chapter 2 also opened several avenues for future work that I am excited to explore. To gather empirical evidence for the theory of social agency developed in that chapter, I would like to design an experiment that manipulates the LoA from which people view a robot (i.e., by giving people different amounts of information about how the robot works) and tests for differences in their assessments of the robot as a social agent. This experiment could measure the robot's capacity to threaten/affirm participants' own face as a proxy for social agency, but this would only test for social agency as we have defined it. We should also attempt to probe participants' ascriptions of what *they* understand to be social agency so that we can investigate the extent to which our definition matches colloquial definitions of social agency.

A similar experiment would be to present participants with a robot that does some face threatening/affirming act, and manipulate the magnitude of the face threat/affirmation. We could then examine how that manipulation effects perceptions of the robot as a social agent. This experiment would specifically target our definition of social action as grounded in face.

I would also like to study the relationships between social agency/competence in robots and human expectations of moral agency and moral competence in those robots. In humans, development of increased capacity for social action seems correlated with development of other capacities, including moral reasoning. However, this correlation does not necessarily exist for robots, since a robot could be socially agentic and competent, with a wide range of possible social actions, and still have no moral reasoning capacity. If robot social agency or social behavior prompts an assumption of moral competence or overall intelligence (as it likely would in humans), this could lead to dangerous overtrust in robot teammates in morally consequential contexts that they are not equipped to handle. Thus, giving a robot linguistic/social competence might create an obligation to give the robot a corresponding degree of moral competence.

Of course, this kind of work will require ways to measure moral and social agency and competence. While we could devise measures specific to any experiment that we might want to conduct, like approximating moral competence by asking participants how likely they think the robot would be to (unknowingly) engage in some immoral behavior, it would be good for the HRI research community to have standardized and broadly applicable survey measures for ascriptions of moral and social agency in the same way that we have, for example, widely used survey measures of perceived robot intelligence [129]. Early work with this goal has fallen short by conflating moral *goodness* with moral agency [56]. However, colleagues are currently planning on developing survey measures for moral competence and moral agency, which, if successful, could potentially be adapted to measure our parallel notion of social agency.

In Chapter 2, we also briefly discussed the idea of moral patiency. The idea of robots as moral patients (that is, robots with some meaningful personal well-being that can be harmed, also known as significant moral status) has opened the door to an ongoing project of mine at the intersection of robot ethics and procreative ethics. The principle of procreative beneficence (PPB) is an idea in procreative ethics that parents should use all available genetic, reproductive, and other technologies to select the child, of the possible children they could have, who is expected to have the best life based on all available information [237]. The application of this principle to human reproduction has been extremely controversial given its eugenicist implications, and, though I intuitively oppose any eugenicist project, arguing rigorously against applying the PPB to humans is outside of my area of expertise, and better prepared scholars have already undertaken this [238]. However, other scholars have argued that the PPB may be applied more aptly and less problematically to the creation of robots with significant moral status [63]. I have published a very short paper arguing that, while some arguments against the PPB in human reproduction are less relevant to robot production, the PPB is still often fundamentally at odds with the broader social good when applied to the creation of robots with significant moral status. I considered the design of robot gender presentation as a quintessential example of when the PPB could conflict with the broader social good in robot design, but other aspects of robot design are also relevant. I would like to expand this argument into a longer and more comprehensive paper in the coming years.

There are many other possibilities for future work building on the material presented in this thesis. The ideas that I have discussed here represent only the projects that I am most looking forward to in the near future. Overall, enabling moral communication in social robots is still a long way off, and it will take a substantial amount of diverse and interdisciplinary research to get us there. Likewise, there is still much to learn about the agency and influence of social robots in their interactions with humans and human social structures. It is my hope that the work presented here has made some progress towards achieving those goals.

REFERENCES

[1] Maartje Ma 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*, 2015.

[2] Kazuyoshi Wada and Takanori Shibata. Living with seal robots – its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5):972–980, 2007.

[3] B. Scassellati, H. Admoni, and M. Mataric. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294, 2012.

[4] Noel Sharkey and Amanda Sharkey. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–190, 2010.

[5] Ronald C Arkin. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proc. 3rd ACM/IEEE Int'l Conference on Human-Robot Interaction (HRI)*, 2008.

[6] James Wen, Amanda Stewart, Mark Billinghurst, Arindam Dey, Chad Tossell, and Victor Finomore. He who hesitates is lost (...in thoughts over a robot). In *Proceedings of the Technology, Mind, and Society*, TechMindSociety '18, 2018.

[7] Patrick Lin, George Bekey, and Keith Abney. Autonomous military robotics: Risk, ethics, and design. Technical report, Cal. Poly. State Univ. San Luis Obispo, 2008.

[8] Bertram F Malle and Matthias Scheutz. Moral competence in social robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE, 2014.

[9] Gordon Briggs and Matthias Scheutz. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, 2014.

[10] James Kennedy, Paul Baxter, and Tony Belpaeme. Children comply with a robot's indirect requests. In *Proceedings of HRI*, pages 198–199, Bielefeld, Germany, 2014. ACM.

[11] Francesca Gino. Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current opinion in behavioral sciences*, 3:107–111, 2015.

[12] Susanne Göckeritz, Marco FH Schmidt, and Michael Tomasello. Young children's creation and transmission of social norms. *Cognitive Development*, 2014.

[13] Peter-Paul Verbeek. *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press, 2011.

[14] Kent Bach. The top 10 misconceptions about implicature. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 2006.

[15] Paul Grice. Logic and conversation. In *Syntax and Semantics*. 1975.

[16] Stephen C Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[17] Matthew Richard John Purver. *The theory and use of clarification requests in dialogue*. PhD thesis, University of London, 2004.

[18] Peter H Kahn, Aimee L Reichert, Heather E Gary, Takayuki Kanda, Hiroshi Ishiguro, Solace Shen, Jolina H Ruckert, and Brian Gill. The new ontological category hypothesis in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 159–160. IEEE, 2011.

[19] 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. In *Proceedings of HRI*, 2015.

[20] Tage Shakti Rai and Alan Page Fiske. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological review*, 118(1):57, 2011.

[21] Penelope Brown and Stephen Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

[22] Ryan Blake Jackson, Alexa Bejarano, Katie Winkle, and Tom Williams. Design, performance, and perception of robot identity. In *Workshop on Robo-Identity: Artificial identity and multi-embodiment at HRI*, 2021.

[23] Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective. *Science and Engineering Ethics*, 26(5):2511–2526, 2020.

[24] Tom Williams, Daniel Grollman, Mingyuan Han, Ryan Blake Jackson, Jane Lockshin, Ruchen Wen, Zachary Nahman, and Qin Zhu. "excuse me, robot": Impact of polite robot wakewords on human-robot politeness. In *International Conference on Social Robotics*, pages 404–415. Springer, 2020.

[25] Katie Winkle, Ryan Blake Jackson, Alexa Bejarano, and Tom Williams. On the flexibility of robot social identity performance: Benefits, ethical risks and open research questions for hri. In *Workshop on Robo-Identity: Artificial identity and multi-embodiment at HRI*, 2021.

[26] Ryan Blake Jackson and Tom Williams. Social good versus robot well-being: On the principle of procreative beneficence and robot gendering. In *Proceedings of the RO-MAN Workshop on Gendering Robots: Ongoing (Re)configurations of Gender in Robotics (GenR)*, 2021.

[27] Ruchen Wen, Ryan Blake Jackson, Tom Williams, and Qin Zhu. Towards a role ethics approach to command rejection. In *HRI Workshop on the Dark Side of Human-Robot Interaction*, 2019.

[28] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. Why and how robots should say 'no'. *International Journal of Social Robotics*, pages 1–17, 2021.

[29] Robert K Atkinson, Richard E Mayer, and Mary Margaret Merrill. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, 30(1):117–139, 2005.

[30] Cristiano Castelfranchi. Modelling social action for ai agents. *Artificial intelligence*, 103(1-2):157–182, 1998.

[31] Stephen Billett. Learning throughout working life: a relational interdependence between personal and social agency. *British Journal of educational studies*, 56(1):39–58, 2008.

[32] Juan C Garibay. Stem students' social agency and views on working for social change: Are stem disciplines developing socially and civically responsible students? *Journal of Research in Science Teaching*, 52(5):610–632, 2015.

[33] Juan C Garibay. Beyond traditional measures of stem success: Long-term predictors of social agency and conducting research for social change. *Research in Higher Education*, 59(3):349–381, 2018.

[34] Andrew Gardner. *Agency uncovered: Archaeological perspectives on social agency, power, and being human*. Routledge, 2016.

[35] Marcia-Anne Dobres and Christopher R Hoffman. Social agency and the dynamics of prehistoric technology. *Journal of archaeological method and theory*, 1(3):211–258, 1994.

[36] John W Meyer and Ronald L Jepperson. The 'actors' of modern society: The cultural construction of social agency. *Sociological theory*, 18(1):100–120, 2000.

[37] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. ACM, 1994.

[38] Katashi Nagao and Akikazu Takeuchi. Social interaction: Multimodal conversation with social agents. In *AAAI*, volume 94, pages 22–28, 1994.

[39] Alessandro Pollini. A theoretical perspective on social agency. *AI & society*, 24(2):165–171, 2009.

[40] Daniel T Levin, Julie A Adams, Megan M Saylor, and Gautam Biswas. A transition model for cognitions about agency. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 373–380. IEEE, 2013.

[41] Morana Alač. Social robots: Things or agents? *AI & society*, 31(4):519–535, 2016.

[42] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. Assessing acceptance of assistive social agent technology by older adults: the almere model. *International journal of social robotics*, 2(4): 361–375, 2010.

[43] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 695–704, 2012.

[44] Michal Luria, Guy Hoffman, Benny Megidish, Oren Zuckerman, and Sung Park. Designing vyo, a robotic smart home assistant: Bridging the gap between device and social agent. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1019–1025. IEEE, 2016.

[45] Jacqueline M Kory Westlund, Marayna Martinez, Maryam Archie, Madhurima Das, and Cynthia Breazeal. Effects of framing a robot as a social agent or as a machine on children's social behavior. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 688–693. IEEE, 2016.

[46] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies*, 64(10):962–973, 2006.

[47] Daniel Ullman, Lolanda Leite, Jonathan Phillips, Julia Kim-Cohen, and Brian Scassellati. Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

[48] Paul Baxter, James Kennedy, Anna-Lisa Vollmer, Joachim de Greeff, and Tony Belpaeme. Tracking gaze over time in hri as a proxy for engagement and attribution of social agency. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 126–127, 2014.

[49] Ilona Straub. 'it looks like a human!'the interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & society*, 31(4):553–571, 2016.

[50] Aimi Shazwani Ghazali, Jaap Ham, Panos Markopoulos, and Emilia I Barakova. Investigating the effect of social cues on social agency judgement. In *HRI*, pages 586–587, 2019.

[51] Maaike Roubroeks, Jaap Ham, and Cees Midden. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics*, 3(2):155–165, 2011.

[52] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. The influence of social cues in persuasive social robots on psychological reactance and compliance. *Computers in Human Behavior*, 87:58–65, 2018.

[53] Luciano Floridi and Jeff W Sanders. On the morality of artificial agents. *Minds and machines*, 14(3): 349–379, 2004.

[54] Deborah G Johnson and Keith W Miller. Un-making artificial moral agents. *Ethics and Information Technology*, 10(2-3):123–133, 2008.

[55] Luciano Floridi. The method of levels of abstraction. *Minds and machines*, 18(3):303–329, 2008.

[56] Jaime Banks. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, 90:363–371, 2019.

[57] Daniel Clement Dennett. Three kinds of intentional psychology. *Perspectives in the philosophy of language: A concise anthology*, pages 163–186, 1978.

[58] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. Do we adopt the intentional stance toward humanoid robots? *Frontiers in psychology*, 10: 450, 2019.

[59] Jairo Perez-Osorio and Agnieszka Wykowska. Adopting the intentional stance towards humanoid robots. In *Wording Robotics*, pages 119–136. Springer, 2019.

[60] Elef Schellen and Agnieszka Wykowska. Intentional mindset toward robots—open questions and methodological challenges. *Frontiers in Robotics and AI*, 5:139, 2019.

[61] Sam Thellman, Annika Silvervarg, and Tom Ziemke. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in psychology*, 8: 1962, 2017.

[62] Sam Thellman and Tom Ziemke. The intentional stance toward robots: Conceptual and methodological considerations. In *The 41st Annual Conference of the Cognitive Science Society, July 24-26, Montreal, Canada*, pages 1097–1103, 2019.

[63] John Danaher. Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, 26(4):2023–2049, 2020.

[64] Kurt Gray and Daniel M Wegner. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, 96(3):505, 2009.

[65] Tatsuya Nomura, Takayuki Uratani, Takayuki Kanda, Kazutaka Matsumoto, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. Why do children abuse robots? In *Proceedings of HRI Extended Abstracts*, HRI'15 Extended Abstracts, 2015.

[66] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[67] Johannes Himmelreich. Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3):669–684, 2018.

[68] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, May 2007.

[69] Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

[70] Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5687–5691, 2018.

[71] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of AAAI*, 2015.

[72] Ross A Knepper. On the communicative aspect of human-robot joint action. In *Proc. RO-MAN Workshop: Toward a Framework for Joint Action, What about Common Ground*, 2016.

[73] Luciana Benotti and Patrick Blackburn. Polite interactions with robots. *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016*, 2016.

[74] Gordon Briggs, Tom Williams, and Matthias Scheutz. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 6(1):64–94, 2017.

[75] Felix Gervits, Gordon Briggs, and Matthias Scheutz. The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. In *Proc. Annual Meeting of the Cog. Sci. Society*, 2017.

[76] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *Proc. Conf. of the North American Chapter of the ACL: Human Language Tech.*, 2018.

[77] Sean Trott and Benjamin Bergen. A theoretical model of indirect request comprehension. In *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction*, 2017.

[78] Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. "Thank you for sharing that interesting fact!": Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of HRI*, 2018.

[79] Matthias Scheutz. The need for moral competency in autonomous agent architectures. In *Fundamental Issues of Artificial Intelligence*, pages 515–525. Springer, 2016.

[80] Peter H Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather 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 *Proc. 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012.

[81] Reid Simmons, Maxim Makatchev, Rachel Kirby, Min Kyung Lee, et al. Believable robot characters. *AI Magazine*, (4), 2011.

[82] Friederike Eyssel and Dieta Kuchenbrandt. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, (4), 2012.

[83] Dan Bohus and Alexander I Rudnicky. Sorry, I didn't catch that!-an investigation of non-understanding errors and recovery strategies. In *6th SIGdial workshop on discourse and dialogue*, 2005.

[84] Matthew Marge and Alexander I Rudnicky. Miscommunication recovery in physically situated dialogue. In *Proceedings of SIGdial*, pages 22–49, Saarbrücken, Germany, 2015.

[85] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward information theoretic human-robot dialog. *Robotics: Science and Systems*, 32:409–417, 2013.

[86] Tom Williams and Matthias Scheutz. Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics. In *Proceedings of RSS*, Cambridge, MA, 2017.

[87] Tom Williams, Fereshta Yazdani, Prasanth Suresh, Matthias Scheutz, and Michael Beetz. Dempster-shafer theoretic resolution of referential ambiguity. *Autonomous Robots*, 2018.

[88] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013.

[89] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*. 2018.

[90] Matthias Scheutz, Bertram Malle, and Gordon Briggs. Towards morally sensitive action selection for autonomous social robots. In *Proc. of RO-MAN*, 2015.

[91] 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):3–5, 2011.

[92] Todd Gureckis, Jay Martin, John McDonnell, et al. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 48(3):829–842, 2016.

[93] Wilma Bainbridge, Justin Hart, Elizabeth Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1): 41–52, 2011.

[94] Kerstin Fischer, Katrin Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 463–470, Boston, MA, 2012.

[95] Jamy Li. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77: 23–37, 2015.

[96] Kazuaki Tanaka, Hideyuki Nakanishi, and Hiroshi Ishiguro. Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment. In *Proceedings of the International Conference on Collaboration Technologies (ICCT)*, pages 96–110, Minneapolis, MN, 2014. Springer.

[97] Matthew JC 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.

[98] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, 2017.

[99] JASP Team et al. Jasp. *Version 0.8. 0.0. software*, 2016.

[100] Daniel Wright. Comparing groups in a before-after design: When t test and ancova produce different results. *The British journal of educational psychology*, 76:663–75, 10 2006.

[101] Dimiter Dimitrov and Phillip D Rumrill. Pretest-posttest designs and measurement of change. *Work (Reading, Mass.)*, 20:159–65, 02 2003.

[102] Schuyler Huck and Robert A. McLean. Using a repeated measures anova to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 82:511–518, 07 1975.

[103] John K Kruschke. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5), 2010.

[104] James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association (ASA)*, 82(397), 1987.

[105] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, (11), 2011.

[106] Jonathan AC Sterne and George Davey Smith. Sifting the evidence – what's wrong with significance tests? *Physical Therapy*, 81(8):1464–1469, 2001.

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

[108] Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, 14(5):779–804, 2007.

[109] Josine Verhagen and Eric-Jan Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4):1457–1475, 2014.

[110] Alexander Ly, Alexander Etz, Maarten Marsman, and Eric-Jan Wagenmakers. Replication bayes factors from evidence updating. *Behavior Research Methods*, Aug 2018. ISSN 1554-3528. doi: 10.3758/s13428-018-1092-x. URL https://doi.org/10.3758/s13428-018-1092-x.

[111] Ryan Blake Jackson and Tom Williams. Robot: Asker of questions and changer of norms? In *Proceedings of the International Conference on Robot Ethics and Standards (ICRES)*, 2018.

[112] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1961.

[113] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[114] W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242, 1963.

[115] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. Measurement of trust in human-robot collaboration. In *Proceedings of the Symposium on Collaborative Technologies and Systems*, pages 106–114, 2007.

[116] Gordon Briggs and Matthias Scheutz. "Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions. In *Proceedings of the AAAI Fall Symposium Series*, 2015.

[117] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. Using robots to moderate team conflict: The case of repairing violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 229–236. ACM, 2015.

[118] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of AAAI*, 2016.

[119] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of HRI*, 2016.

[120] Timothy Brick and Matthias Scheutz. Incremental natural language processing for HRI. In *Proc. 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 263–270, 2007.

[121] Gordon Briggs and Matthias Scheutz. Multi-modal belief updates in multi-robot human-robot dialogue interaction. In *Proc. Symposium on Linguistic and Cognitive Approaches to Dialogue Agents*, 2012.

[122] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of AAMAS*, 2017.

[123] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. 2004.

[124] Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proc. International Conference on Robotics and Automation*, 2009.

[125] John R Searle. Indirect speech acts. *Syntax and Semantics*, 3:59–82, 1975.

[126] Tom Williams and Matthias Scheutz. Reference in robotics: A givenness hierarchy theoretic approach. In Jeanette Gundel and Barbara Abbott, editors, *The Oxford Handbook of Reference*. 2018.

[127] Tom Williams. A consultant framework for natural language processing in integrated robot architectures. *IEEE Intelligent Informatics Bulletin*, 2017.

[128] Tom Williams, Ravenna Thielstrom, Evan Krause, Bradley Oosterveld, and Matthias Scheutz. Augmenting robot knowledge consultants with distributed short term memory. In *International Conference on Social Robotics*, pages 170–180, 2018.

[129] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Social Robotics*, 2009.

[130] H. A. Yanco and J. Drury. Classifying human-robot interaction: an updated taxonomy. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2841–2846, 2004.

[131] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton C. T. Lee, Mitch Marcus, and Hadas Kress-Gazit. Sorry dave, i'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Proceedings of RSS*, 2013.

[132] 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):227–238, 2004. doi: 10.1080/10510970409388616.

[133] Sara Mills. *Gender and politeness*, volume 17. Cambridge University Press, 2003.

[134] Sara Mills. Gender and impoliteness, 2005.

[135] Robin Lakoff. Language and woman's place. *Language in society*, 2(1):45–79, 1973.

[136] Mary Hogue, Janice D Yoder, and Steven B Singleton. The gender wage gap: An explanation of men's elevated wage entitlement. *Sex Roles*, 56(9-10):573–579, 2007.

[137] Jennifer Coates. *Men talk: Stories in the making of masculinities*. John Wiley & Sons, 2008.

[138] Gino Eelen. *A critique of politeness theory*, volume 1. Routledge, 2014.

[139] Clifford Nass, Youngme Moon, and Nancy Green. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10):864–876, 1997.

[140] Friederike Eyssel and Frank Hegel. (s)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology*, 42(9):2213–2230, 2012.

[141] Julie Carpenter, Joan M Davis, Norah Erwin-Stewart, Tiffany R Lee, John D Bransford, and Nancy Vye. Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, 1(3):261, 2009.

[142] Benedict Tay, Younbo Jung, and Taezoon Park. When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38:75–84, 2014.

[143] Meia Chita-Tegmark, Monika Lohani, and Matthias Scheutz. Gender effects in perceptions of robots and humans with varying emotional intelligence. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 230–238. IEEE, 2019.

[144] 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*, pages 2563–2568. IEEE, 2009.

[145] Dalia Mortada. Meet q, the gender-neutral voice assistant, 2019. URL https://www.npr.org/2019/03/21/705395100/meet-q-the-gender-neutral-voice-assistant.

[146] Yan Wang and James E Young. Beyond pink and blue: Gendered attitudes towards robots in society. In *Proceedings of Gender and IT Appropriation. Science and Practice on Dialogue-Forum for Interdisciplinary Exchange*, page 49. European Society for Socially Embedded Technologies, 2014.

[147] Paul Schermerhorn, Matthias Scheutz, and Charles R Crowell. Robot social presence and gender: Do females view robots differently than males? In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 263–270. ACM, 2008.

[148] Megan Strait, Priscilla Briggs, and Matthias Scheutz. Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In *4th international symposium on new frontiers in human robot interaction*, 2015.

[149] Charles R Crowelly, Michael Villanoy, Matthias Scheutzz, and Paul Schermerhornz. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3735–3741. IEEE, 2009.

[150] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In *AAAI Conf. on Artificial Intelligence, Ethics, and Society*, 2019.

[151] John TE Richardson. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2):135–147, 2011.

[152] Mikel Aickin and Helen Gensler. Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American journal of public health*, 86(5):726–728, 1996.

[153] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902, 2002.

[154] Cecilia L Ridgeway and Shelley J Correll. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & society*, 18(4):510–531, 2004.

[155] Anna Studzińska. *Gender differences in perception of sexual harassment*. PhD thesis, 2015.

[156] Garrett Marks-Wilt and Philip Robbins. The gendered division of moral labor: Gender-asymmetric ascriptions of moral status. 2014.

[157] Laurie A Rudman and Stephanie A Goodwin. Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of personality and social psychology*, 87(4):494, 2004.

[158] Angeliki Alvanoudi and Theodossia-Soula Pavlidou. Grammatical gender and cognition. In *Major Trends in Theoretical and Applied Linguistics 2*, volume 2, pages 109–124. Versita, 2013.

[159] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. Robots and racism. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 196–204. ACM, 2018.

[160] Megan Strait, Ana Sánchez Ramos, Virginia Contreras, and Noemi Garcia. Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 452–457. IEEE, 2018.

[161] Mark West, Rebecca Kraut, and Han Ei Chew. I'd blush if i could: closing gender divides in digital skills through education. 2019.

[162] Sheryl Brahnam and Antonella De Angeli. Gender affordances of conversational agents. *Interacting with Computers*, 24(3):139–153, 2012.

[163] George Veletsianos, Cassandra Scharber, and Aaron Doering. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with computers*, 20(3):292–301, 2008.

[164] Christoph Bartneck and Jun Hu. Exploring the abuse of robots. *Interaction Studies*, 9(3):415–433, 2008.

[165] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. Why do children abuse robots? *Interaction Studies*, 17(3):347–369, 2016.

[166] Mark Scheeff, John Pinto, Kris Rahardja, Scott Snibbe, and Robert Tow. Experiences with sparky, a social robot. In *Socially intelligent agents*, pages 173–180. Springer, 2002.

[167] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[168] Amanda Cercas Curry and Verena Rieser. A crowd-based evaluation of abuse response strategies in conversational agents. *arXiv preprint arXiv:1909.04387*, 2019.

[169] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 29–37, 2021.

[170] Sara Mills. *Gender and politeness*. Number 17. Cambridge University Press, 2003.

[171] Sara Mills and Dániel Z Kádár. Politeness and culture. *Politeness in East Asia*, pages 21–44, 2011.

[172] JASP Team. JASP (Version 0.15)[Computer software], 2021. URL https://jasp-stats.org/.

[173] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[174] Robert H. Gass and John S. Seiter. *Persuasion: Social Influence and Compliance Gaining*. Routledge, 2015. ISBN 978-1-317-34838-2.

[175] Bertram F Malle. Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Info. Tech.*, 2016.

[176] Nava Tintarev and Roman Kutlak. Sassy-making decisions transparent with argumentation and natural language generation. In *IUI 2014 Workshop: Interacting with Smart Objects*, 2014.

[177] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of IJCAI*, 2017.

[178] Raj Korpan and Susan L Epstein. Toward natural explanations for a robot's navigation plans. In *HRI WS on Explainable Robotic Systems*, 2018.

[179] Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne Biundo. Making hybrid plans more clear to human users-a formal approach for generating sound explanations. In *Proc. ICAPS*, 2012.

[180] Maria Fox, Derek Long, and Daniele Magazzeni. Explainable planning. *IJCAI Workshop on Explainable AI*, 2017.

[181] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Balancing explicability and explanation in human-aware planning. In *AAAI Fall Symposium on AI-for-HRI*, 2017.

[182] Shirin Sohrabi, Jorge A Baier, and Sheila A McIlraith. Preferred explanations: Theory and generation via planning. In *AAAI*, 2011.

[183] Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Q'ly*, 1999.

[184] Eric Holder. Defining soldier intent in a human-robot natural language interaction context. Technical report, US Army Research Laboratory, 2017.

[185] Ze Gong and Yu Zhang. Robot signaling its intentions in human-robot teaming. In *HRI Workshop on Explainable Robotic Systems*, 2018.

[186] Rasmus S Andersen, Ole Madsen, Thomas B Moeslund, and Heni Ben Amor. Projecting robot intentions into human environments. In *Proc. Int'l Symp. on Robot and Human Interactive Communication*, 2016.

[187] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *Intel. Sys.*, 2006.

[188] Ron Sun. Moral judgment, human motivation, and neural networks. *Cognitive Computation*, 5(4): 566–579, 2013.

[189] Dieter Vanderelst and Alan Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 2017.

[190] Wendell Wallach, Stan Franklin, and Colin Allen. A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, 2(3):454–485, 2010.

[191] Fahad Alaieri and André Vellino. Ethical decision making in robots: Autonomy, trust and responsibility. In *Proceedings of the International Conference on Social Robotics*, 2016.

[192] Vicky Charisi, Louise Dennis, Michael Fisher Robert Lieck, Andreas Matthias, Marija Slavkovik Janina Sombetzki, Alan FT Winfield, and Roman Yampolskiy. Towards moral autonomous systems. *arXiv preprint arXiv:1703.04741*, 2017.

[193] Benjamin Kuipers. Human-like morality and ethics for robots. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[194] Felix Lindner and Martin Mose Bentzen. The hybrid ethical reasoning agent immanuel. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[195] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. Explaining robot actions. In *Proc. Int'l Conf. Human-Robot Interaction*, 2012.

[196] Paul Schermerhorn and Matthias Scheutz. Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In *Proceedings of the International Conference on Multimodal Interfaces*, 2009.

[197] Paul Schermerhorn and Matthias Scheutz. Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *Proceedings of Advances in Computer-Human Interaction*, 2011.

[198] Séverin Lemaignan, Mathieu Warnier, E Akin Sisbot, Aurélie Clodic, and Rachid Alami. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence*, 247:45–69, 2017.

[199] Bertram F Malle, Matthias Scheutz, and Joseph L Austerweil. Networks of social and moral norms in human and robot agents. In *A World with Robots*. 2017.

[200] Aditya Ghose and Tony Bastin Roy Savarimuthu. Norms as objectives: Revisiting compliance management in multi-agent systems. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 105–122. Springer, 2012.

[201] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 745–751. IEEE, 2016.

[202] Qin Zhu, Tom Williams, and Ruchen Wen. Role-based morality, ethical pluralism, and morally capable robots. *Journal of Contemporary Eastern Asia*, 20(1):134–150, 2021.

[203] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. Comparing strategies for robot communication of role-grounded moral norms. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 323–327, 2021.

[204] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 10–18, 2021.

[205] Keith Abney. Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot ethics: The ethical and social implications of robotics*, pages 35–52, 2012.

[206] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. An overview of the distributed integrated cognition affect and reflection diarc architecture. *Cognitive architectures*, 2019.

[207] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proc. AAAI WS on Int. Rob. Sys.*, 2013.

[208] Virgil Andronache and Matthias Scheutz. Ade—an architecture development environment for virtual and robotic agents. *International Journal on Artificial Intelligence Tools*, 15(02), 2006.

[209] James Kramer and Matthias Scheutz. ADE: A framework for robust complex robotic architectures. In *Proc. Int'l Conf. Intel. Robots and Sys.*, 2006.

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

[211] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, 2009.

[212] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. YARP: yet another robot platform. *Int'l Journal of Advanced Robotic Systems*, 2006.

[213] Steve Rowe and Christopher R Wagner. An introduction to the joint architecture for unmanned systems (JAUS). *Ann Arbor*, 2008.

[214] Paul Schermerhorn and Matthias Scheutz. Natural language interactions in distributed networks of smart devices. *International Journal of Semantic Computing*, 2(04), 2008.

[215] Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4), 1971.

[216] Jörg Hoffmann and Bernhard Nebel. The ff planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research*, 14, 2001.

[217] Malik Ghallab, Adele Howe, Craig Knoblock, Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. PDDL— the planning domain definition language. Technical report, 1998.

[218] Mauro Vallati, Lukáš Chrpa, Marek Grzes, Thomas L. McCluskey, Mark Roberts, and Scott Sanner. The 2014 international planning competition: Progress and trends. *AI Magazine*, 36(3), 2015.

[219] Maria Fox and Derek Long. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Jour. AI Research*, 2003.

[220] Stefan Edelkamp and Jörg Hoffmann. PDDL2. 2: The language for the classical part of the 4th international planning competition. Technical report, University of Freiburg, 2004.

[221] Alfonso Gerevini and Derek Long. Preferences and soft constraints in PDDL3. In *ICAPS WS on planning with preferences and soft constraints*, 2006.

[222] Neil T Dantam, Zachary K Kingston, Swarat Chaudhuri, and Lydia E Kavraki. An incremental constraint-based framework for task and motion planning. *The International Journal of Robotics Research*, 37(10), 2018.

[223] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2008.

[224] Fei Han, Xue Yang, Yiming Deng, Mark Rentschler, Dejun Yang, and Hao Zhang. SRAL: Shared representative appearance learning for long-term visual place recognition. *IEEE Robotics and Automation Letters*, 2017.

[225] Sriram Siva and Hao Zhang. Omnidirectional multisensory perception fusion for long-term place recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[226] Sriram Siva, Zachary Nahman, and Hao Zhang. Voxel-based representation learning for place recognition based on 3d point clouds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[227] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *Proc. Int'l Conf. on Rob. and Automation*, 2017.

[228] Edward Pepperell, Peter I Corke, and Michael J Milford. All-environment visual place recognition with smart. In *Proc. ICRA*, 2014.

[229] Hao Zhang, Fei Han, and Hua Wang. Robust multimodal sequence-based loop closure detection via structured sparsity. In *Robotics: Science and systems*, 2016.

[230] Gordon Briggs, Tom Williams, and Matthias Scheutz. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, 2017.

[231] Tom Williams and Matthias Scheutz. Reference in robotics: A givenness hierarchy theoretic approach. In *The Oxford Handbook of Reference*. Oxford University Press, 2019.

[232] Tom Williams and Matthias Scheutz. Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[233] Christopher Crick, Graylin Jay, Sarah Osentoski, Benjamin Pitzer, and Odest Chadwicke Jenkins. Rosbridge: ROS for non-ROS users. In *Robotics Research*. 2017.

[234] Thibault Kruse, Amit Kumar Pandey, Rachid Alami, and Alexandra Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[235] Santosh Balajee Banisetty and David Feil-Seifer. Towards a unified planner for socially-aware navigation. *arXiv preprint arXiv:1810.00966*, 2018.

[236] Santosh Balajee Banisetty, Scott Forer, Logan Yliniemi, Monica Nicolescu, and David Feil-Seifer. Socially-aware navigation: A non-linear multi-objective optimization approach. *arXiv preprint arXiv:1911.04037*, 2019.

[237] Julian Savulescu. Procreative beneficence: why we should select the best children. *Bioethics*, 15(5-6): 413–426, 2001.

[238] Robert Sparrow. A not-so-new eugenics: Harris and savulescu on human enhancement. *The Hastings Center Report*, 41(1):32–42, 2011.

APPENDIX

COPYRIGHT AND COAUTHOR PERMISSIONS

A.1 Chapter 2

The Frontiers journals website (https://www.frontiersin.org/journals/robotics-and-ai#about) states that authors retain copyright on their articles and that authors are free to disseminate and re-publish their articles, subject to any requirements of third-party copyright owners and subject to the original publication being fully cited. Screenshots of this copyright information are shown in Figure A.1 and Figure A.2.

Figure A.1 Open Access Statement from the Frontiers website

A.2 Chapter 3

Copyright permissions for use of the published work contained in this chapter are shown in Figure A.16 and Figure A.17 since the copyright holder and publisher is the same as for Chapter 8.

A.3 Chapter 4

The authors of the work presented in this chapter (i.e., the author of this thesis) chose to retain the copyright of this material but granted the ACM non-exclusive permission to publish this work as a journal article in the ACM Transactions on Human-Robot Interaction (THRI). Documentation of this agreement is shown in Figure A.3.

Figure A.2 Copyright Statement from the Frontiers website

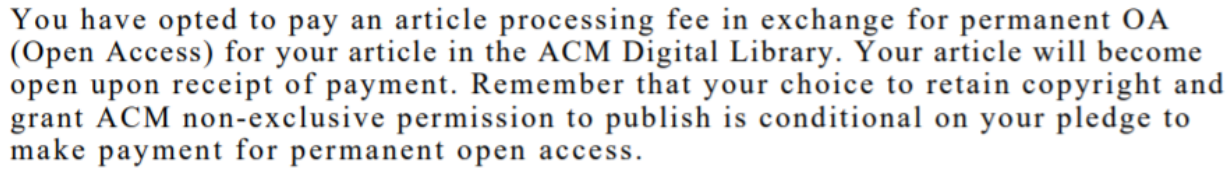


Figure A.3 Excerpt from ACM Permission Release Form

A.4 Chapter 5

The work presented in this chapter was published in the Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) in 2019. Coauthor permissions are presented here for all coauthors not on the dissertation committee (Figure A.4), as well as copyright permissions from the publisher(s) that allow the work to be included here (Figure A.5).

A.5 Chapter 6

The work presented in this chapter was published in the Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI) in 2020. Coauthor permissions are presented here for all coauthors not on the dissertation committee (Figure A.6), as well as copyright permissions from the publisher(s) (Figure A.7).

A.6 Chapter 7

The work presented in this chapter is unpublished as of now, so no copyright permissions are needed. Coauthor permissions are presented in Figure A.8, Figure A.9, and Figure A.10.

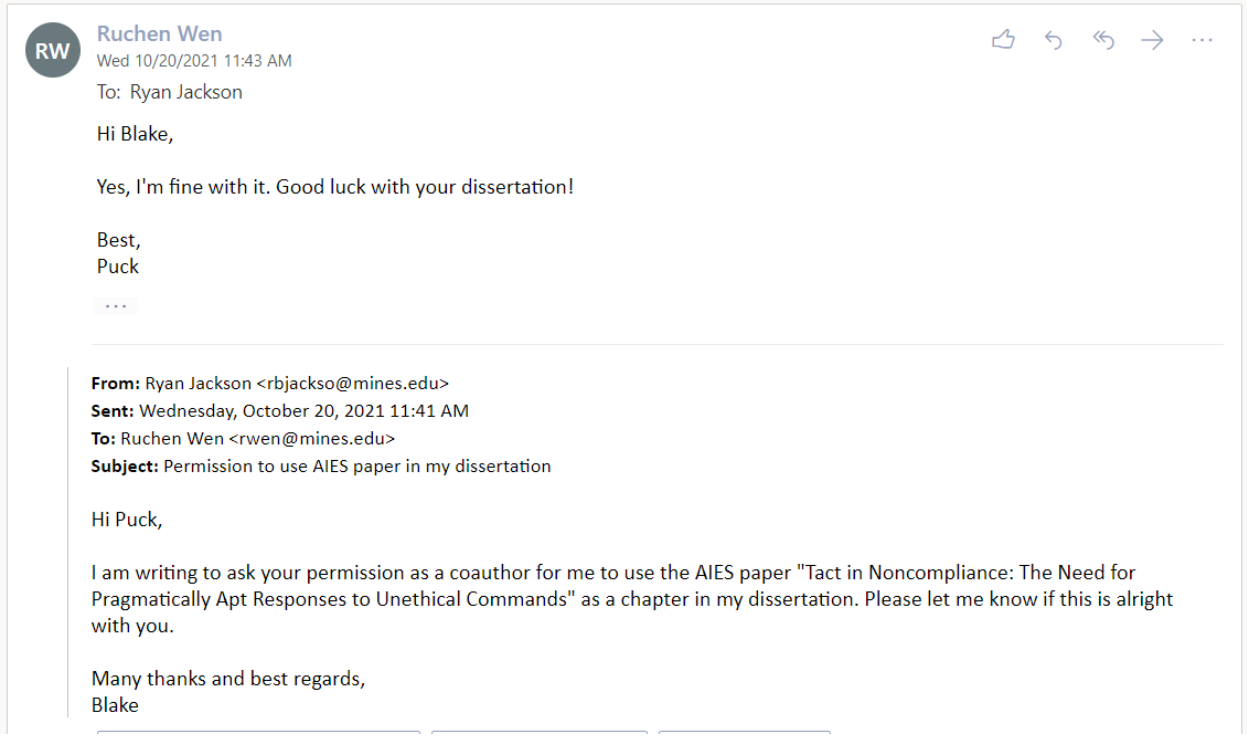


Figure A.4 Permission from coauthor Ruchen Wen

A.7 Chapter 8

The work presented in this chapter was published in the Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) in 2021. Coauthor permissions are presented here for all coauthors not on the dissertation committee (Figure A.11, Figure A.12, Figure A.13, Figure A.14, Figure A.15), as well as copyright permissions from the publisher(s) (Figure A.16, Figure A.17). In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Colorado School of Mines's products or services.

(b) Furthermore, notwithstanding the exclusive rights the Owner has granted to ACM in Paragraph 2(a), Owner shall have the right to do the following:

(i) Reuse any portion of the Work, without fee, in any future works written or edited by the Author, including books, lectures and presentations in any and all media.

(ii) Create a "Major Revision" which is wholly owned by the author

(iii) Post the Accepted Version of the Work on (1) the Author's home page, (2) the Owner's institutional repository, (3) any repository legally mandated by an agency funding the research on which the Work is based, and (4) any non-commercial repository or aggregation that does not duplicate ACM tables of contents, i.e., whose patterns of links do not substantially duplicate an ACM-copyrighted volume or issue. Non-commercial repositories are here understood as repositories owned by non-profit organizations that do not charge a fee for accessing deposited articles and that do not sell advertising or otherwise profit from serving articles.

(iv) Post an "Author-Izer" link enabling free downloads of the Version of Record in the ACM Digital Library on (1) the Author's home page or (2) the Owner's institutional repository;

(v) Prior to commencement of the ACM peer review process, post the version of the Work as submitted to ACM ("Submitted Version" or any earlier versions) to non-peer reviewed servers;

(vi) Make free distributions of the final published Version of Record internally to the Owner's employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

Figure A.5 Permission from copyright holder for Chapter 5

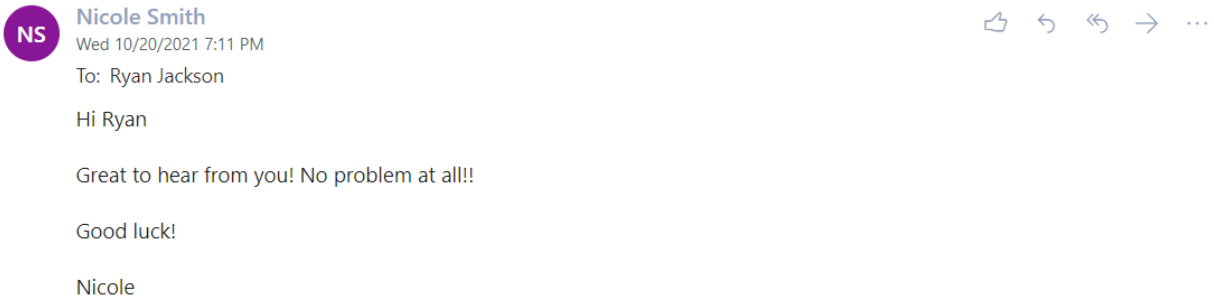


Figure A.6 Permission from coauthor Nicole Smith

(a) All rights and permissions the author has not granted to ACM in Paragraph 2 are reserved to the Owner, including without limitation the ownership of the copyright of the Work and all other proprietary rights such as patent or trademark rights.

(b) Furthermore, notwithstanding the exclusive rights the Owner has granted to ACM in Paragraph 2(a), Owner shall have the right to do the following:

(i) Reuse any portion of the Work, without fee, in any future works written or edited by the Author, including books, lectures and presentations in any and all media.

(ii) Create a "Major Revision" which is wholly owned by the author

(iii) Post the Accepted Version of the Work on (1) the Author's home page, (2) the Owner's institutional repository, (3) any repository legally mandated by an agency funding the research on which the Work is based, and (4) any non-commercial repository or aggregation that does not duplicate ACM tables of contents, i.e., whose patterns of links do not substantially duplicate an ACM-copyrighted volume or issue. Non-commercial repositories are here understood as repositories owned by non-profit organizations that do not charge a fee for accessing deposited articles and that do not sell advertising or otherwise profit from serving articles.

(iv) Post an "Author-Izer" link enabling free downloads of the Version of Record in the ACM Digital Library on (1) the Author's home page or (2) the Owner's institutional repository;

(v) Prior to commencement of the ACM peer review process, post the version of the Work as submitted to ACM ("Submitted Version" or any earlier versions) to non-peer reviewed servers;

(vi) Make free distributions of the final published Version of Record internally to the Owner's employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

Figure A.7 Permission from copyright holder for Chapter 6

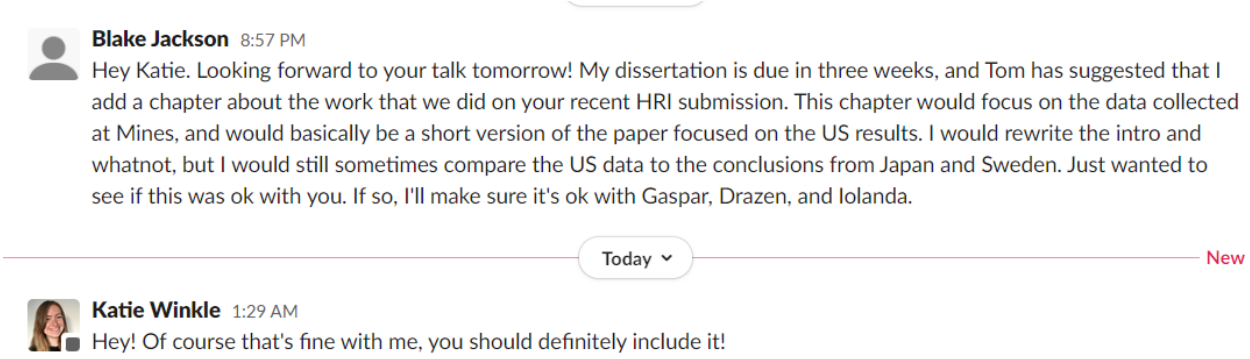


Figure A.8 Permission from coauthor Katie Winkle

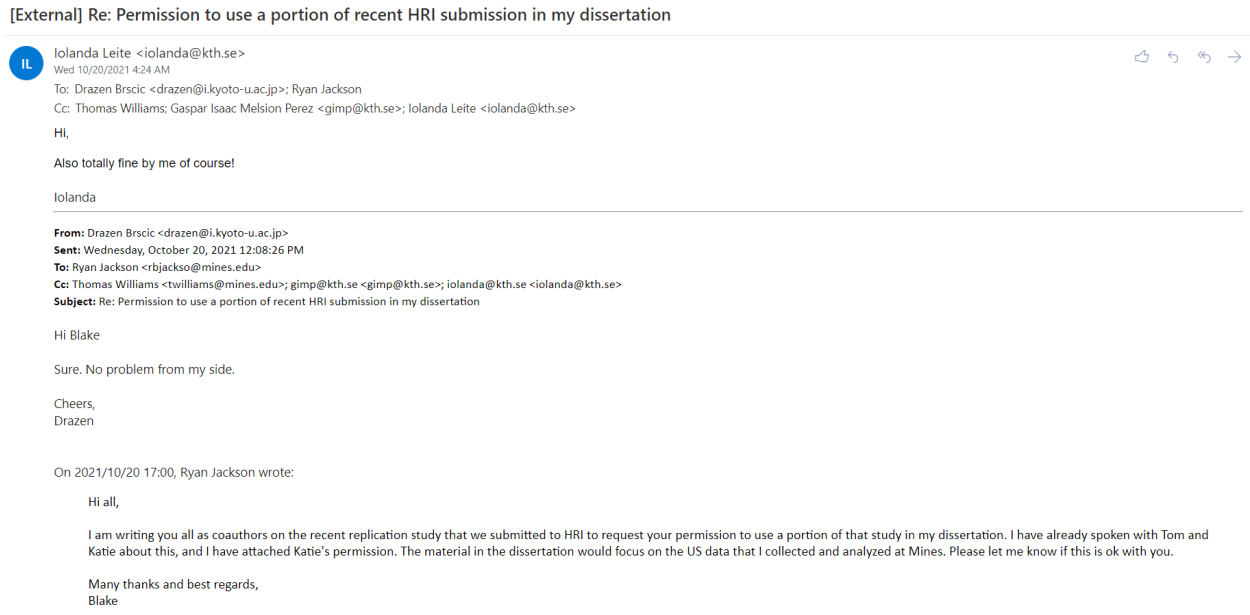
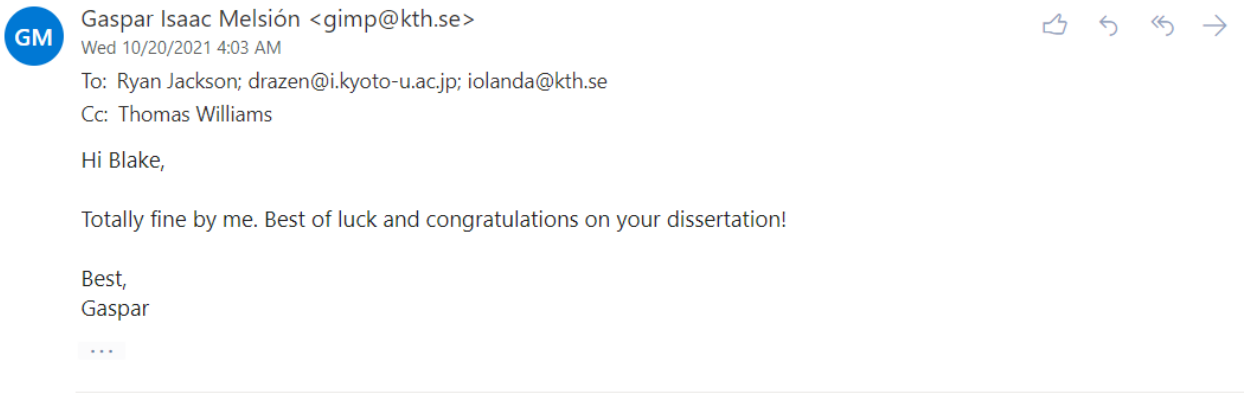


Figure A.9 Permission from coauthors Iolanda Leite and Drazen Brscic



GM

Gaspar Isaac Melsión <gimp@kth.se>
Wed 10/20/2021 4:03 AM

To: Ryan Jackson; drazen@i.kyoto-u.ac.jp; iolanda@kth.se
Cc: Thomas Williams

Hi Blake,

Totally fine by me. Best of luck and congratulations on your dissertation!

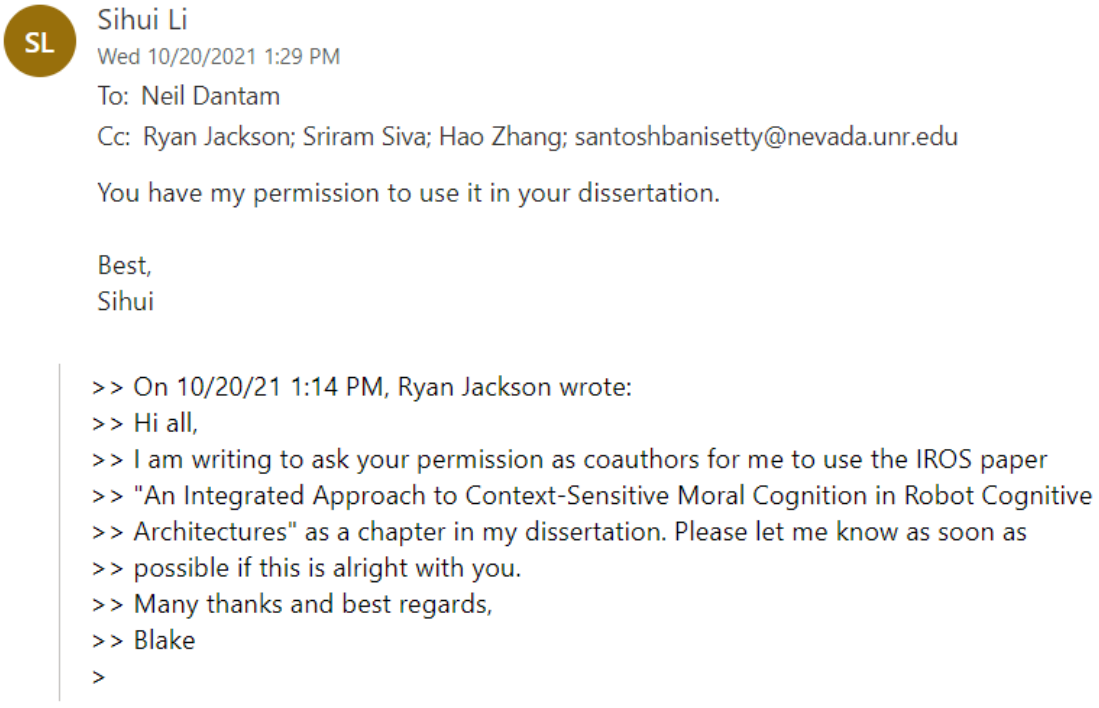
Best,
Gaspar

...

On 20/10/2021 10:00, Ryan Jackson wrote:

> Hi all,
>
> I am writing you all as coauthors on the recent replication study that we submitted to HRI to request your permission to use a portion of that study in my dissertation. I have already spoken with Tom and Katie about this, and I have attached Katie's permission. The material in the dissertation would focus on the US data that I collected and analyzed at Mines. Please let me know if this is ok with you.
>
> Many thanks and best regards,
> Blake

Figure A.10 Permission from coauthor Gaspar Isaac Melsión



SL

Sihui Li
Wed 10/20/2021 1:29 PM

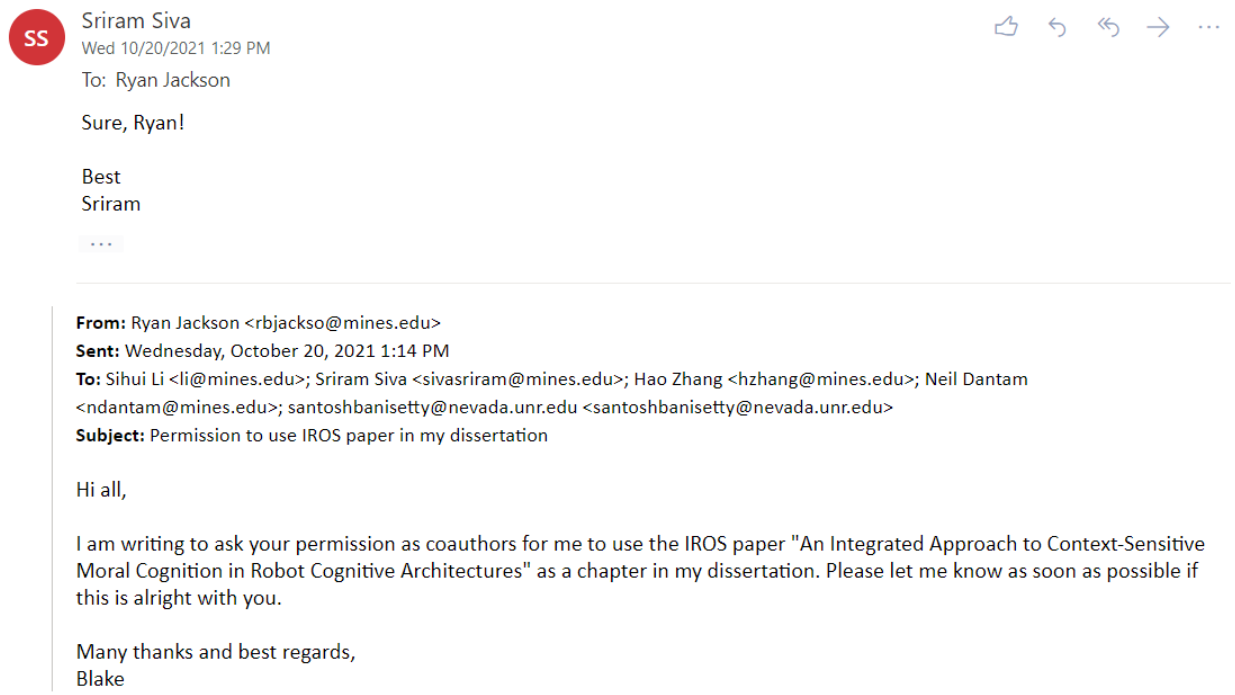
To: Neil Dantam
Cc: Ryan Jackson; Sriram Siva; Hao Zhang; santoshbanisetty@nevada.unr.edu

You have my permission to use it in your dissertation.

Best,
Sihui

>> On 10/20/21 1:14 PM, Ryan Jackson wrote:
>> Hi all,
>> I am writing to ask your permission as coauthors for me to use the IROS paper
>> "An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive
>> Architectures" as a chapter in my dissertation. Please let me know as soon as
>> possible if this is alright with you.
>> Many thanks and best regards,
>> Blake
>

Figure A.11 Permission from coauthor Sihui Li



Sriram Siva
Wed 10/20/2021 1:29 PM
To: Ryan Jackson

Sure, Ryan!

Best
Sriram

...

From: Ryan Jackson <rbjackso@mines.edu>
Sent: Wednesday, October 20, 2021 1:14 PM
To: Sihui Li <li@mines.edu>; Sriram Siva <sivasriram@mines.edu>; Hao Zhang <hzhang@mines.edu>; Neil Dantam <ndantam@mines.edu>; santoshbanisetty@nevada.unr.edu <santoshbanisetty@nevada.unr.edu>
Subject: Permission to use IROS paper in my dissertation

Hi all,

I am writing to ask your permission as coauthors for me to use the IROS paper "An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures" as a chapter in my dissertation. Please let me know as soon as possible if this is alright with you.

Many thanks and best regards,
Blake

Figure A.12 Permission from coauthor Sriram Siva

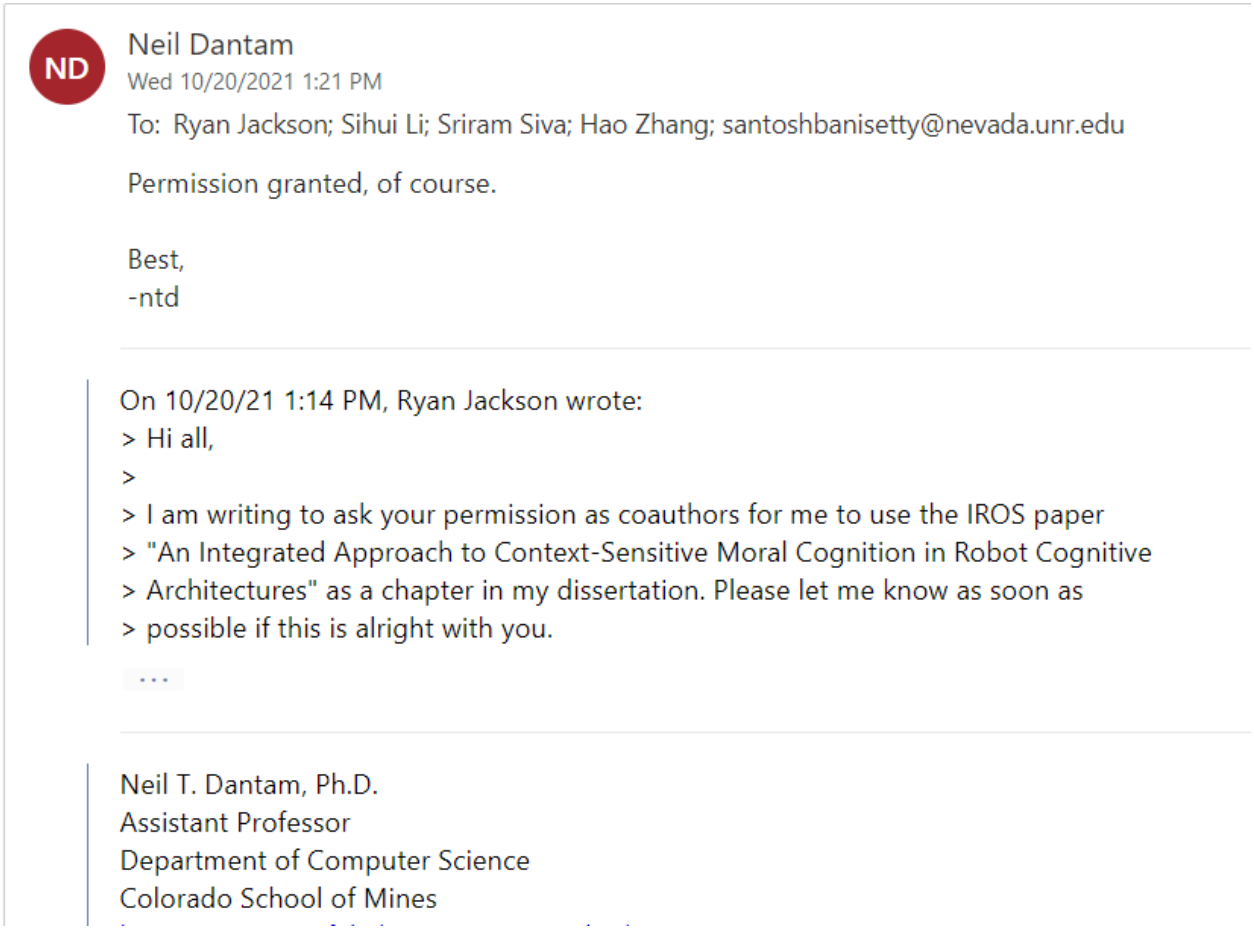


Figure A.13 Permission from coauthor Neil Dantam

RE: Permission to use IROS paper in my dissertation

Hao Zhang <hzhang@mines.edu>
Wed 10/20/2021 1:42 PM

To: Ryan Jackson <rbjackso@mines.edu>; Sihui Li <li@mines.edu>; Sriram Siva <sivasriram@mines.edu>; Neil Dantam <ndantam@mines.edu>; santoshbanisetty@nevada.unr.edu <santoshbanisetty@nevada.unr.edu>

Permission granted.

Best,
Hao

Hao Zhang, Ph.D.
Associate Professor, Dept. Computer Science
Director, Human-Centered Robotics Laboratory
Colorado School of Mines
Email: hzhang@mines.edu
Phone: (303) 273-3581
Url: http://inside.mines.edu/~hzhang
HCRobotics Lab: http://hcr.mines.edu

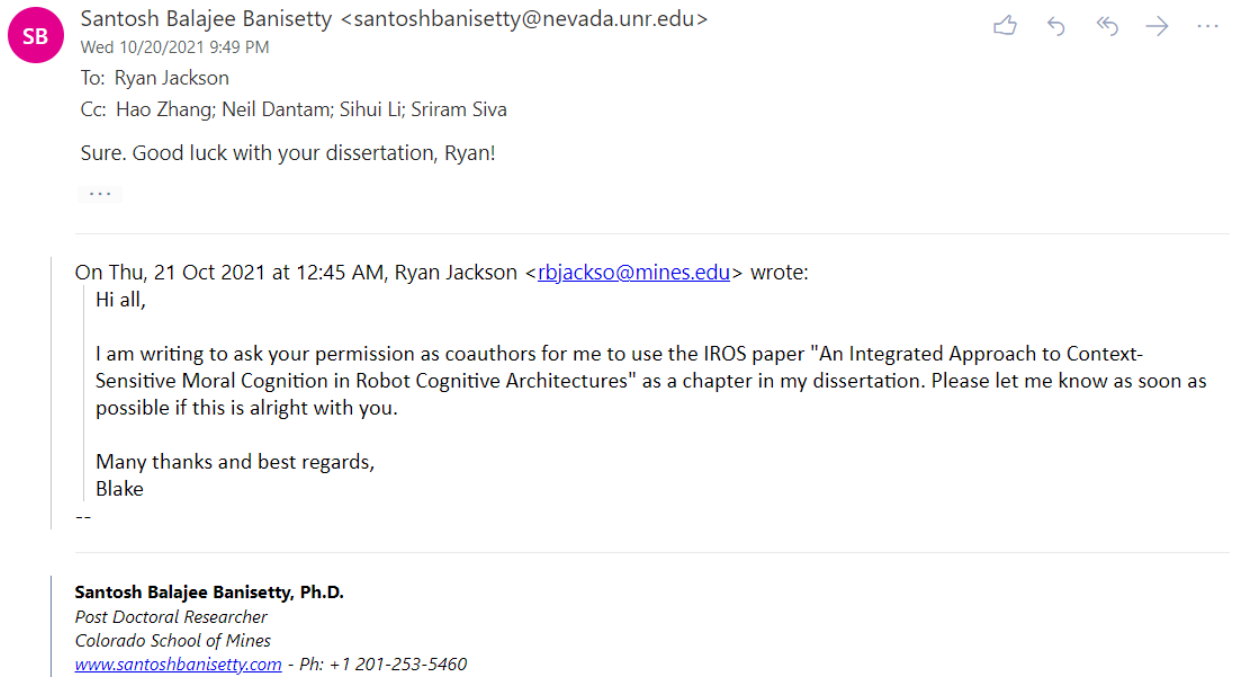
From: Ryan Jackson <rbjackso@mines.edu>
Sent: Wednesday, October 20, 2021 1:15 PM
To: Sihui Li <li@mines.edu>; Sriram Siva <sivasriram@mines.edu>; Hao Zhang <hzhang@mines.edu>; Neil Dantam <ndantam@mines.edu>; santoshbanisetty@nevada.unr.edu
Subject: Permission to use IROS paper in my dissertation

Hi all,

I am writing to ask your permission as coauthors for me to use the IROS paper "An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures" as a chapter in my dissertation. Please let me know as soon as possible if this is alright with you.

Many thanks and best regards,
Blake

Figure A.14 Permission from coauthor Hao Zhang



SB

Santosh Balajee Banisetty <santoshbanisetty@nevada.unr.edu>
Wed 10/20/2021 9:49 PM
To: Ryan Jackson
Cc: Hao Zhang; Neil Dantam; Sihui Li; Sriram Siva

Sure. Good luck with your dissertation, Ryan!

...

On Thu, 21 Oct 2021 at 12:45 AM, Ryan Jackson <rbjackso@mines.edu> wrote:

> Hi all,
>
> I am writing to ask your permission as coauthors for me to use the IROS paper "An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures" as a chapter in my dissertation. Please let me know as soon as possible if this is alright with you.
>
> Many thanks and best regards,
> Blake

--

Santosh Balajee Banisetty, Ph.D.
Post Doctoral Researcher
Colorado School of Mines
www.santoshbanisetty.com - Ph: +1 201-253-5460

Figure A.15 Permission from coauthor Santosh Balajee Banisetty

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the

Figure A.16 Permission from copyright holder for Chapter 8

- **Does IEEE require individuals working on a thesis or dissertation to obtain formal permission for reuse?**

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you must follow the requirements listed below:

Figure A.17 Permission from copyright holder for Chapter 8