Abstract

Investigating Human Perceptions and Responses to Robot Failures Across Functional, Social, and Moral Contexts in Human-Robot Interactions

Nicholas C. Georgiou

2026

Robots are increasingly becoming a part of human environments, where they have the potential to assist people in their daily lives in an abundance of ways. However, when robots interact within these settings, they are bound to fail at some point. A robot may fail due to a wide variety of reasons, whether it be due to an uncertainty in the robot's sensor reading, an inaccuracy in the robot's world model, or due to the unpredictability of human behavior in the robot's workspace. To design robots that can appropriately interact within environments filled with people, it is critical to understand how people perceive and respond to robots when they fail.

This dissertation investigates people's perceptions and responses to robot failures in human-robot interaction contexts along three different dimensions: functional, social, and moral. To examine failures within these dimensions, we conduct a range of controlled, human-subjects experiments that begin with simple, task-based failures and that progressively move towards failures that have more social and moral implications.

In our first human-subjects experiment, we investigate how people provide feedback to a variety of task-based robot failures in a card-selection task. We find significant variation in how people evaluate the robot's performance and show that this

variance can influence how effectively the robot performs the task when trained with different feedback strategies.

Building on these findings, we conduct a pair of studies with children aged four to seven years and with adults, in order to investigate how user characteristics influence variability in people's responses to robot failures. We find that users' age, as well as a robot's social responses following its own failures (i.e., providing incorrect advice), affects how people trust the failing robot.

Finally, we extend our investigation to a much more severe robot failure with inherent moral implications, in which a robot intentionally commits physical harm (i.e., pushing down a human). In this final experiment, we showcase the importance of prior expectations when people evaluate harmful behavior, as the ways in which the robot's capabilities were framed *before* witnessing the failure significantly influenced people's moral judgments of the robot after the failure.

Overall, this dissertation contributes to our knowledge of how people respond to robot failures and can help inform the design of robots that interact with regular people.

Investigating Human Perceptions and Responses to Robot Failures Across Functional, Social, and Moral Contexts in Human-Robot Interactions

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Nicholas C. Georgiou

Dissertation Director: Brian Scassellati

May 2026

"Failure is not the opposite of success; it's part of success." – Ariana Huffington

Contents

List of Figures

List of Tables

Acknowledgment

First, I would like to thank my advisor, Scaz. Thank you for the opportunity all those years ago when you first accepted me into this program and thank you for continuing to support me, since. Thank you for the all of the flexibility that you granted me throughout this PhD. I was able to research a wide range of HRI topics throughout my time at Yale, and to be a part of some really great collaborations that you made possible.

I would also like to thank the members of my committee: Drazen, Tesca, and Marynel. Thank you for all of the meetings that we had, all of the time and energy that you put into me, and all of the advice that you provided along the way. A particularly special shoutout to Drazen, who would e-meet with me all of the way from Japan.

To all of my collaborators and labmates, I sincerely thank you for all of your support. A big thank you to Tamar Kushnir and Tess Flanagan, a great collaborator and friend, both of whom I have collaborated with since my first semester at Yale. Thank you to all of the grad students in the robotics labs at Yale and thank you to all of the undergraduates that contributed to work in this dissertation.

I would also like to thank all of the friends and family that supported me throughout my PhD. A huge thank you goes out to all of the friends at Yale that I have made over the years. All the great times that we shared have helped me get to the finish line. To my intramural soccer and basketball teammates, I am so grateful for all of the

moments that we got to share on and off the field and the court. Our championship runs were special. To the Greek graduate student community at Yale, thank you for always reminding me of home. To all of my Computer Science friends outside of the lab and all of the friends that I have made in SEAS, a huge thank you for all of the support and companionship. To all of my close friends in the plant biology department at Yale, thank you for taking me in as a honorary PMB student. I will always cherish the trivia nights at ERB, the Survivor watch parties, the holiday parties, the hikes, and all of the hangouts that we had. To my friends from back home, my family friends, my family back in Greece, and Morgan's family (that has treated my like their own) thank you all for the support and encouragement throughout these years.

There are not enough words to describe how critical my "three ladies" have been in getting me here today, and I cannot thank them enough for their endless support, love, and patience with me. Mama, you are an endless source of support and motivation for me and I know that you are always there for me. Thank you for everything that you do for our family. Joanna, thank you for your love and support. As your older brother, I always want to set a good example for you and help you whenever I can, but many times you are the one inspiring and helping me. Morgan, you are a constantly shining light and a calming force that has helped me get through it all. I am so thankful that we shared our Yale PhD experience together. You (and Archie – our cat) bring me happiness every day. I could not have done this without the three (four with Archie) of you.

Finally, a tremendous thank you to my dad, who motivates me and inspires me every day. Losing you three years into my PhD was the hardest thing I've ever gone through. I love you and I think about you every day. I did it. Dr. Georgiou, just like you ♡.

Funding I would also like to thank the funding bodies that have made this work possible. These include the National Science Foundation (NSF) (Grants 1813651, 2106690, 1928448, and 1955653, SL-1955280) and the Office of Naval Research (ONR) (award N00014-18-1-2776 and N00014-24-1-2124).

Chapter 1

Introduction

Robots are becoming a larger presence in human environments, where they are bound to interact with people. Interaction may be a primary function of a robot's role, like when a tutoring robot conducts lessons with students or when an industrial robot collaborates with co-workers. Alternatively, an interaction may occur simply because a robot's role requires it to be co-located with people, such as when a household robot performs tasks in the same space as users. Ideally, robots that interact with people will seamlessly integrate into our environments and complete the tasks that we need of them without any difficulty. But, since human environments are unstructured, dynamic, and filled with people that are unpredictable and diverse, human-robot interactions are bound to not go as planned all of the time.

This dissertation revolves around the moments when robots behave in ways that they should not during human-robot interactions (HRI). When a robot behavior is not ideal, and it does not satisfy task objectives or rules, meet user expectations, or follow design specifications, we will refer to such a behavior as a failure. When robots interact in human environments, failure is inevitable at some point, whether it be due to an imprecise sensor reading, an inaccurate world model, or an unpredictable user. Failures can take many forms and they can vary widely, as they can range

from a small functional mistake such as a robot selecting a suboptimal action when performing a task to a much more serious event, like a robot causing physical harm.

When robots fail, they can have a significant impact on users, since they have the potential to not only influence how people perceive a robot, but also how people behave towards a robot. For example, a robot's failures can lead to an erosion of trust from its users, along with a lower willingness to re-engage with the robot in the future. For these reasons, it is critical to understand how different HRI contexts affect regular people's perceptions and responses to robot failures. By doing so, these insights can better inform the design of robots that can appropriately interact with people [104].

However, this problem is particularly challenging. A robot's failures may have varying implications depending on the context surrounding the failure. Furthermore, people have different expectations, beliefs, and backgrounds which affect the ways that they view robots and evaluate their actions. We still do not have a complete understanding of how different failure contexts influence people perceptions and responses to robot failures.

The work in this dissertation investigates novel failure scenarios in human-robot interactions. We take a multifaceted approach in our research methodologies in order to gain knowledge into how people perceive and respond to robot failures. We examine a range of failures that may occur when robots are placed into human environments, study how contextual factors surrounding robot failures shape people's perceptions and responses, and offer insights to inform the development of robots that interact with regular people.

This dissertation begins with a review of relevant literature involving robot failures in human-robot interaction in Chapter 2. The chapter begins with a brief overview of how other researchers have defined and categorized failures in HRI. The chapter continues with an overview of the different types of failures that may occur during

HRI and introduces a multidimensional framework for how humans perceive robot failures in HRI. The chapter also discusses how the context surrounding robot failure (i.e., nature of the failure, robot behavior following failure, robot characteristics, user characteristics) influences people's perceptions and responses.

In Chapters 3-5, we present human-subjects experiments designed to broaden our understanding of how different types of robot failures affect people's perceptions and responses. A key component of these experiments involves intentionally manipulating a robot's behavior to fail in particular ways, allowing us to study its effects on human perceptions and responses in human-robot interactions. As such, in each experiment, we carefully select the type of failure(s) and the context surrounding the failure that the robot commits. In each chapter, we primarily focus on broadening our understanding of the effects of robot failures within one of three different dimensions: functional, social, and moral. Our experiments begin by investigating the effects of simple, task-based robot failures in the functional dimension and progressively move towards failures that have more social and moral implications.

First, in Chapter 3, we examine a setting in which a human evaluates a robot's performance as it commits a range of task-based, functional failures. To do this, we conduct an in-person study ($N = 36$) in which people evaluate a robot on a scale of 1-10 each time the robot succeeds or fails at a card-selection task. We find that successful robot actions (i.e., selecting the correct set of cards) lead to similar feedback across participants, but that robot failures lead to much more variation in people's responses. Interestingly, all participants in our experiment provide some form of partial credit to the robot after it fails, but participants differ in how they attribute this partial credit. After collecting each person's evaluative feedback, we use simulations to investigate how the robot would perform if it was trained with each participant's extrapolated feedback strategy (i.e., a linear regression fit to what they value when providing partial-credit). Using each feedback strategy, we train a sepa-

rate multi-layer perceptron (MLP) to predict a proxy reward label for every possible robot action in a given state. We find that some participants' feedback strategies produce labels that inform the model to select actions that achieve significantly higher task performance than other feedback strategies. In summary, people value different factors when providing partial credit after a robot fails, and robots that learn directly from human feedback should be equipped to deal with this phenomenon, as a failure to account for this can lead to ineffective task performance.

Next, in Chapter 4, we investigate the social implications of robot failure, by examining how user characteristics (i.e., age) influence people's trust in a robot that fails. To do this, we conduct a pair of 2x3 between-subjects studies: one with school-aged children between the ages of 4 and 7 years old (N=168) and one with adults (N=168). In these studies, participants play a collaborative word-guessing game with a partner, in which the participants' goal is to select the correct names for unknown objects, and the partner sometimes gives wrong advice on certain objects. We manipulate the type of partner (human vs. robot) and the type of response by the partner following wrong advice (mistaken vs. apologetic vs. uncooperative). We find that older children, between the ages of 6 and 7, are less trusting of their partner after it errs than adults and than younger children, between the ages of 4 and 5. Additionally, older children maintain their trust in a robot that apologizes for a longer time period than they do in a human that apologizes. We find that people's age and their partner's failure responses can significantly affect peoples' trust throughout an interaction. In turn, extra consideration must be taken when designing robots that interact with children, as they may perceive or react to failing robots differently, based on their age.

Finally, in Chapter 5, we explore a setting in which a robot commits a physical, moral transgression against a human (in a video of a robot pushing down a human). Although it is not pleasant to think about, a robot behavior leading to some form

of physical harm to a person is entirely possible when robots operate in the same environments that people do. Since prior work suggests that people can view robots as morally accountable, we investigate what factors affect people's moral judgments towards a robot that did something morally wrong. To do this, we perform a 2x4 between-subjects, online study in which participants ($N = 720$) first read a backstory about an agent and then watch the agent commit a physical transgression (i.e., pushing down a human). We manipulate the type of transgressor (either human or robot) and the kind of backstory that is told about the transgressor, whether it is a simple, default introduction, or stories about the transgressor's physio-emotional capabilities (e.g., feeling pain or hunger), socio-emotional capabilities (e.g., feeling love or shame) or cognitive capabilities (e.g., thinking or reasoning). We find that participants believe that both the human and the robot intend to push down the human, but that the human transgressor is perceived to have higher morality than the robot. The type of backstory about the transgressor also significantly affects perceived morality. Interestingly, we find that robots with more emotional backstories, such as those that highlighted physio-emotional or socio-emotional capabilities, are perceived to have higher moral status than other robots, across measures such as desire, moral knowledge, and emotional knowledge. In summary, designers should be cognizant of how a robot is described to and perceived by users, as this can significantly affect how people view a robot's morality after it does something morally wrong.

Chapter 6 provides a discussion of the work presented throughout the dissertation, including the contributions of our work, as well as limitations and directions for future work. In this chapter, we also offer conclusions for this dissertation.

Chapter 2

Robot Failures in Human-Robot Interactions

This chapter provides an overview of robot failures in human-robot interactions (HRI). We define robot failures in human-robot interactions and discuss how researchers have previously classified robot failures in HRI. We then introduce a multidimensional framework with three dimensions related to how people perceive and respond to robot failures: functional, social, and moral. Last, we discuss how the context surrounding robot failures can influence people's perceptions and responses across these three dimensions.

2.1 What is a Robot Failure in HRI?

As robots are increasingly deployed in human environments, robot failures during human-robot interactions are inevitable. But what does it mean for a robot failure to occur in HRI?

2.1.1 Definition

We define a robot failure as *any robot behavior that deviates from ideal or expected performance, according to a set of objective task-related rules or ground truth, the expectations or desires of users, or the design put forth by developers.*[1] Our definition is inspired by Brooks' 2016 dissertation on a human-centric approach to autonomous robot failures [38] and by Honig and Oron-Gilad's 2018 literature review on understanding and resolving failures in human-robot interaction [104], which defined failure as "a degraded state of ability which causes the behavior or service performed by the system to deviate from the ideal, normal, or correct functionality". In this dissertation, any robot failure that is perceived and evaluated by a human, such that it affects the human's cognitive or behavioral responses towards the robot, is a part of a human-robot interaction.

Under our definition, a robot behavior can be classified as a failure in both objective and subjective terms. In objective cases, successful or optimal robot behavior is defined by a ground truth, set of rules, or design principles. When a robot behaves sub-optimally or violates these rules or principles, the behavior can be considered a failure. However, even when a robot performs a task optimally according to objective metrics, it may still be perceived as a failure depending on how end users view and respond to the robot's behavior.

2.1.2 Prior Robot Failure Classifications

Traditionally, robot failures were studied through a robot-centric perspective, focused on how they affected the robot's ability, or inability, to perform tasks effectively and efficiently. Robots typically had very specific roles that could help automate monotonous or dangerous tasks, such as production robots in automotive plants [244]

[1]Note that some work uses the terms *error*, *mistake*, or *fault* with our definition. For the sake of this dissertation, all of these terms fall under the broader term of failure.

or unmanned ground vehicles (UGVs) in search and rescue or military operations [43]. Therefore, failure classifications were functional in nature, focused on robot behaviors that degraded task performance. These classifications helped designers and operators understand common task-related failures and guided improvements to make robots more reliable and efficient in the future.

In 2005, Carlson et al. [43] introduced a robot failure classification that categorized failures in unmanned ground vehicles (UGVs), or autonomous and teleoperated land robots. They organized failures into two main categories based on cause: the physical category defined by robot component deficiencies and the human category defined by design faults or teleoperator errors. Carlson et al. also categorized failures by their repairability, defined by how easily the robot could be repaired following a failure, and by their impact, defined by the extent to which the robot's capability to perform its task was degraded. In these contexts, the consequences of robot failures were considered exclusively from a robot-centric, functional perspective, focusing on the robot's operational performance and its capacity to continue completing tasks.

Inspired by that work, in 2012, Steinbauer et al. [245] published a survey about common sources of robot failures that occurred during RoboCup, an international annual robotics and artificial intelligence research competition that involves the development of autonomous, humanoid robot soccer teams. The survey provided researchers with a better understanding of why state-of-the-art robots were failing when operating autonomously in dynamic, complex environments. Similarly to Carlson's 2005 survey, Steinbauer et al. classified robot failures by cause, into the categories of hardware (e.g., sensors, manipulators), software (e.g., perception, decision making), interaction (e.g., environment, other robots), and algorithms (e.g., behavior execution, localization). Again, failures were approached from a robot-centric and functional perspective, focusing on the underlying causes of failures that degraded robot's task performance.

In their 2015 systematic analysis of video data (N=201 videos) from five different HRI user studies, Giuliani et al. [91] adopted a more human-centered perspective on robot failure analysis in several ways. First, their analysis focused on user studies involving interactions between robots and non-expert users. Second, they examined the potential social consequences of robot failures, by including social norm violations as one of their robot failure categories (defined as situations in which the robot did not adhere to the underlying social script of the interaction, typically occurring at planning time), in addition to a technical failures category (described as technical shortcomings of the robot that mainly emerged at execution time). Third, they measured the effects of robot failures on users by identifying social signals (e.g., head and body movements, speech, hand gestures) that people exhibited in response to failures. Giuliani et al.'s analysis demonstrated a clear shift in perspective, showing that robot failures can take on social meaning and elicit distinct user responses, as opposed to prior surveys that primarily emphasized the impact of robot failures on the robot's task-related performance metrics.

This human-centered perspective is further emphasized in Honig and Oron-Gilad's 2018 review (N=52 papers) on understanding and resolving failures in HRI [104]. In their survey, Honig and Oron-Gilad argued that focusing on untrained, non-expert users is essential for robot failure analysis in HRI, as robots are progressively utilized in everyday human settings. In particular, they stressed the importance of investigating the cognitive factors that shape how users perceive and respond to robot failures. Thus, along with Giuliani et al.'s analysis, these works marked a transition from primarily robot-centric failure analysis towards approaches that accounted for users' perceptions and responses.

Recent failure classifications in HRI increasingly adopt a human-centered approach, explicitly considering how failures affect users and their perceptions. In 2020, Tolmeijer et al. [260] classified robot failures based on whether they were caused by

the robot's design, a system level issue, a behavior misaligned with the user's expectations, or the user behaving inappropriately. Their primary goal was to identify failures that could erode user trust and to offer mitigation strategies that the robot could use to recover that trust. Similarly, in 2024, Cameron et al. [41] analyzed domestic robot failures by their causes and outcomes, focusing on how failures influenced users' perceptions of the robot's trustworthiness. In both cases, the emphasis on trust highlights the importance of understanding the social consequences of robot failures for end users.

Finally, a 2021 taxonomy by Tian and Oviatt [259] fully embraced a human-centric approach to robot failure classification by categorizing failures solely based on how they influence users' perceptions and responses. They defined two main categories: performance errors, which degrade a user's perception of a robot's capability in achieving a task (e.g., perceived intelligence and competence), and social errors, which degrade a user's perception of a robot's socio-affective competence and their relationship with it. This taxonomy stressed the importance of evaluating robot failures through multiple user-centered lenses, while categorizing robot failures based on their impact on users' perceptions, rather than their cause.

In summary, robot failure classifications have evolved over time, beginning with mostly robot-centric, task-related approaches to now incorporating considerations on user perception, interpretation, and response. Yet, existing classification approaches still face limitations, as they often rely on rigid failure categories that do not fully represent or encapsulate the highly circumstantial, overlapping ways in which robot failures are experienced and interpreted by users. A robot failure may simultaneously have functional, social, or potentially even moral implications on user perceptions and responses.

To illustrate this problem, imagine a food delivery robot that brings food and drink orders to customers' homes. Suppose the robot fails by miscalculating the handover

of the order to the customer and drops some of it on the ground. The robot's failure can carry multiple implications that do not fall into just one category, depending on the context.

Under older, purely robot-centric classifications, this event would be treated primarily as a functional failure: the robot did not successfully execute its task, and the focus would be on diagnosing what went wrong and how to prevent similar failures in the future. However, more recent user-centered classifications broaden this view by considering how such failures affect users, for example by labeling the dropped order as a functional failure if it leads the customer to perceive the robot as incompetent or unintelligent. But many of these approaches do not account for the fact that the same failure may simultaneously affect multiple dimensions of the user's cognition. If the robot's failure creates a decrease in trust from the user, such that the user does not want to reuse the robot delivery system again, the robot's failure also carries a social dimension. If the failure causes harm to the customer (e.g., by spilling a hot drink on the customer), such that it affects the user's moral judgments towards the robot (i.e., assigning blame), then the failure also has a moral dimension. While a robot failure may affect all three of these dimensions (functional, social, moral) simultaneously, it may also primarily engage one or two. In the next section, we introduce and describe a flexible, multidimensional framework that characterizes robot failures in terms of how they are interpreted by users.

2.2 Dimensions of Failure

In this section, we further elaborate on the three dimensions of failure introduced in our food delivery robot example: functional, social, and moral. Each dimension characterizes a robot failure based on its influence on users' perceptions and responses to the robot, and reflects a different form of cognition involved in how humans interpret

and react to robot failures.

Importantly, we do not adopt a robot-centric approach to define these dimensions; *how* a robot fails does not determine which dimensions are implicated. For example, even if a robot fails by violating a social norm (which might traditionally be labeled as a social failure), this does not necessarily mean that the failure will have a social dimension from the user's perspective. Similarly, if a robot fails due to a breakdown (often characterized as a functional or technical failure), this does not necessarily imply that the user will interpret the failure along the functional dimension. In our framework, the dimensions that are relevant to a particular failure are determined solely by how the user perceives and interprets the failure.

These three dimensions are not mutually exclusive. One or more dimensions may be simultaneously relevant when interpreting a robot failure, with the degree of relevance for each dimension varying depending on the context of the failure. Some failures may be primarily functional, with minimal social or moral implications, while others may strongly influence user's perceptions across all three dimensions (see the robot delivery example in Section 2.1.2, last paragraph). In addition, the dimensions may interact or be correlated for a given failure. For example, a failure that strongly implicates the moral dimension may also carry significant social consequences, as a robot perceived to have intentionally caused harm is likely to be viewed as untrustworthy.

In the following subsections, we describe each failure dimension and provide illustrative examples. To ground our discussion, we continue with the food delivery robot example throughout the section.

Functional

The *functional* dimension refers to users' cognitive evaluations of a robot's ability to perform a task, and is therefore constructed from perceptions of the robot's task

performance. This dimension captures how users asses a robot's efficiency and effectiveness, relative to their expectations of the robot's competence, intelligence, and ability to successfully complete a task.

Within the functional dimension, users may consider questions such as: How well did the robot perform this task? Did it perform better or worse than expected? Did it achieve its goal? In the context of the food delivery example, functional evaluations might include questions such as: How well did the robot do in its food delivery?

The functional dimension becomes particularly salient in emerging paradigms such as human-in-the-loop machine learning and reinforcement learning from human feedback, in which untrained users are expected to provide evaluative assessment of a robot's performance.

Social

The *social* dimension of robot failure refers to users' cognitive evaluations related to users' relationships with the robot and the robot's ability to adhere to social norms. This dimension captures evaluations such as likability and trustworthiness, which may be influenced by users' expectations of the robot's emotional intelligence and ability to foster and maintain interpersonal relationships.

Within the social dimension, users may ask questions such as: Do I like the robot? Can I trust the robot? Is the robot my friend? Related to our food delivery example, one might ask: Will I trust this robot to deliver food to me next time?

Many failures will have a social dimension when they involve violations of social norms (e.g., a robot making a joke during a serious moment or a robot interrupting a user at an inappropriate time). However, failures that are purely functional in nature (e.g., a robot dropping food on the ground) can also have a social dimension if they affect how much users like or trust the robot. Thus, it is the impact of the failure on the user, rather than the nature of the failure itself, that determines whether the

social dimension is relevant.

The social dimension is especially critical as robots occupy roles that require building bonds and fostering relationships with users, like companion or collaborator robots.

Moral

The *moral* dimension of robot failure refers to users' cognitive evaluations related to their moral and ethical judgments towards the robot. This includes perceptions of blame, responsibility, and intentionality, which may be influenced by users' expectations of the robot's moral knowledge. Moral evaluations often involve concepts of harm, fairness, and justice.

Within the moral dimension, users may ask: Is the robot morally responsible for its actions? Should it be punished for its actions? Related to our food delivery example, one might ask: Did this robot intend to spill hot coffee on me? Should it be punished for what it has done? Does it know that it is wrong to do what it did?

The moral dimension is imperative as robots increasingly interact with untrained users and may take actions that harm or adversely affect users.

Summary

In summary, the functional, social, and moral dimensions of failure involve different perceptions of the robot after it fails. The functional dimension is purely related to the robot's performance, the social dimension is related to the robot's social relationships with the user or with others, and the moral dimension is related to moral judgments of the robot. A summary of these dimensions can be found in Table 2.1.

In the remainder of this chapter, we discuss how these three failure dimensions apply to the existing literature on robot failures in human-robot interaction. We summarize key findings from prior research that highlight how contextual factors

Failure Dimension	Definition	Human Perception Examples
Functional	People's perceptions related to a robot's performance.	Robot's effectiveness, efficiency, capability
Social	People's perceptions related to their relationship with the robot and the robot's adherence to social norms.	Robot's trustworthiness, likability, friendliness
Moral	People's moral judgments of the robot.	Robot's understanding of right and wrong, punishment, intent

Table 2.1: Dimensions of Robot Failure Summary

influence the ways in which robot failures are perceived and responded to by humans. These insights motivate the research questions and studies presented throughout this dissertation.

2.3 Context Surrounding Failure

In this section, we examine how the context surrounding a robot failure is critical in shaping user perceptions and responses. We review prior research demonstrating how factors such as the nature of the failure, the robot's characteristics, user characteristics, and the robot's behavior following the failure influence responses to robot failures. Throughout this discussion, we apply our multidimensional framework to interpret these findings. Overall, this section highlights the importance of context in human-robot interactions and underscores the need for continued research to better understand how robot failures affect people.

2.3.1 The Nature of the Failure

Early work examining the effects of robot failures in human-robot interactions investigated *if* robot failures would influence people's perceptions of a robot. In a pioneering study, Salem et al. [228] compared the effects of a robot that did not fail to a robot

that "failed", by making gestures that were not aligned with its speech, while performing a joint task with participants. Interestingly, the robot that failed was perceived as more likable and more desirable to live with in the future than the non-failing robot, even though the failing robot led to decreased task performance. Studies by Mirnig et al. [185] and Ragni et al. [213] reinforced these findings, concluding that a failing robot was perceived as more likable than a non-failing robot. They suggested that this may be explained by the Pratfall effect [12], a psychological phenomenon which proposes that competent individuals are perceived as more likable after making mistakes. However, the influence of robot failure on perceived likability is not always positive. Gideoni et al. [88] demonstrated that robot failures can decrease a robot's perceived likability if failures are personally relevant to the participant. Additionally, in Ragni et al.'s experiment, despite results showing failure positively affecting robot likability, failures also led to decreases in the robot's perceived reliability, competence, and intelligence [213].

Together, these findings illustrate that robot failures can simultaneously affect multiple failure dimensions and that context is critical in how people interpret failures. Ragni et al. demonstrated that a failure that decreased perceptions of the robot's competence in the functional dimension also increased perceptions of the robot's likability in the social dimension. Gideoni et al. demonstrated that the positive effects to perceived likability are context dependent, and that the nature of the failure can make the difference. Perceptions are shaped not only *if* a failure occurs, but also *how* it occurs.

Severity & Failure Type

The severity of a robot's failure can influence how people perceive and respond to a robot. For example, in Brooks et al.'s study [39], participants read about two vacuum robot failures: one less severe failure which involved the robot partially, but

not fully, cleaning the floor, and one more severe failure, which involved the robot knocking over a plant and damaging the carpet. Brooks et al. showed that the severity of the failure significantly influenced how negatively people reacted towards the robot's failure, with more severe failures leading to more negative reactions (on survey questions related to the robot's perceived competence, trust, responsibility, and more). Similarly, Khavas et al. [124] showed that people's trust in a drone falls significantly more when a drone experiences a more severe failure (i.e., collision with a wall), when compared to a less severe one and Stiber et al. [246] showed that more severe robot failures in a Programming by Demonstration (PbD) task resulted in faster, more intense behavioral responses from users.

Furthermore, Garza [83] investigated failure severity through the lens of harm, examining how different levels of property harm and personal risk to the participant affected perceptions and responses. More severe failures led to a higher decrease in trust in the robot than less severe failures, although they noted no significant difference between high and low severity personal risk conditions, which both led to large decreases in trust. These findings are supported by Adubor et al. [2] who found that robot failures related to personal risk tend to be perceived as more severe than those related to property risk. Additionally, Rezaei et al. [216] found that a robot failure that violated moral trust significantly influenced participants' trust in a robot more than a failure that violated performance trust. Interestingly, Khavas et al. [123] found that moral violations by robot teammates led to sharper decreases in trust than human teammates' moral violations, suggesting that robots that fail morally may even be held to a higher ethical and moral standard than humans are. These findings indicate that the moral implications of a failure can have high adverse consequences on participants' perceptions towards a robot that fails.

Together, these works demonstrate that failure severity can influence people's perceptions of robot failures across all three failure dimensions.

Timing & Frequency

The timing of robot failures can shape how people perceive and respond to the robot. Desai et al. [59] investigated how the timing of a robot's failures influenced users' trust. Participants were tasked with operating a robot through a course and were able to control how much autonomy the robot had over its behaviors. Participants were placed in one of four conditions, denoted by when in the interaction the robot would fail (i.e., never, beginning, middle, end). Failures later in the interaction had significantly more negative effects on trust than those that occurred earlier. Lucas et al. [167] reinforced these findings in their study investigating failure timing on a robot's persuasion on a ranking task. Robot failures that occurred later in the interaction had a more negative effect on the robot's persuasive influence than robot failures that occurred earlier in the interaction. Additionally, Luebbers et al. [168] found that participants demonstrated a recency bias when observing robot failures and disproportionately weighed a robot's most recent behaviors when judging the robot. Participants' perceptions of trustworthiness were especially affected by recent robot failures.

The frequency of robot failures can also impact people's perceptions and responses to the robot. In a human-robot collaboration setting, Van Waveren et al. [268] observed that an accumulation of less severe robot failures damages trust more than one severe failure. Liu et al. [163] also found that the recurrence of failures can lead to evolving user behavioral responses and reactions. The first few failures correlated with user confusion, but a continuation of failures later correlated with frustration.

These studies suggest that timing and frequency can influence both the functional and social dimensions of a failure.

Robot & Human Roles

The roles adopted by both the robot and the human can influence how people perceive and respond to robot failures. Chi and Malle [50] conducted a between-subjects experiment in which participants were assigned to an interactive teacher or supervisor role while training a virtual healthcare robot assistant. Participants in the teacher role were more resistant to initial trust loss following a robot failure, more willing to trust the robot on subsequent tasks, and more likely to attribute the robot's improvement to their own efforts compared to those in the supervisor role. In a related line of work, Karli et al. [114] examined how power dynamics associated with a robot's role influenced people's trust and compliance after robot failure. Participants were more willing to comply with a supervisor robot than a subordinate robot, even when the robot was failing. They also trusted a supervisor robot that attempted to repair trust verbally more than a subordinate robot using the same repair strategy.

Relational expectations and power dynamics contribute to the context that people use in their evaluations of robot failures. Thus, these findings demonstrate the importance of human and robot roles in shaping user perceptions and responses after failures, particularly in the social dimension.

2.3.2 Robot Characteristics

Robot characteristics are often related to the robot's degree of anthropomorphism, including human-like qualities expressed through its physical appearance, behavior, and overall characterization.

Kontogiorgos et al. [137] manipulated the embodiment of a physical agent by using a smart speaker or a human-like robot to collaborate with participants on a cooking task. The agents committed various failures of the same severity and frequency, such as providing incorrect guidance, failing to respond, or disengaging from the task. Participants rated the human-like robot more positively than the

smart speaker in terms of perceived intelligence, competence, and social presence. Participants' intention to interact with the smart speaker also decreased following failures, while this decline was not observed for the human-like robot. Participants' behavioral responses also differed; they used more speech signals toward the speaker but directed greater visual attention toward the human-like robot [135].

Despite Kontogiorgios et al.'s findings, an online study by Choi et al. [51] concluded that a robot's human-like appearance does not always result in more positive user perceptions following failure. Participants imagined interacting with a humanoid or a non-humanoid robot that committed process failures (e.g., slow service, awkward behavior) or outcome failures (e.g., delivering the incorrect food order). Participants were more dissatisfied when humanoid robots committed process failures compared to non-humanoid robots that committed the same failures. Choi et al. attributed this effect to increased social expectations inherently associated with humanoid robots, which can explain why apologies from humanoid robots were more effective in increasing perceived warmth than apologies by non-humanoid robots. Therefore, while a human-like appearance can make a robot more relatable and likable, it can also put additional pressures on the interaction for the robot to perform in a certain way, affecting user interpretations of robot failure.

Beyond physical appearance, user expectations can form through previous interactions or framing of the robot's capabilities. For example, Kassem et al. [116] found that human teachers stopped teaching robots with lower pre-existing skills sooner than those with higher initial proficiency, suggesting that expectations biased their subsequent behaviors. This lack of patience with perceived less skilled robots was also seen in a study by Morales et al. [188], where participants who had previously witnessed a robot commit a failure were significantly less willing to help it in the future compared to participants who had not witnessed the failure [188]. More generally, Washburn et al. supports these conclusions by showing that the framing of a

robot's functionality before interaction can affect how people's trust changes after a failure [277]. Rossi et al. [225] found that robots displaying theory of mind behaviors was perceived as more reliable and was trusted more than the robot without a theory of mind. These ideas are further expanded by Claure et al., who saw that framing of the robot's perceived agency and intention can affect how fairly people perceive a robot after it commits harm [54]. Similarly, Stower et al. [248] found that a robot's perceived risk-tolerance and task competence at a task affected user's trust in a robot that fails and succeeds. Last, Asavanant et al. [14] found that a robot's perceived likability affected how users perceived potential personal space violations by a robot.

Together, these findings demonstrate that perceived robot characteristics influence how failures are evaluated across functional, social, and moral dimensions.

2.3.3 User Characteristics

As robots are placed in more diverse environments and interact with users who differ in age, culture, personality, and beliefs, concerted efforts have been made to investigate how individual characteristics shape responses to robot failures.

Age

Although much of the existing robot failure work focuses on adults (18+ years of age), the field has begun to explore the effects of robot failures on children [156, 84, 32, 249, 282, 207, 249] as robots take on roles like companions [195, 200] or tutors [232, 119]. Developmental differences may influence how children interpret and respond to robot failures, though this area remains underexplored.

In one of the first studies of its kind, Lemaignan et al. [156] investigated the effects of "unexpected robot behaviors" (failures, by our definition) on children's (4-5 years old) perceptions of a robot. Pairs of children interacted with a robot that delivered dominoes to them while they played a game, while occasionally committed failures

such as getting lost or disobeying. Children who interacted with the failing robot showed higher engagement than those who interacted with a non-failing robot. While Lemaignan et al show the effects of age on behavioral responses, Geiskkovitch et al. [84] present how failures can be affected in the social dimension. They found that when young children (3-5 years old) interacted with two robots: one that committed informational failures and one that did not, they trusted the non-failing robot more than the failing one, although this difference was only significant on a simple object selection task. Although these results displayed negative impacts of failures on children's perception of robots, robot failures may have positive effects, as seen in a study by Bowman et al. [32]. Eight to eleven year old children were placed in a learning-through-teaching role. Instead of disengaging with a robot student that made "atypical failures", as expected from the results of Lemaignan et al., the children generated more teaching strategies and demonstrated greater learning gains when they taught these robot students. Importantly, children can differentiate between human and robot groups and may perceive robot and human failures differently. Stower et al. [249] found that three to six year old children evaluated a failing robot differently from a failing human; children were more likely to believe that the human failed on purpose than the robot.

Children's understanding of the world evolves across developmental stages, yet there is still not a complete understanding of how these changes shape responses to robot failures. In their study investigating the effects of different child age groups (i.e., 6-7 years, 8-10 years, and 10-12 years), Wrobel et al. [282] found that age group may significantly influence children's responses, with children at different developmental stages focusing more on different types of failures. Although these studies all indicate that age should be factored in understanding user perceptions and responses to robot failures, more research is needed to fully elucidate the developmental intricacies.

Culture

Culture can influence user perceptions of robot failures, as shown in Zhang and Lee's cross-national, online study with 330 participants from the U.S. and 368 participants from South Korea [289]. Zhang and Lee examined the effects of different types of robot failure (e.g., logic, semantic, syntax) and trust-repair strategies on participants' trust. They found significant cross-cultural differences, with failures being significantly more detrimental to competence-based trust among Korean participants than among American participants. Additionally, cultural differences influenced trust-repair strategy perceptions; compared to Americans, Koreans more positively received the internal-attribution robot apology (i.e., blaming itself for the failure) than denial. Koreans and Americans also blamed the robot for its failures differently, reflecting the contrast in Korean and American values, as Korea is a more collective society compared to American ideals of individualism.

Personality & Beliefs

Individual differences in personality traits and beliefs also influence how people perceive a robot that fails. Esterwood et al. [67] observed that a person's pre-existing attitudes toward robots can influence which trust repair strategies are most effective to them. People's general disposition and attitude towards other people can also indicate how they will perceive a robot failure, as Aliasghari et al. [5] found that a person's disposition to trust other people correlated with how much they trusted a failing robot. Rossi et al. [226] extended this work to an emergency scenario, where they found that personality traits such as conscientiousness and agreeableness, in addition to a person's disposition to trust other humans, were significantly correlated with participants' tendency to trust a robot.

2.3.4 Robot Behavior Following Failure

Robot behavior following a failure that is intended to reduce its negative effects is commonly referred to as a mitigation or failure-recovery strategy. Within HRI research, a substantial sub-field focuses on how robots should behave after a failure to repair lost trust, oftentimes referred to as trust-repair strategies (see [68] and [17] for detailed reviews). In this section, we will provide an overview of the potential effects that apologies, explanations, and other social behaviors can have on users' perceptions and responses to a robot after it fails.

Apologies

An apology is a behavioral construct that typically involves an admission of wrong-doing, acceptance of responsibility, restitution, and/or an expression of sincere regret [236]. In human-human interactions, apologies are a powerful social tool with the potential to mitigate the negative consequences of failure, such as repairing lost trust following harmful or inconvenient behavior [230]. Apologies can be verbal (e.g., saying "I am sorry") or non-verbal (e.g., giving somebody a hug) and their impact can depend on contextual factors, such as the relationship between the individuals involved, the severity of the failure, and the perceived sincerity of the apology.

The impact of robot apologies has been widely studied in HRI. Chang et al. [46] found that a robot apology can help mitigate or recover lost trust with a user following a failure. However, social nuance plays an important role in determining an apology's effectiveness, as shown by Pompe et al.'s study [211] that found that remorseful robot apologies rebuilt lost trust more strongly than apologies without remorse or the absence of an apology altogether. Similarly, Maehigashi et al. [173] found that contextual details matter in non-verbal robot apologies, as the details of the movement mattered when a robot bowed during an apology. The robot's behavior following the apology matters as well – Nesset et al. [197] noted that making a promise

and then breaking it led to a higher decrease in trust than offering a general apology. Mitigation strategy preferences can also vary between individuals, as Lee et al. [152] concluded that some participants prefer an apology while others prefer compensation. These works demonstrate that an apology must be refined in style and delivery to best regain user trust. By influencing trust, robot apologies primarily engage the social dimension of failure.

Explanations

Explanations, communicative behaviors that provide users with information about a robot's actions, can play a significant role in shaping how robots are perceived [122, 121, 196, 157, 58, 136, 7, 271, 3, 158]. Lyons et al. [169] found that when robots provide explanations for why a failure occurred, participants experience a smaller decrease in trust towards the robot. LeMasurier et al. [157] found that robots that provided explanations before a failure fostered higher perceived intelligence and trust than robots that provided explanations after a failure. Hald et al. [100] observed that explanations were not always sufficient to increase lost trust following a robot failure, while Kox et al. [141] noted that transparency after failure helped keep trust higher than when there was a lack of transparency. However, Aroyo et al. found that increased transparency can reduce interaction quality in some cases [13]. Overall, explanations can influence both the functional and social dimensions of failure, although their impact depends heavily on contextual factors.

Other Behaviors

Other robot behaviors following a failure, such as humor, profanity, or blame, have also been investigated. Green et al. [96] found that robot humor following a failure can influence people's perceptions of the robot's competence. Shippy et al. [233] revealed that, while robot profanity had limited impact after a failure, it can be

perceived as relatable or humorous. Finally, van der Hoorn et al. [267] shared that robots that attributed blame to themselves, rather than to a human collaborator, were perceived more positively. Although these behaviors are social in nature, they can affect both the functional and social dimensions related to failure.

2.4 Summary

In summary, the context surrounding a robot's failure can significantly influence how people perceive and respond to the robot, and therefore which failure dimensions are affected. In the following chapters, we present a series of human-subjects experiments designed to expand our understanding of how people perceive and respond to robot failures within the three dimensions of failure that we have described in this chapter: functional, social, and moral. Each chapter primarily focuses on the effects of robot failure within one of the three dimensions and adds to our knowledge of how human perceptions and responses are influenced within these different HRI failure contexts.

We begin by examining how people provide evaluative feedback in an interactive human-robot teaching setting. This study contributes to our understanding of how individuals differ in their interpretations and evaluations of robot failures in the functional dimension. We continue by investigating failures in the social dimension, by examining changes in people's trust in a collaborative setting where the robot provides advice. In this context, we investigate the influence of peoples' age by investigating differences between children aged 4-7 years old and adults, as well as different robot behaviors following failure. Lastly, we investigate failures in the moral dimension by examining how a robot's backstory influences moral judgments after the robot causes physical harm. In studies investigating the social and moral dimensions, we directly compare the robot's failures to a human's.

Chapter 3

The Functional Dimension: Humans' Evaluative Feedback to Robot Failures

As discussed in the previous chapter, robot failures can influence how people perceive and respond to a robot with respect to its task performance. However, there remains limited understanding of how individuals differ in their evaluations of robot failures. In this chapter, we broaden our understanding of the effects of robot failures within the functional dimension by investigating how people provide scalar evaluative feedback in an interactive human–robot teaching context. This context is used for investigating the functional dimension because it is centered on how people evaluate the robot's performance. Specifically, we study how participants ($N = 36$) assign numerical feedback to a robot as it attempts a card game task and commits various task-related failures.[1] We found that participants employed different partial credit feedback strategies for robot failures during the task (i.e., participants varied in how

[1]Portions of this chapter were originally published as: **Nicholas C. Georgiou**, Shuangge Wang, Joel Banks, Kate Candon, Dražen Brščić, and Brian Scassellati. (2025). When Teaching A Robot, People Employ Different Feedback Strategies: Some Are More Effective Than Others. In *Proceedings of the 47th Annual Conferences of the Cognitive Science Society.* [86]

they scored the same robot failure actions). We then used the feedback from each participant to generate extrapolated feedback strategies. In simulations, we found that training a supervised machine learning model with these different extrapolated feedback strategies influenced how well the model was able to perform the task. Models trained with labels from some reasonable strategies significantly outperformed models trained with labels from other reasonable strategies. Participants' familiarity with machine learning, artificial intelligence, and the task did not significantly affect how well the model trained with their extrapolated feedback performed on the task. These findings suggest that people will value different factors when responding to robot failure and this observation can have implications for transferring learning algorithms into the real world.

3.1 Introduction

Numerical evaluations that assess the quality of an action or behavior have been utilized as a user-friendly modality through which people can teach a machine learner [257, 130, 48]. When users are tasked with providing numeric feedback for evaluation, they must determine a feedback strategy, i.e., *what* and *how* different factors weigh into the score that they provide each time. Feedback strategies will likely vary between teachers, as people have different expectations, experiences, prior knowledge, and mental models related to teaching.

Nonetheless, as long as a teacher is providing consistent feedback that is aligned with the learning goal, what we refer to as a *reasonable* feedback strategy, one would expect that such a strategy would be sufficient to teach the learner the task. However, the downstream effects of different reasonable feedback strategies that people use are not obvious. It is unknown how different feedback strategies while training will impact how effectively and efficiently a learning algorithm performs a task. This question is

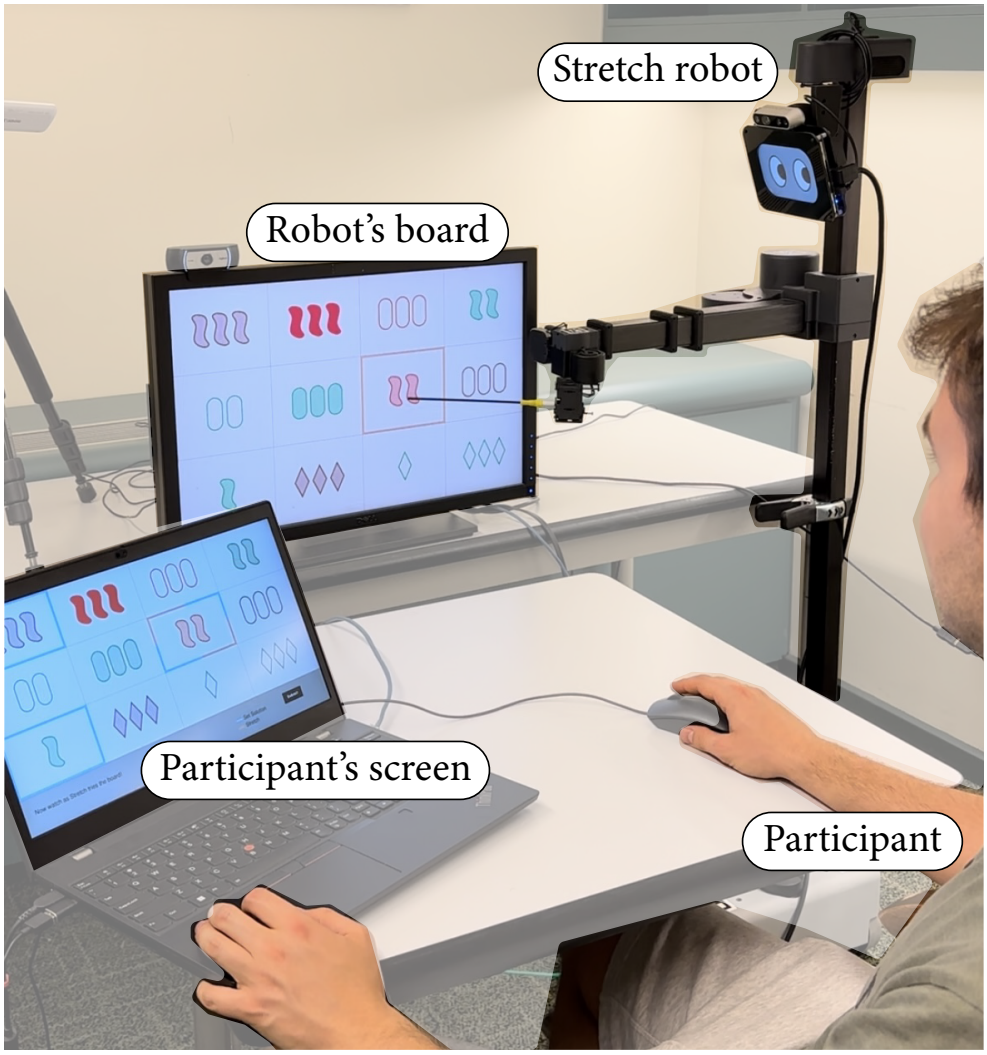

Figure 3.1: Experimental setup.

particularly important to study as machines are put into situations where they will be expected to learn directly from human feedback from different teachers.

To explore these ideas, we conducted an in-person user study in which participants were tasked with providing feedback to a robot through numerical evaluations. The experimental setup can be seen in Fig. 3.1. We first explored the feedback different participants provided throughout the task. Using the collected data, we fit a linear regression to each participant's collection of feedback to highlight the different feedback strategies that people employed. We considered these linear regressions to be *extrapolated feedback strategies* because they can provide an estimate for how

participants would score unseen actions. To then explore the downstream effects of different reasonable strategies, we ran simulations in which each participant's extrapolated feedback strategy was used to train a supervised machine learning model to predict proxy reward labels for every possible robot action in a given state, and we measured how these predicted labels affected the models' task performance. Lastly, we discuss what our results show about the different feedback strategies that humans used to teach a robot a particular task, as well as potential real-world implications for our findings.

3.2 Related Work

Traditionally, machine learning (ML) algorithms have relied on ground truth training data, defined or curated by programmers, to learn a task. In reinforcement learning, this ground truth takes the form of pre-specified environmental rewards, while in supervised learning, it consists of verified, labeled data points. However, as robots are deployed in new environments and expected to learn new tasks, developers cannot manually design environmental rewards or curate data for every possible task. Instead, new human-in-the-loop ML paradigms aim to address this challenge by enabling end users to train robots through their own feedback [189].

Evaluative feedback – values provided by a human teacher to assess a learner's behaviors and actions – can be used as a reward in interactive reinforcement learning or as labeled training data in interactive machine learning (ML) [189]. While some work has offered approaches to incorporate human-provided feedback into the robot learning loop, such as TAMER [129], COACH [172], and REPaiR [120], these methods primarily focus on algorithm development. Most work does not focus on conducting controlled user studies to systematically investigate *how* users provide feedback when teaching a robot, even though understanding and analyzing these differences is critical

for effectively incorporating users' feedback into the ML loop [256].

Recent work in human-robot interaction (HRI) has looked into how people's feedback changes over time [42], how robot competency affects feedback quality [274], how binary feedback differs from scalar feedback between users [288], and how feedback can be broken into multiple dimensions [106]. Despite these advancements, insights into how people's underlying feedback strategies differ and how these strategies affect learning are understudied.

Prior work has studied how and when people employ different training strategies, defined by how much they focus on allocating penalty vs. reward [165, 164]. Other work focuses on modeling humans' preference-based feedback [151] and evaluating how synthetic and human labelers differ when training preference-based reinforcement learning models [183]. Much of the work in this area studies how to learn varying preferences among people (i.e., different end goals with respect to what the user wants). Instead, we are studying how people provide feedback in the same task (i.e., same end goal) via different approaches, as well as effects of these approaches on learning.

Based on the gaps identified in prior work, our guiding research questions were:

RQ1: When given the same teaching objective with the same end goal, do the feedback strategies that people employ to teach a robot vary?

RQ2: Do the feedback strategies extrapolated from different human teachers impact how effectively and efficiently an ML algorithm can learn to complete the task?

3.3 Methodology - Data Collection

To investigate our research questions, we asked participants to provide feedback to a robot performing a task. Importantly, we required participants to provide numeric evaluations, but did not instruct them on the strategy they should use to determine

their scores. The robot did not learn during the interaction, but the data was used to train models in simulation (see Section 3.6). The following subsections describe the details of our data collection methodology.

3.3.1 Teaching Task - Set

We chose Set, a card selection game, as our teaching task. The main objective of Set is to select three cards that meet certain requirements, classifying them as a *set*, out of a board of 12 cards. Each card has four feature dimensions: number, shape, color, and fill. Each feature dimension can have one of three distinct values (see Fig. 3.2 for examples). To be considered a set, for each of the four dimensions, the three cards must either have the same value or three different values.

For our version of Set, we use three row by four column boards, where each board includes only one set, that is, only one solution[2]. We represent the board space as a $3 \times 4 \times 3^4$ dimensional space. There are $\binom{12}{3} = 220$ three-card selections for a Set board with twelve cards, with only one of these selections constituting a success (the selection of the set).

Set was carefully selected to help us explore the potential effects of different strategies on learning. Set is very controllable — it is straightforward to verify solutions and to generate unique states that satisfy various constraints. The game can be explained quickly, yet is cognitively challenging. It is simple enough that there are a limited number of likely feedback strategies, but complex enough that there is not one strategy more obvious than others.

Based on our personal experience playing Set and on a pilot study, we hypothesized that there were two main components that people would focus on when determining what feedback score to provide to the robot: 1) the number of cards, c, from the board solution included in the three-card selection, and 2) the number of feature dimensions

[2]We constrained the boards to have only one solution to keep the focus on measuring how participants evaluated robot actions towards the same goal for each board.

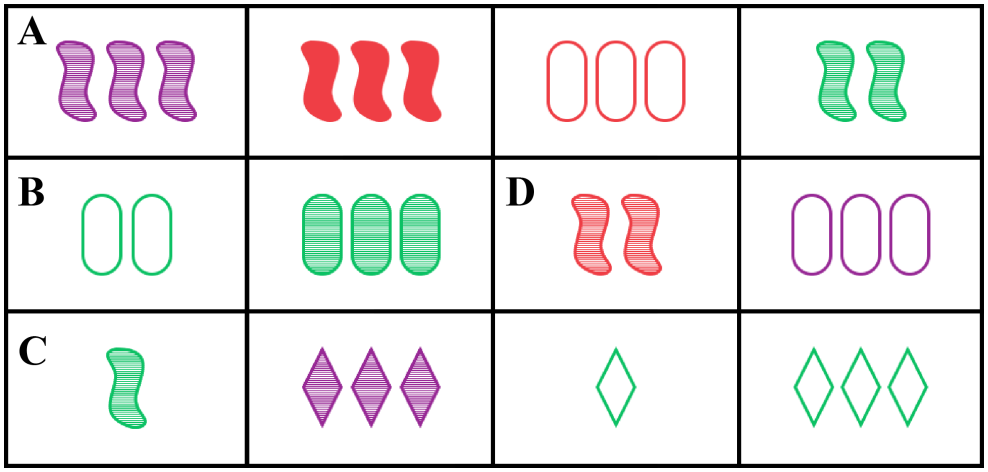

Figure 3.2: Set board example. Cards A, B, and C are not a set. Despite each having a different number, they break the requirements of a set for three dimensions (shape: 2 squiggle, 1 oval; color: 2 green, 1 purple; fill: 2 striped, 1 outlined). The action of selecting cards A, B, and C can be represented as $[c = 2, f = 1]$: two cards are from the solution, one of the dimensions (number) meets the set requirements. Cards A, C, and D are a set because they have the same shape and fill and have different colors and numbers. Cards A, C, and D can be represented as $[c = 3, f = 4]$.

in the selection, f, that fulfill the Set rules criteria. A three-card selection with a high c value indicates that the selection is close to the specific solution for a current board, whereas a high f value indicates that a selection is close to fulfilling the criteria for feature dimensions with respect to the rules of the game. Scoring proportionally to c and/or f are all reasonable strategies, and it is not obvious which strategy people would use, nor which is more effective in providing feedback for machine learning.

Importantly, in Set, c and f can vary independently. A three-card selection can include cards from the board solution (high c), but have no feature dimensions achieving the rules of Set (low f), or vice-versa. Alternatively, a three-card selection could be high in both c and f (e.g., two cards from solution, third card breaks Set rules on only one feature dimension), or could be low in both c and f (e.g., three cards not in solution and breaking rules on all four dimensions).

Therefore, we have $c \in \mathbf{C} = \{0, 1, 2, 3\}$ and $f \in \mathbf{F} = \{0, 1, 2, 3, 4\}$. By definition, $c = 3$ if and only if $f = 4$, because if the three cards from the solution are selected, then they meet the Set requirements on all four dimensions. Thus, $c = 3$ and $f = 4$

is the success action for the robot. There are $|\{\mathbf{C}\backslash\{3\}\} \times \{\mathbf{F}\backslash\{4\}\}| = 12$ failure variations of three-card selections. From this point forward, a three-card selection from the robot will be represented in terms of c and f. See Fig. 3.2 and Fig. 3.3 for examples.

3.3.2 Participants

We recruited 45 participants for data collection. Nine participants were excluded due to technical issues. The final 36 participants reported their age ($M = 25.83$ years, $SD = 4.65$ years) and gender (15 male, 21 female). 27 participants were undergraduate or graduate students. When self-reporting familiarity with ML, 8 participants reported that they knew it well, 4 knew a fair amount, 14 knew a little, and 10 had heard of it. For familiarity with AI, 6 reported that they knew it well, 7 knew a fair amount, 17 knew a little, and 6 had heard of it. For prior experience with Set, 13 reported that they had played before, 22 had not, and 1 was unsure.

All participants provided consent for the study and audio-visual recording. Participants received $15 as compensation. The study was reviewed and approved by an Institutional Review Board. The study took approximately one hour.

3.3.3 Robot

We used a Hello Robot Stretch 2 [118] in our Set task. Stretch is a lightweight mobile robot with a manipulator. We replaced the default end effector gripper with a simple pointer to suit our task.

To make Stretch appear more social, we added a small HDMI monitor to its head that displayed two eyes, based on the Shutter robot [258], as seen in Fig. 3.1. With the help of this monitor, the Stretch performed anthropomorphic behaviors, such as blinking, nodding as a greeting or as a response to the participant's feedback, peering towards the participant while selecting a card, expressing happiness or sadness, and

turning/tilting its head to and from the participant and Set board.

3.3.4 Procedure

In our data collection study, participants provided numeric feedback to the Stretch robot as it was playing Set. Participants provided a score to the robot after each three-card selection on a board. All data collection sessions were conducted by the same experimenter, who used a script for consistency. The procedure consisted of five phases.

Pre-Interaction Phase: Participants completed a consent form and demographics survey outside of the study room.

Tutorial Phase: The participant was brought into the study room by the experimenter and completed a tutorial explaining the rules of finding a set on a laptop in the room. After the participant completed the tutorial, the experimenter explained how to play Set. For each round, participants were shown a 3×4 card board on the laptop screen (participant's screen in Fig. 3.1), which contained only one solution. When participants selected cards on the laptop screen via mouse clicks, they were highlighted in blue. After they submitted their three-card selection, the system confirmed whether or not the participant found the set. If the participant did not find the set within one minute, the solution was revealed to the participant. After the participant found the set or their time expired, they rated the perceived difficulty of the board on a scale from 1 (very easy) to 10 (very hard).

Robot Introduction Phase: The experimenter explained to the participant that their goal was to help the robot learn Set by providing feedback as it tried to solve each board. Participants were told to provide feedback after each three-card selection made by the robot. They were asked to provide a score on a scale from 1 to 10, with 1 being very poor and 10 being excellent, via the laptop keyboard. They were also told that they should feel free to speak with Stretch throughout the study.

They were told that the robot would not adjust its behavior based on their feedback during the session, but that the feedback would be used to help the robot learn in the future.

After the interaction protocol was explained, the participant pressed a button on the laptop to wake up the robot. The robot opened its eyes and turned its head to the experimenter. The experimenter greeted the robot by saying “Hello Stretch". In response, the robot displayed happy eyes and nodded its head. The robot then turned its head to the participant, displayed happy eyes and nodded its head. The experimenter told the participant to press a button to begin the feedback interaction and left the room.

Robot Interaction Phase: The participant and Stretch then went through eight different Set boards. For each board, the participant first submitted a card selection, followed by a difficulty score. The correct solution was then highlighted in blue on the participant's screen. Next, Stretch selected three cards on its own board, which was identical to the participant's, but did not show the true solution. The robot's board was displayed on the monitor in front of the participant and to the side of Stretch (see Fig. 3.1). The robot's arm was used to select cards on the board. When pointing to a card, the robot's head faced the board, the robot's arm moved to the desired card, the robot's face rotated and its eyes peered towards the participant, and its wrist rotated such that its pointer appeared to touch the card. At the end of this motion, the representative card on both the robot's and participant's boards were highlighted in orange. After Stretch selected three cards, it turned back to the user and waited for a ring or buzzer noise that indicated whether its selection was correct or incorrect. The robot reacted with happy or sad eyes depending on the noise. The cards that the robot selected were predefined (see Robot Action Sequences section). The participant provided a feedback score to Stretch for each three-card selection that the robot made. The robot nodded to the participant to acknowledge the receipt of

Board	$[c, f]$	**Robot Action Sequence** $[c, f]$
1	$[0, 0]$	[0,0] → [0,0] → [0,0] → [3,4]
A	$[0, \uparrow]$	[0,0] → [0,1] → [0,2] → [0,3] → [3,4]
B	$[2, \uparrow]$	[2,0] → [2,1] → [2,2] → [2,3] → [3,4]
C	$[\uparrow, 0]$	[0,0] → [1,0] → [2,0] → [3,4]
D	$[\uparrow, 3]$	[0,3] → [1,3] → [2,3] → [3,4]
E	$[\uparrow, \uparrow]$	[0,1] → [1,2] → [2,3] → [3,4]
F	$[\uparrow, \downarrow]$	[0,3] → [1,2] → [2,1] → [3,4]
G	$[\downarrow, \uparrow]$	[2,1] → [1,2] → [0,3] → [3,4]

Table 3.1: Robot Action Sequences. The second column indicates the type of improvement exhibited by the robot during the sequence. A number indicates that the c or f value stayed consistent until the set was found, an up arrow indicates that the value increased, and a down arrow indicates that the value decreased. The third column shows the detailed sequence of robot actions, represented in the $[c, f]$ space.

the feedback. Stretch attempted the board, failing either three or four times, until it finally found the set. Then, the participant moved on to the next board. Stretch waited for the participant to complete their turn, and then it attempted the new board that the participant had just attempted. After all boards were completed, Stretch closed its eyes and put its head down, indicating that it had returned to sleep.

Post-Interaction Phase: The experimenter used semi-structured questions to ask about the participant's approach to determine feedback scores for the robot. Then, the participant completed survey questions about their interaction.

3.3.5 Robot Action Sequences

For each Set board, there was a predefined sequence of three-card selections for the robot, which we will call a *robot action sequence*. Except for the first tutorial board, the robot showed different types of improvements before finally arriving at the solution. The types of improvements were driven by c and f such that at least one of c or f increased throughout the sequence. Each participant saw the same eight boards, each with a fixed robot action sequence (see Table 3.1). After Board 1, Boards A-G were presented in randomized order. See Fig. 3.3 for examples.

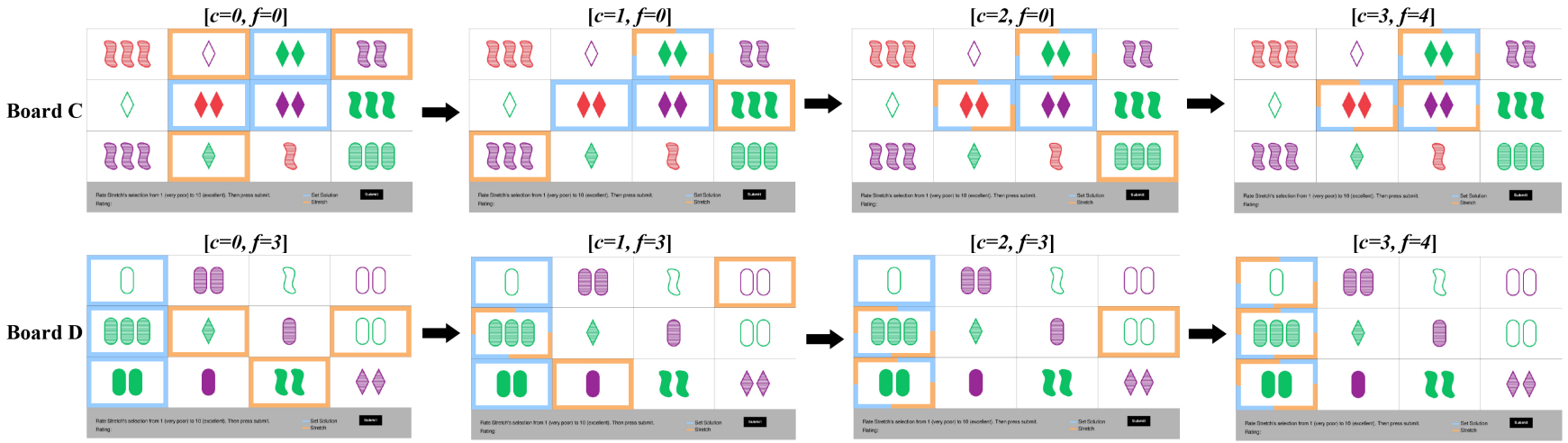

Figure 3.3: Robot action sequences for Boards C and D. Sequences progress left to right. The cards in the board solution (i.e., the set) are outlined in blue. For each action, the robot selects three cards, which are outlined in orange, or in both orange and blue if the card is also in the solution. Each robot action is represented by c, how many cards from the solution it contains, and f, how many of the feature dimensions satisfy the Set rules. Each action's $[c, f]$ is shown above the board. Board C shows the robot getting closer to the board solution in terms of c, but consistently doing poorly in terms of f. Board D shows the robot also getting closer to the board solution in terms of c, but consistently doing well in terms of f.

Figure 3.4: Distribution of participants' feedback scores.

3.4 Results - Data Collection

In this section, we present results on the numeric feedback provided by participants during their interaction with Stretch.

3.4.1 Feedback Score Statistics

Overall, participants provided a mean score of 5.53 ($SD = 3.18$) for the robot's actions. Participants used most of the unique numbers ($M = 7.33, SD = 1.55$, min $= 4$ unique numbers, max $= 10$ unique numbers) on the feedback scale. On successes, participants gave the robot perfect scores most of the time ($M = 9.87$, $SD = 0.58$). On failures, participants provided a variety of scores with a mean of $M = 4.20$ ($SD = 2.36$). The distribution of participants' feedback scores, broken down by success and failure, are shown in Fig. 3.4.

Duration	Boards (N)	Difficulty Rating	Robot Actions (N)	Feedback Score
$< 30s$	71	3.07 ± 1.53	302	5.43 ± 3.31
$\geq 30s, < 60s$	44	5.27 ± 1.69	194	5.59 ± 3.12
incomplete	173	7.35 ± 2.00	728	5.55 ± 3.15

Table 3.2: Mean ± SD of self-reported difficulty ratings and feedback scores provided to the robot, grouped by how long it took each participant to find the solution on each board.

3.4.2 Difficulty Ratings

Participants provided an average difficulty score of 5.92 ($SD = 2.56$) across all boards. Table 3.2 shows the difficulty ratings and feedback scores, broken down by how long it took participants to complete a board (or not). Pearson's correlation showed a strong positive correlation between board completion time (incomplete = 60s) and difficulty rating ($r = 0.72, p < 0.001$), but no correlation between board completion time and feedback score provided to the robot ($r = 0.03, p = 0.37$) or between difficulty rating and feedback score ($r = 0.01, p = 0.65$).

3.4.3 Participant Feedback Strategies

To examine the feedback strategies that participants used to determine feedback scores, we fit a linear regression for each participant that predicted the feedback score as a dependent variable, with c and f as independent variables (β_c and β_f are the learned coefficients). These participant models approximated feedback scores well with an average $R^2 = 0.84$ ($SD = 0.09$, min = 0.61, max = 0.98). Every participant model was statistically significant ($p < 0.001$), suggesting that these regressions reasonably align with what participants considered when determining their scores. Each participant's β_c and β_f can be seen in Fig. 3.5. This plot demonstrates that participants weighed factors differently when determining what score to provide as feedback to the robot. Additionally, Fig. 3.5 illustrates that there is no apparent pattern with respect to self-reported familiarity with ML. There was also no observed pattern in

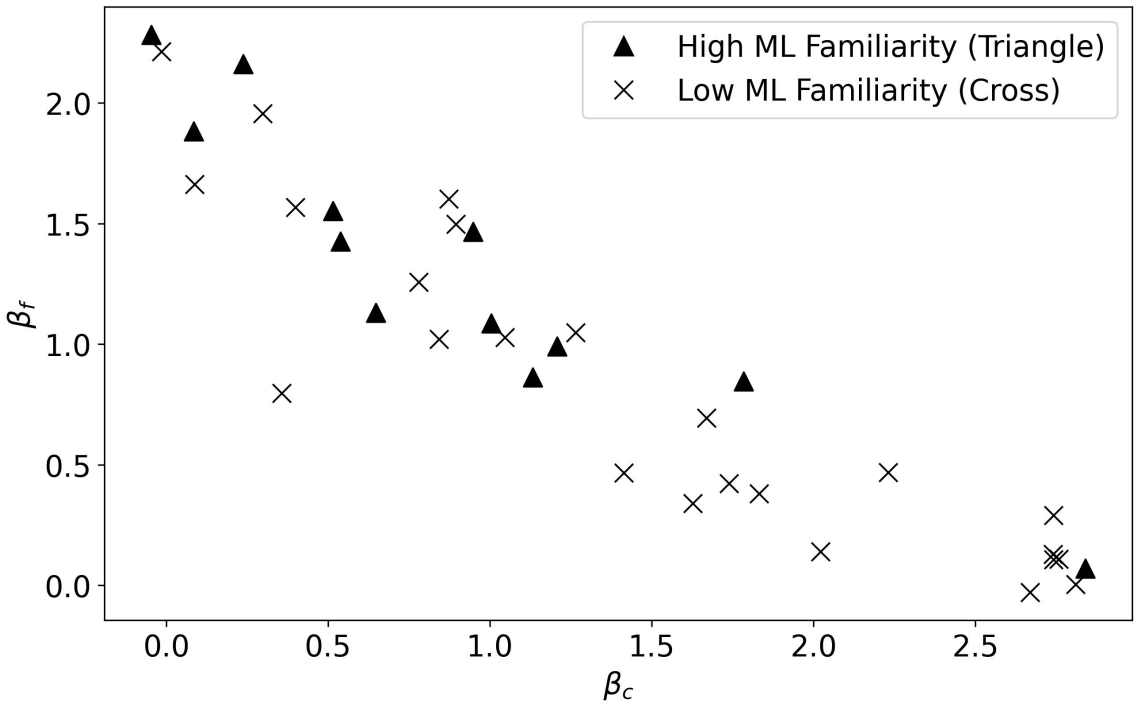

Figure 3.5: Participants' linear regression coefficients, labeled by self-reported familiarity with ML. High familiarity = knew it well or a fair amount; low familiarity = knew it a little or had heard of it.

plots with familiarity with AI and Set.

Further indicative of the idea that different people used different teaching strategies, not all participants' β_c and β_f were found to be significant ($p < 0.05$) in the linear regression predicting their feedback scores. More specifically, seven participants only valued β_f, nine only valued β_c, and 20 valued both. These different strategies were consistent with how participants explained their reasoning for choosing what score to provide for feedback in their post-interaction interviews.

3.5 Methodology - Simulations

We wanted to investigate how training using different strategies would affect an ML algorithm's ability to perform the task. The linear regressions from the previous section use handcrafted input features, c and f. However, we wanted to train models

without this handcrafted domain knowledge. As the input dimensionality increases, the amount of data required to train ML algorithms increases [140], so it becomes intractable to collect sufficient data from real humans in an in-person user study. Therefore, we used the linear regressions as *extrapolated feedback strategies* (one per participant) to automatically generate proxy feedback scores for given board-action pairs. For each extrapolated feedback strategy, we trained an individual multilayer perception (MLP[3]) to predict the proxy feedback score for a given board-action pair. We used a reduced space of the Set board by having 8 cards to make data curation more tractable. We sampled a training dataset of 10 million boards, each containing only one *valid* set with random cards and robot action. To evaluate the model's accuracy on a board, we searched the action space for the action that generated the highest proxy feedback score using the MLP, and checked if that action matched the true solution.

3.6 Results - Simulations

In this section, we present the results from MLPs that were trained using different extrapolated feedback strategies. There were very large differences in task performance, depending on which extrapolated feedback strategy was used to train the MLP. The average post-training accuracy on 1000 unseen boards for the trained MLPs was 62.9% (SD = 41%, max = 100%, min = 3.2%). The average accuracy throughout training (where accuracy was reported every 1000 batches) also greatly varied between extrapolated feedback strategies for training (M = 37.3%, SD = 27%, max = 77%, min = 1.6%) The smoothed accuracies during training (window size=10 batches) re-

[3]The state and action representations are converted into a one-hot encoding and concatenated before being fed into the MLP. The MLP outputs a scalar feedback score, which is evaluated using mean squared error (MSE) loss. The MLP consists of four hidden layers with 512, 256, 128, and 64 neurons, respectively, and employs ReLU activation [80]. The model is trained for 15 epochs with a batch size of 512, and optimization is performed using the Adam optimizer with a learning rate of 0.001 [128].

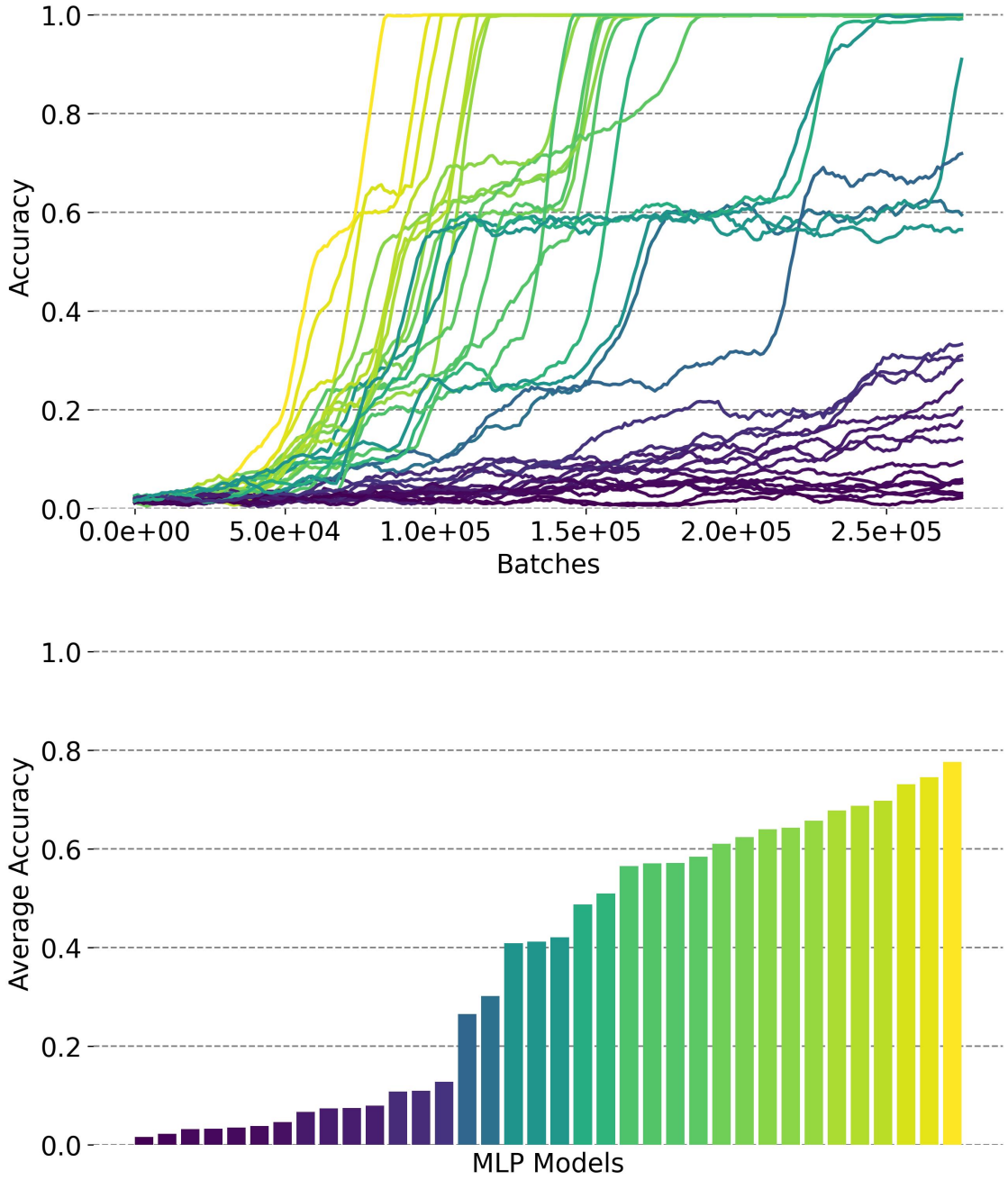

Figure 3.6: Smoothed accuracy on unseen test boards during training over time (top) and time-averaged accuracy (bottom) of each MLP model, trained using a different extrapolated feedback strategy, labeled by the same color in both figures.

ported every 1000 batches can be seen in Fig. 3.6, and bar graphs showing the average accuracy of each MLP can be seen in Fig. 3.6.

3.6.1 Familiarity with ML, AI, and Set

To explore if the participants' self-reported familiarities with ML, AI, or the task (i.e., Set) had a significant impact on how well the MLP trained with their extrapolated feedback strategy performed the task, we mapped each MLP to the participant whose extrapolated feedback strategy was used to train it. Then, we grouped the MLPs' post-training accuracy results by the participant's familiarity (or not) with ML, with AI[4], or with prior experience (or not) playing Set. We performed Mann-Whitney U tests[5]. We did not find significant differences when comparing the accuracy of the MLPs separated by familiarity with ML ($p = 0.13, U = 99.5$), with AI ($p = 0.25, U = 114.5$), or prior Set experience ($p = 0.27, U = 111.0$).

3.7 Discussion

In this chapter, we investigated how people provided numeric feedback to evaluate a robot attempting a task. Importantly, all participants provided partial credit to the robot (i.e., they did not provide the same value for every failure action of the robot). This is important because machine learning algorithms prefer granular feedback to learn more effectively and efficiently. Additionally, the ways in which participants provided this partial credit differed among participants – they valued different factors when they provided a score for the robot's failure actions. In our simulations, the drastic differences in task performance of some models compared to others suggest that not all partial-credit feedback strategies were equally effective or efficient, even

[4]To be considered familiar with ML or AI, a participant had to self-report that they knew it well or they knew it a fair amount.

[5]We used this test because the Shapiro-Wilk Test of Normality showed the groups' accuracies to be non-normally distributed.

if they were seemingly reasonable and consistent.

We found that the teacher's perceived difficulty of the task exhibited low correlation with the feedback score they provided. This suggests that their own performance on a given board did not significantly influence their feedback. It could be interesting to explore if (and how much) people's difficulty perceptions affect their feedback in other machine learning tasks, as well.

We realize that there may have been other reasonable feedback strategies that we did not consider in our simulations and that the specifics of Set do not necessarily generalize to other tasks. However, this limitation does not diminish the broader point motivating our study: many complex, real-world tasks allow for different interpretations of what constitutes a reasonable feedback strategy. We found that when people provide evaluative feedback, the strategies they use to determine the score can vary. Set provided us with a simple, controlled way to examine this variation. Taken together, our study highlights the following phenomenon: different teachers may employ consistent, reasonable feedback strategies that have substantially different effects on how well an ML algorithm performs a task.

This phenomenon has several real-world implications. Most importantly, when a specific learning algorithm is chosen to learn a task from numeric evaluations, the strategy employed by the teacher is critical. It should not be assumed that just because a teacher's feedback is consistent and seemingly coherent, the strategy will result in effective task performance. We found that even if a participant was familiar with ML, AI, or the task, this prior knowledge did not translate into higher task performance than those with less self-reported familiarity. Developers should not assume that people with a better understanding of the task or the technology will necessarily be more effective teachers.

3.8 Summary

In summary, this chapter investigated how people evaluated task-based robot failures in a human-robot teaching setting. We found that participants provided partial credit for different types of failures and varied significantly in the factors that they prioritized when providing their feedback. When these participant-specific strategies were extrapolated and used to train a machine learning algorithm, some strategies resulted in significantly more effective and efficient task performance than others. Together, these results provide insights into how people evaluate robot failures within the functional dimension and underscore the need for robots to accommodate variability in how users interpret and respond to functional failures.

While this chapter focused on how people evaluate robot failures in terms of task performance, many robot failures have consequences that extend beyond functional outcomes. In the next chapter, we turn to the social dimension of robot failure, investigating how failures affect users' trust and responses to robots.

Chapter 4

The Social Dimension: Children's and Adults' Trust After Successive Robot Failures

In the previous chapter, we investigated the effects of robot failures in the functional dimension by examining how people's evaluative feedback varied in response to task-based robot failures in an interactive human-robot teaching setting. In this chapter, we turn our focus to the effects of robot failures in the social dimension by investigating the effects of age on people's trust in a failing robot during a collaborative word-learning game[1]. This context primarily focuses on a component of the social dimension because it is centered on how much people trust the robot's advice after it starts to fail. In the real world, we expect children to learn new words, skills, and ideas from various technologies. When learning from humans, children prefer people who are reliable and trustworthy, yet children also forgive people's occasional mistakes or failures. In this chapter, we explore the following question: Are the dynamics of chil-

[1]Portions of this chapter were originally published as: Teresa Flanagan, **Nicholas C. Georgiou**, Brian Scassellati, Tamar Kushnir. (2024). School-age children are more skeptical of inaccurate robots than adults. In *Cognition*, Volume 249, August 2024, 105814. [75]

dren learning from technologies, which can also fail, similar to learning from humans? We tackle this question by focusing on early childhood, an age at which children are expected to master foundational academic skills. In this project, 168 4–7-year-old children (Study 1) and 168 adults (Study 2) played an online word-guessing game over video with either a human or robot partner. The partner first gave a sequence of correct answers, but then followed this with a sequence of wrong answers, with a reaction following each one. Reactions varied by condition, either expressing an accident, an accident marked with an apology, or an unhelpful intention. We found that older children were less trusting than both younger children and adults and were even more skeptical after failures. Trust decreased most rapidly when failures were intentional, but only children (and especially older children) outright rejected help from intentionally unhelpful partners. As an exception to this general trend, older children maintained their trust for longer when a robot (but not a human) apologized for its mistake. Our work suggests that educational technology design cannot be one size fits all but rather must account for developmental changes in children's learning goals.

4.1 Introduction

Every day, we put our trust in technologies – we ask smart speakers, like Amazon Alexa, what the temperature is, we use apps to learn new languages, we seek help from chatbots, like ChatGPT, to write our code, emails, and even papers. When we are adults, these instances of trust may seem like small additions to what we already know about the world, but what if we were expected to trust technologies as we are forming the foundations of our knowledge? In this project, we address this question by investigating whether 4- to 7-year-old children, and by comparison adults, trust an interactive technology that starts out accurate and helpful but becomes inaccurate

and unhelpful mid-way through a collaborative word learning game.

Educational technologies for young children became popular in the 1990s with TV programs like Baby Einstein and have drastically increased in multiple mediums in the 30 years since. Though parents and educators were initially excited for the possibilities of broadening access to learning opportunities for young children, research warned that relying too heavily on TV and other technologies (including eBooks and some tablet applications) did more harm than good [52]. For example, numerous studies on the "video deficit effect" [9, 21, 251, 262] show that children under 3-years-old do not encode simple events (new words, locations of hidden objects) from video, but can easily learn when the same events are shown to them live. Even though video deficits decline with age, research continues to show that best practice for media use is as a supplement to social interactions with adults, rather than as stand-alone experiences [191, 198, 250, 252]. The message is clear: social learning is best for children, and technology is only effective when it is interactive (i.e., can contingently communicate either verbally or nonverbally), rather than passively used.

Robots are an interesting case for educational technologies because they are readily perceived by children (and even sometimes by adults) to be agentic – not exactly human, but not exactly objects either [25, 36, 49, 71, 77, 82, 110, 111, 112]. Promisingly, new work shows that robots make better teachers for young children than TV and other non-interactive media. For example, toddlers' vocabulary skills improve when taught by a social robot [190], preschoolers learn novel words and facts taught by robots [34, 37, 139, 160], and children even engage with robots in uniquely human forms of social learning, such as imitation (though not as much as they are willing to do so for humans [241]).

Thus, by mimicking the qualities of human teachers, social robots (and, more recently, AI applications without human-like form but with interactive, agentic qualities; see [11, 115]) have the potential to transform education in early childhood. But

for this potential to be realized, we need a better understanding of how to design systems that match the expectations of young social learners, who have evolved learning mechanisms responsive to uniquely human social cues and forms of communication [261].

In this work, we explore whether one such characteristic of human social learning applies to robots: that learning is built on a foundation of trust which can be maintained even after occasional failures. In any social interaction where information is shared – whether it be between speakers and listeners or between teachers and learners – there is a tacit acknowledgement that communicators will be truthful and maximally informative, and thus learners (or listeners) can trust the quality of the information they receive [30, 97, 98, 149]. While young children are inclined to trust information from others [107], they also have an emerging ability to selectively block information from teachers who violate this acknowledgment (e.g., by being repeatedly inaccurate or uncertain [101, 131, 133, 240]).

Equally important for human social learning is the ability to understand the communicative signals of occasional failures without losing trust. From the second year of life, young children distinguish between errors that are intentional and accidental [24, 44, 87]. As such, preschoolers and elementary-aged children will forgive an accidental transgression [8, 49, 179]. By 4-years-old, children will still learn from a teacher who is occasionally wrong, as long as the teacher has also been occasionally right [208, 221]. Preschoolers will also continue to trust a person that admits ignorance about some things but shows confidence in their knowledge of other things [143, 144]. Preschoolers also understand that displays of remorse after a mistake, such as an apology, can be an indication that the transgressor feels guilty and wants to rectify the mistake [238, 239]. Accordingly, 4-year-olds prefer and forgive a person that gives an apology after a transgression [203, 266].

However, children are not forgiving of all failures. For example, 3–6-year-olds will

promote punishment on those that intentionally cause harm [49, 179]. Preschoolers also distrust and negatively evaluate people that have a history of harmful behavior [64] and people that intentionally share false information [61, 217]. Children also do not treat all apologies as the same, but this seems to change with age. Specifically, 5–6-year-olds are less forgiving of a person that repeatedly gives the same apology for the same offense [272] and 7–9-year-olds distinguish between willing and coerced apologies [237]. Together this work suggests that young children expect and prefer their human teachers to be truthful and well intentioned, while also understanding that humans cannot be perfect. Can lessons from this body of work be applied to learning from robots? Recent work has shown that children will selectively trust robot teachers: 3–5-year-olds will learn new words from a robot that has previously accurately labeled objects, but not from a robot that has previously inaccurately labeled objects [37, 160]. But it remains an open question as to whether children's social expectations and preferences for human teachers extend to robot teachers. Specifically, are children inclined to trust a robot, even if they are unaware of the robot's competency, and how would children respond to a robot that communicatively signals that a failure is accidental (or intentional) or that it even feels remorseful for the failure?

Designing learning systems that mimic human social error-signaling cues (such as marking accidental errors or even apologizing for them) could help mitigate two types of problems: First, it could be beneficial so that learners do not immediately stop trusting technology that occasionally breaks, glitches, or otherwise makes mistakes. But also, socially signaling imperfection could mitigate issues that arise when children (or adults) assume that technologies, like internet search engines, voice assistants, or large language models, are more reliable sources of information than humans by virtue of having easy access to more information [56, 89, 218, 273]. Either way, understanding the dynamics of trust in learning from social robots across development

is an important first step.

To date, questions have primarily been concerned about how much adults trust robots that make errors. The gist of this body of work is that adults distinguish social from technical errors and prefer technologies that act more agent-like. For example, when a robot or chat bot makes technical errors (e.g., typos), adults do not view the technology as human-like and are less likely to maintain trust [40, 281]. However, when the robot makes social mistakes (e.g., does not follow the rules, makes incongruent gestures, or is overtly mean), adults prefer the robot and think it has human-like qualities [185, 228, 234, 287]. Some research suggests that adults also prefer when a robot attempts to rectify its mistakes (e.g., by apologizing or offering compensation [127, 152, 284]).

There are several reasons to think that children will not necessarily respond the same way as adults to robot failures. For one thing, young children are overall more willing to treat robots as agents than adults [76, 110, 215]. Young children, therefore, might similarly respond to a robot's failure as they would for a person's failure. Furthermore, children's belief in a technology's agentic capabilities declines with age [77], so we may even see age-related changes in how children respond to a robot's failure. We may also see changes in learning that correspond to changes in learning goals across the critical transition from play-based learning to formal, classroom learning. In our formal education system in the U.S., 6-to 7-years-old (first grade) are expected to master foundational skills necessary for learning to read, write, and understand symbolic and numerical systems [55, 223]. Educational technologies, therefore, could be treated differently for children at this age compared to younger children who use technologies more for entertainment purposes. Taking this work together, it is reasonable to anticipate age differences in trust of technologies, both between children and adults and between children of different ages.

In this project, we explored 4–7-year-old children's (Study 1) and adults' (Study

2) trust in technologies during an educational, collaborative game. In the following studies, children and adults played an online, word-learning game with either a human or robot partner[2]. The game involved 8 trials in which, for each trial, participants had to guess the "correct" label of a novel object. Participants always had zero knowledge of the correct label, as we used random novel objects with randomly selected correct labels from the Novel Object and Unusual Database [105]. Critically, before the participants made their own guess, they first heard their partner say what they think is the correct label. We measured trust, therefore, by whether participants picked the label their partner suggested. Importantly, participants got immediate feedback on their and their partner's choice, either indicating that the chosen label was right or wrong. The feedback after each trial, therefore, would theoretically inform participants whether they should trust the human or robot's suggestion on the next trial.

We were first interested in children and adults' baseline trust in their technological partner. Prior work has explored whether children are willing to trust a robot after they see the robot accurately or inaccurately label familiar objects [37, 160], but it is also important to explore whether children and adults will default to trusting an unfamiliar robot, as they do with other people [107]. Considering this, we intentionally did not give participants any information about their partner's word knowledge. Therefore, participants' responses on the first trial will demonstrate whether children and adults are inclined to trust an unfamiliar robot partner. We can also then see how trust changes as children and adults gain more positive information about their partner through playing the game. So, for trials 1–4, the robot or human suggested the correct label – if the participant picked the label given by the partner, the participant got the question right; if the participant picked one of the other two labels, the participant got the question wrong.

[2]The partner was not co-located with the participant, but rather interacted with the participant through video recordings (the participants were not told that these videos were prerecorded).

Our primary interest was whether children and adults lose trust in their partner once it makes a mistake, and whether this depends on the way the partner responds to the loss. To investigate this, for trials 5–8, the robot or human suggested the wrong label, so that if the participant picked the label suggested, the participant got the question wrong. Children and adults saw the partner consistently react to the inaccuracy in one of three ways: the Mistaken partner expressed self-awareness by responding to the loss with shaking their head and saying, "oops I made a mistake" The Apologetic partner responded the same as the Mistaken partner but added an apology ("I am so sorry") after admitting the mistake. The Uncooperative partner expressed a deceitful intention by responding with pointing their finger and saying, "haha I told you the wrong one."

After all 8 trials, we followed with an interview probing participants' judgments of their partner as well as the other partner that they did not play with. Prior work by Brink et al. [37] found a moderate relationship between children's trust in a robot and believing that the robot has psychological agency (e.g., ability to think, know good from bad). We aim to explore the relationship of psychological agency and trust in light of mistakes. We also include other judgments of agency (e.g., emotional, social, sensory, and competence, similarity to humans) to investigate whether these also relate to children's trust in mistaken technologies.

4.2 Study 1

4.2.1 Methods

The study was approved by a university Institutional Review Board. The deidentified data set, analysis code, and preregistration for the study are available on the project's Open Science Framework (OSF) page at https://osf.io/n4d23/.

Participants

The final sample consisted of 168 4–7-year-old children (M = 5.96, SD = 1.09, 48% females) recruited from two lab databases, one in a small city in the Northeastern United States (64% of children) and the other from a city in the Southeastern United States (36% of children). The preregistration had an initial goal of 156 participants (26 in each condition), but an additional 12 participants were collected to have a better distribution of age and gender in each sample. The initial goal of 156 participants was determined through an a priori power analysis using G*Power (version 3.1.9.6; [70]), which found that 154 participants (25.67 in each condition) would be required to achieve 80% power for detecting a medium effect (0.25) at a significance criterion of $\alpha = 0.05$ for an ANCOVA statistical test (number of groups = 6, numerator degrees of freedom = 2). With our new sample number, we conducted a sensitivity power analysis using GPower and found that 168 participants with a significance criterion of $\alpha = 0.05$ and power = 0.80 would result in a medium effect size (0.24) for the same type of ANCOVA statistical test.

There were 28 children in each condition. Of those that reported, 69% children were White, 17% were Bi- or Multi-racial, 10% were Asian, 3% were Black/African American, 1% was Hispanic/Latino. The majority of children's primary caregivers held a college degree or above (98%) and had a household income of $50,000 or above (91%). Ten additional children participated but were excluded from the final sample due to internet issues (N = 3), the child wanting to stop playing the game (N = 5), or the child being distracted (e.g., looking away from the computer, talking over videos or the experimenter) throughout the majority of the game (N = 2).

Materials

The videos of the robot partner involved a Nao humanoid robot. The Nao robot is 58 cm tall, can speak, and has legs, arms, a torso, and a head with eyes and a mouth.

The videos of the human partner involved a young adult woman with blonde hair in a red shirt. Careful attention was paid to making the robot behaviors and vocalizations similar to those of the human partner in the study. To make the partners seem more agentic, the partner's mannerisms included idle behaviors (e.g., slight arm, hand, or head movements) and other possible behaviors to express emotion (e.g., raising its arm towards its chin in a pensive pose; head facing downwards to express disappointment; a thumbs up to express happiness; and finger pointing to express deceit). The videos of the partners were pre-recorded, but this detail was not mentioned to participants.

There were 8 novel objects used during the word-guessing game, and each object had 3 novel labels (24 novel labels total). The novel objects and labels were obtained through the Novel Object and Unusual Name (NOUN) Database [105]. The entire study, including the word-guessing game and agency questionnaire was developed on Unity and hosted on GitHub Pages as a Unity WebGL build. The web application consists of a graphical user interface (GUI), which includes embedded pictures and videos, that enables the participant to interact with the game and to answer questions.

Procedure

The study was done online via Zoom. Each child sat with their guardian by the computer and the experimenter displayed her screen. When the experimenter displayed her screen, she got confirmation from the child and the child's guardian that they could both see the full screen. Children were then first given a warm-up questionnaire – children were asked how much they like certain foods – to ensure that children could see the screen, to make children feel comfortable in the online testing environment, and to familiarize children with a scale later used in a questionnaire. Children were randomly assigned to one of the six conditions in a 2×3 design (Partner type: Human or Robot; Response type: Mistaken or Apologetic or Uncooperative). The study proceeded in two parts: the word-guessing game, then the questionnaire.

First, children were told that they are going to play a word-guessing game with a partner (Anne or Nao). Children were told that in the game they will see an object, Anne/Nao will tell them that she/it thinks it is called, and then the child can make their own guess. Then participants were introduced to the partner via video: the partner waved, said hello and their name (Anne or Nao), and then said, "let's play the game". This was done to demonstrate that the robot can move and talk on its own, while also establishing that the partner is aware of the game being played. While the video of the partner was still up on the screen, a button that said "next" appeared on the screen and the experimenter repeated the instructions of the game (e.g., "I am going to hit next and you and Nao are going to play the game, where you will see an object, hear what Nao thinks it is called, and you will make your guess.")

There were 8 trials in the word-guessing game. For each trial, the participant saw a novel object, three possible novel labels, and a video of the partner on the screen. The partner always gave the first response by saying what she/it thinks the object is called (e.g., "I think it is a lorp"). Then there was a blue arrow that pointed to the label the partner suggested. The participant was then given a chance to endorse the partners' response or pick a different label (e.g., "What about you? Do you think it is called a blap, a lorp, or a tunk? Anne/Nao thinks it is called a lorp."). The order of the partner's suggested label (left, middle, right) was predetermined by a randomized generator and was fixed across participants (L, M, R, L, M, L, M). For trials 1–4, if the participant picked the same label as the partner, a bell-ring noise played, and the screen displayed a green check mark and a green arrow pointed to the label the partner suggested. The partner responded to the outcome with encouragement (lifted hands, saying "yay great job"). The experimenter told the child that they and the partner guessed the right answer (e.g., "That was correct. You and Anne/Nao guessed the lorp and that was correct."). If the participant picked a different label, a buzzer noise played and the screen displayed a red X mark, a green arrow pointed to the

label that the partner suggested, and a red arrow pointed to the label that the child picked. The partner responded to the outcome with discouragement (shaking head, saying "oh no"). The experimenter told the child that they guessed the wrong answer, but that the partner guessed the right answer (e.g., "That was incorrect. You guessed the blap and that was incorrect. Anne/Nao guessed the lorp and that was correct."). Starting after the participant's response on Trial 5, and continuing through Trial 8, the feedback was reversed. If participants picked the same label as the partner, then a buzzer noise played and the screen displayed a red X mark, a red arrow pointed to the label that the partner suggested, and a green arrow pointed to one of the other two labels. The experimenter told the child that they and the partner guessed the wrong answer (e.g., "That was incorrect. You and Anne/Nao guessed the koba and that was incorrect."). If the participant picked a different label (either one of the two labels that the partner did not suggest), then a bell ringing noise played and the screen displayed a check mark, a red arrow pointed to the label that the partner suggested, and a green arrow pointed to the label that the child picked. The experimenter told the child that they guessed the right answer, but that the partner guessed the wrong answer (e.g., "That was correct. You guessed the bosa and that was correct. Anne/Nao guessed the koba and that was incorrect.").

Critically, the partner's response on Trials 5–8 was to her/its own incorrect answer, rather than to the outcome, and varied by condition. In the Mistaken response condition, the partner shook her/its head and said, "oops I made a mistake." In the Apologetic response condition, the partner responded the same as the Mistaken response condition, but added an apology: the partner shook her/its head and said, "oops I made a mistake. I am so sorry." In the Uncooperative response condition, the partner pointed her/its arm and said, "ha ha I told you the wrong one."

After the word-guessing game, participants were told that they were done with the game and were now going to answer a few questions about what they think

about their 'partner' and the game they played (see full questionnaire and coding in Appendix A). The questionnaire consisted of questions regarding the partner's mental capabilities (can think for herself/itself), emotions (has feelings like happy and sad), sensations (can see and hear the things around it), friendliness (can be your friend), epistemic knowledge (knows the answers to a lot of questions) and moral knowledge (knows the difference between good and bad). For each question, the screen displayed a picture of the partner, the question, and three possible answers (not at all, a little bit, or a lot). Question order was randomized. Participants were also asked about the partner's ontological status ("Is Anne/Nao more like a person or a computer?"). The screen displayed a picture of the partner, the question, a picture of a human lady at one end of the screen and a picture of a computer at the other end, with tick marks in between as possible answers (person a lot, person a little bit, in the middle, computer a little bit, computer a lot).

After participants answered the questions for the partner they played with, they were then shown the other partner (participants in the Human Partner type were shown the robot partner, participants in the Robot Partner type were shown the human partner). A short video of the partner played in which the partner waved and said hello and their name. Then participants were asked the same questions about the new partner.

At the end of the study, we investigated participants' reasoning about the partner's inaccuracy. Participants were asked, "Remember in the word-guessing game with Anne/Nao. In the beginning Anne/Nao told you the right answers but then Anne/Nao started telling you the wrong answers. Why do you think Anne/Nao told you the wrong answers?" Participants' responses were recorded.

Coding

In the word-guessing game, we measured participants' endorsement of the partner at each trial. Participants received a score of 1 if they picked the same label as the partner and a score of 0 if they picked a different label.

For the agency questionnaire, participants' responses to the mental, emotional, sensory, friendliness, epistemic knowledge, and moral knowledge questions were measured in a 3-point scale coded as 0 (not at all), 1 (a little bit), and 2 (a lot). Participants' response to the ontological status question was measured in a 5-point scale coded as 0 (computer a lot), 1 (computer a little bit), 2 (in the middle), 3 (person a little bit), 4 (person a lot).

Participants' responses to the open-ended question were categorized into any of the following categories, not mutually exclusive. Intention: reference to the partner's actions as either not intentional (e.g., "it was an accident"), intentionally harmful (e.g., "he tried to trick me"), or intentionally helpful (e.g., "she wants us to think for ourselves"). Mechanical Property: reference to the partner's mechanical properties (e.g., "he's a robot", "it's broken"). Competence: reference to the partner's competence (e.g., "she doesn't know the answer"). Game Difficulty: reference to the game being difficult (e.g., "the game got trickier"). Self blame: reference to the child's role (e.g., "I hit the wrong bottom"). Adult learning: reference to it being an opportunity for the adult to play and learn (e.g., "so I could figure it out myself"). Wrong answer: restating that the partner gave the wrong answer (e.g., "because it was wrong"). Physiology: reference to the partner's physical state (e.g., "she was tired"). Other: any unrelated answer (e.g., "bad") or the child saying they don't know. Two condition blind coders independently categorized all explanations for each response (agreement κs ≥ 0.264, ps < 0.001) and any discrepancies were resolved through discussion.

4.2.2 Results

We were interested in whether children endorsed the partner's suggested word choices during each phase of the game and whether this differed by the type of partner (human or robot), the partner's response, or both. In the preregistration, we initially planned to compare children's responses for each trial, but we decided it would be more beneficial to see how children's responses change by trial as well. The following analyses, therefore, are exploratory, but we report the preregistered analyses in Appendix A and we summarize any notable findings in the main manuscript.

The Accuracy Phase included all endorsements prior to the partner's first inaccuracy, so does not include the condition-dependent responses (Trials 1–5). For the Accuracy Phase, therefore, we ran a mixed-effects model with Endorsement as the dependent variable, with partner type (robot or human), Trial (Trials 1–5), and age (in years) as the independent variables, and participant ID as a random intercept. The Inaccuracy Phase included endorsements after the first inaccuracy and through all remaining trials, and thus included all the trials on which children had received condition-dependent feedback of the partner's response to her/its inaccuracy (Trials 6–8). For the Inaccuracy Phase, therefore, we ran a similar mixed-effects model as the Accuracy Phase but included Response type (Mistaken or Apologetic or Uncooperative) as an independent variable. For both of the models, we only included interactions in the model if they were previously found to be significant.

Figure 4.1 shows the rates of endorsement of the partner's word choice across the Accuracy phase from Trials 1 to 5. On Trial 1 – after just a brief introduction and prior to corrective feedback – the majority of children followed both the human and robot partner's suggested word choice (Human: 60.7%, $SD = 0.49$; Robot: 61.9%, $SD = 0.49$; binomial $ps < 0.0001$). The final model for the Accuracy Phase included partner type, Trial, and age as factors – interactions were removed from the final model as they were initially found to be insignificant, $ps \geq 0.080$. We did not find

a significant main effect of partner type, children trusted the human and robot at similar rates, $\chi^2(1) = 0.38$, $p = .541$. We found a main effect of age, $\chi^2(1) = 7.03$, $p = .008$, such that older children were less likely to endorse the partner's word choices than younger children, $OR = 0.74$, 95% CI (0.59, 0.93). Finally, we found a main effect of Trial, $\chi^2(4) = 55.14$, $p < .001$, such that it took two instances of accuracy for children's trust in the partner to significantly increase above the initial level, whether they were human or robot: Trial 3 versus Trial 2, $OR = 2.96$, $p = .001$, 95% CI (1.64, 5.34). Trust did not differ between the other sequential Trials, $ps \geq 0.237$. Trust remained high throughout the first half of the game (67.9% - 100%), binomial $ps < 0.0001$.

Fig. 4.2 shows the rates of endorsement of the partner's word choice at Trial 5 and during the Inaccuracy phase from Trials 6 to 8. When the partner starts giving the wrong answer, we found that children's trust varied by the number of inaccuracies, the partner and response type, as well as children's age. In general, children were still willing to trust the partner after one instance of inaccuracy (57.1% -67.9%), Trial 6 binomial $ps < 0.004$. The final model for the Inaccuracy Phase included partner type, Response Condition, Trial, and age as factors as well as two three-way interactions between partner type, Response Condition, and age and between Response Condition, Trial, and age – all other possible two- and three-way interactions were removed from the final model as they were initially found to be insignificant, $ps \geq 0.051$.

Running our model, we found a main effect of age, such that younger children trusted the partner's word choice more than older children, $\chi^2(1) = 9.79$, $p = .002$. We also found a main effect of Trial, $\chi^2(2) = 43.44$, $p < .0001$, such that children's endorsements significantly declined from Trial 6 to 7, $OR = 0.17$, $p < .0001$, 95% CI (0.09, 0.36), but not from Trial 7 to Trial 8, $OR = 0.44$, $p = .051$, 95% CI (0.19, 1.00). The percentage of children trusting the partner at Trial 7 (25% - 42.9%) and Trial 8 (17.9% - 25%) were at chance, binomial $ps > 0.069$, or less than chance (for the

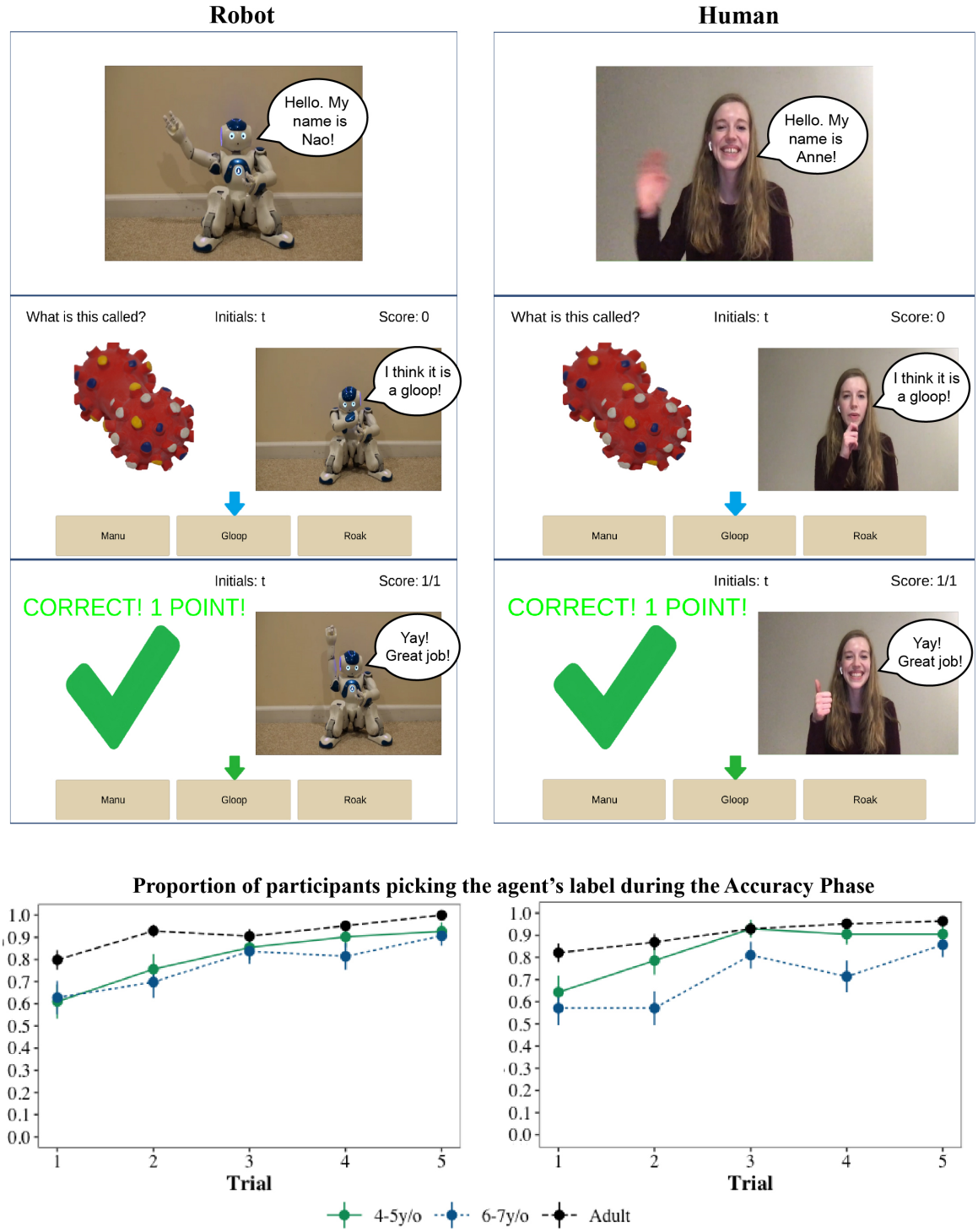

Figure 4.1: Example of the word-guessing game during the Accuracy Phase and participant's responses at each Trial (1–5), split by Partner type (Robot and Human) and age (4–5-year-olds, 6–7-year-olds, adults). Bars represent standard error. For visualization, children's age is grouped categorically, but analyses involve age as a continuous variable.

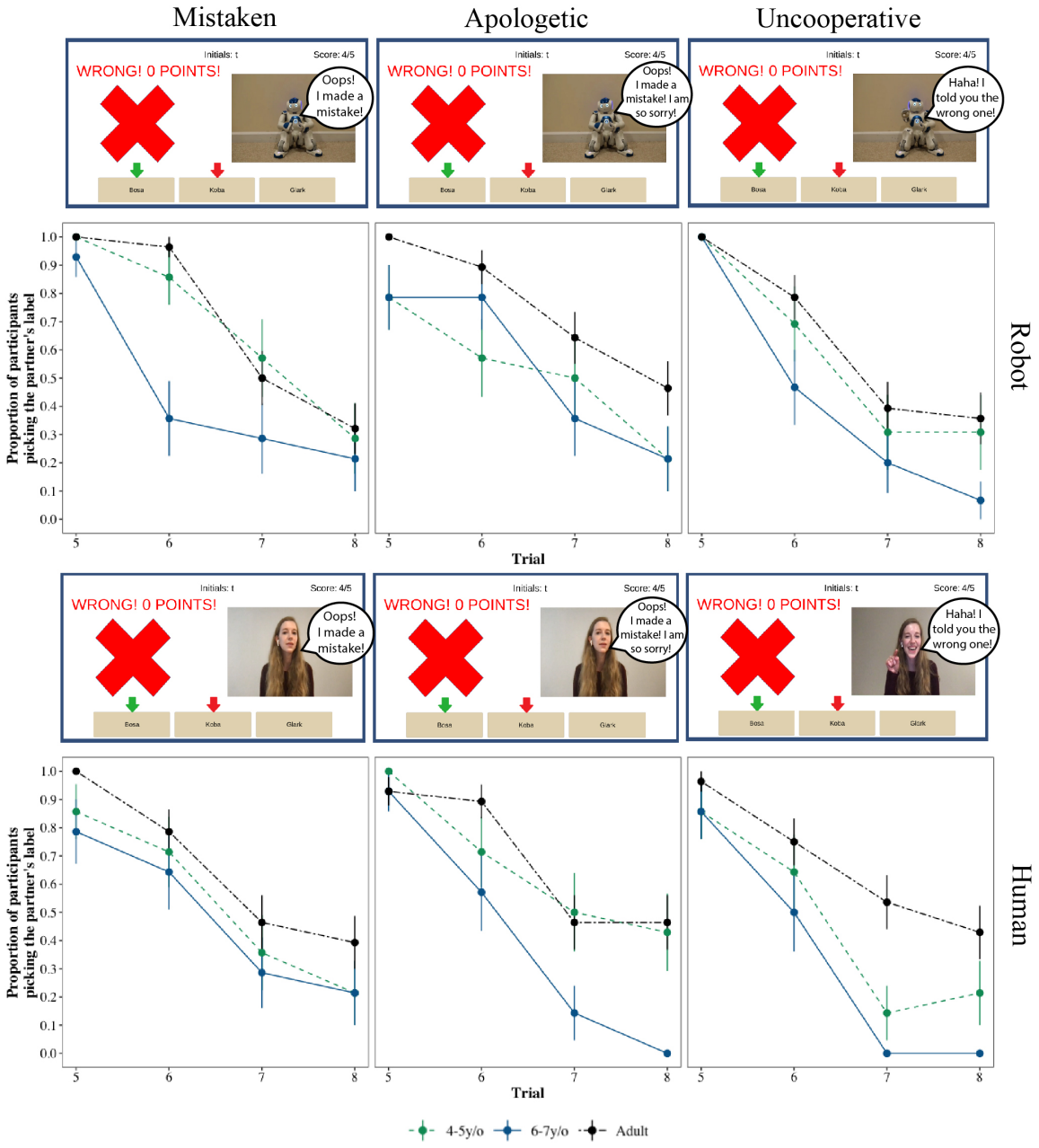

Figure 4.2: Example of the word-guessing game during the Inaccuracy Phase and participant's responses at each Trial (5–8), split by Partner type (Robot and Human), Response type (Mistaken, Apologetic, and Uncooperative) and age (4–5-year-olds, 6–7-year-olds, adults). Bars represent standard error. For visualization, children's age is grouped categorically, but analyses involve age as a continuous variable.

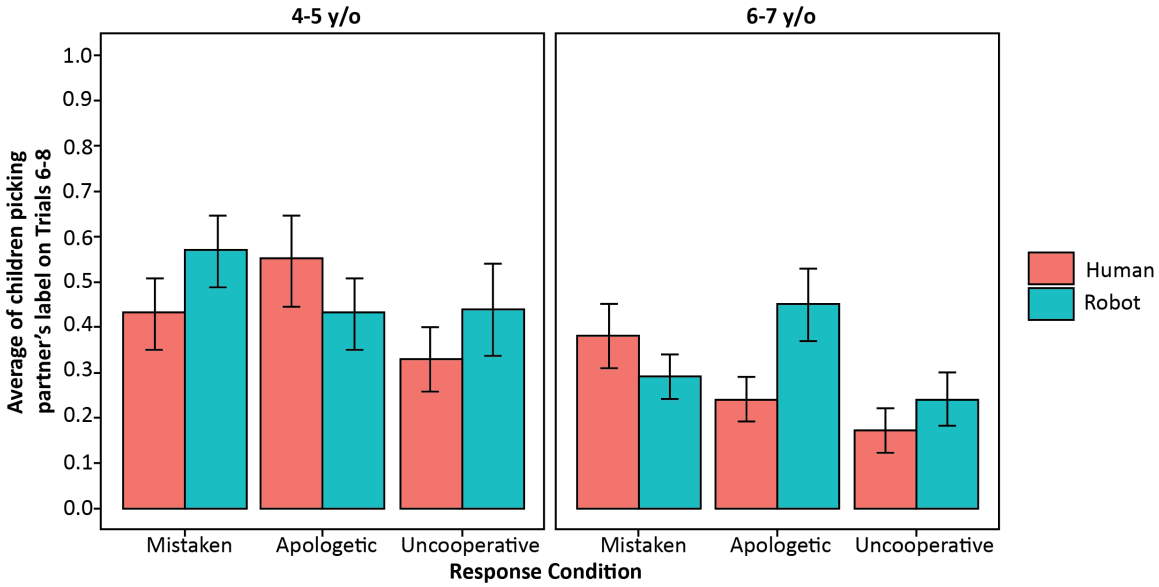

Figure 4.3: Mean proportion of children picking the partner's label on Trials 6–8, split by age (4-5-year-olds and 6-7-year-olds), Response type (Mistaken, Apologetic, and Uncooperative) and Partner type (Robot and Human). Bars represent standard error.

Human Uncooperative at Trial 7 (7.1%, SD = 0.26) and Trial 8 (10.7%, SD = 0.32), binomial $ps < 0.014$. We did not find a main effect of partner type, $\chi^2(1) = 1.80$, $p = .180$, or a main effect of Response Condition, $\chi^2(2) = 4.65$, $p = .098$. We also did not find any significant two-way interactions within the three-way interactions included in our model, $ps \geq 0.337$. However, children's trust in partners who admit mistakes diverged with age, as indicated by two three-way interactions found in the model.

First, we found a significant interaction between partner type, Response condition, and age, $\chi^2(2) = 6.53$, $p = .038$ (see Fig. 4.3). We looked at this interaction in two ways: how children's trust in each condition changed with age and how trust differed between conditions for each age group. For the former, we ran an additional model with the three-way interaction with age as a continuous variable; for the latter, we ran a similar model but with age as a categorical variable (4–5-year olds and 6–7-year-olds). All models included Trial and participant ID as random variables. Results of the two models can be summarized as follows: We found that older children were less trusting of a robot that makes a mistake than younger children, $OR = 0.46$, $p = .026$,

95% CI (0.27, 0.78). Older children were also less trusting of a human that apologizes than younger children, $OR = 0.50$, $p = .046$, 95% CI (0.30, 0.83). Furthermore, older children were more trusting of an apologetic robot than an apologetic human, $OR = 3.34$, $p = .034$, 95% CI (1.09, 10.20). Together, these findings suggest that by school age, children are more skeptical of unintentional mistakes made by a robot, but apologies mitigate this effect.

Second, we found a significant interaction between Response condition, Trial, and age, $\chi^2(2) = 15.62$, $p = .004$. Running this additional model, we investigated the differences in children's endorsement by age at each Trial for each Response type, averaged over partner types. At Trial 6, older children were less likely to endorse a Mistaken partner's word choices than younger children, $OR = 0.33$, $p = .029$, 95% CI (0.15, 0.69). We did not find a significant difference in age at Trial 6 for the other Response conditions or at the other trials for any Response condition, $ps \geq 0.073$. This demonstrates that the above finding – that older children are less forgiving than younger children of mistakes – takes effect immediately, after just one instance in accuracy without apology. Furthermore, supplemental analyses exploring differences at each Trial (see Appendix A) suggest that children across the age range distinguished between intentional and unintentional inaccuracies, preferring to trust partners whose errors were clearly unintentional. Specifically, children in the Uncooperative conditions were less likely to trust the partner at Trial 7 than children in the Mistaken, $OR = 0.29$, $p = .028$, or Apologetic Conditions, $OR = 0.30$, $p = .033$ (see Appendix A for full model results). This aligns with children's open-ended judgments, such that children who played the game with the Uncooperative partner were more likely to say that the partner had harmful intentions than children in the Mistaken or Apologetic conditions (see Appendix A).

Results for children's responses to the agency questions are reported in Appendix A, but we give a brief summary of the findings here. In general, children thought

that the human partner was more like a person than the robot. In line with prior work [77], younger children said that the robot was more like a person than older children. Younger children also said that the robot could have feelings and could think for itself more than older children. Unlike prior work [37], we did not find an influence of children's belief that the robot had psychological agency (e.g., could think, know the difference between good and bad) on children's overall trust in the robot. However, we did find that children's belief that the robot knows the answers to a lot of questions increased their overall trust in the robot.

4.2.3 Discussion

Children in the modern world are surrounded by technologies, particularly in their education. It is critical, therefore, to investigate whether children are trusting of technologies and whether trust can be maintained when the technologies are inaccurate. In this study, we found that 4–7-year-old children are inclined to trust their partner, human or robot, when learning new words, even though children had no prior information on their partner's competency ([107] for a perspective on young children's default bias to trust). Naturally, children's trust increased as they saw more instances of the partner being accurate [101, 132, 133, 240]. When the partner was inaccurate, however, children were sensitive to whether the error was intentional or accidental. Notably, both when the partner was accurate or intentionally inaccurate, children's trust did not differ between the human or robot partner, suggesting that children expect robots to be helpful and accurate teachers just like humans. However, when robots made accidental errors, we found that children's trust differed with age. The age-related differences highlight the different ways educational technologies are used in early childhood. In general, we found that older children were less trusting of their partner throughout the entire game, even when the partner was accurate. Considering that children were not given any prior information about their partner before

the game, it may be that younger children are inclined to trust unfamiliar partners [107] while older children may require more information about their partner before deciding to trust. Similarly, the role of anchoring [18], the tendency to rely on the first piece of information encountered when making a decision, may play a role here, with the robot's initial correct advice influencing younger and older children differently. It would be beneficial for work to explore this possibility more directly, as this is relevant to the way children are engaging with educational technologies in their everyday lives: children are often introduced to technologies with little information about the technologies' competence or source of knowledge.

The change in age could also be related to differences in how children engage with educational programs. In preschool, children may use robots and other technology mainly for fun, not for the explicit goal of learning [231]. As children transition to formal education, they are expected to master foundational skills, such as reading words. As such, children at this age are being evaluated on their academic performance and are increasingly aware of, and sensitive to, their own and others' academic competencies [26, 227, 247]. Older children in our study, therefore, likely took the game more seriously as part of their learning and academic performing. This expectation could have made children more restrictive regarding who they trusted.

Furthermore, and most notably, older children were sensitive to the robot's response to the unintentional inaccuracy while younger children seemed to judge the robot's and the human's response similarly. Specifically, older children were more forgiving of a robot that apologizes for its mistake than a human who gives the same apology. An apology from a person can indicate two things: a genuine feeling of remorse or following a social norm. We expect people to apologize after a transgression, so older children may have viewed the human's apology as conforming to norm rather than as sincere [237, 272]. However, we may not expect a machine-like robot to apologize or even be aware of the social norms of apologies. For this reason,

older children may have viewed an apology from the robot as a sincere expression of remorse, and perhaps even a promise to not make the same mistake again.

The developmental changes in children's trust of a robot partner call to question how adults would respond. Adults are also using technologies in their daily lives, but are less likely to view technologies as agents than children [76, 110, 215]. Furthermore, adults likely use technologies to build upon their existing knowledge (e.g., asking ChatGPT to write an email), rather than using technologies to teach the foundations of their knowledge (e. g., learning new words). For each of these reasons, we may see adults trusting technologies differently than the children in our study. We may also see adults responding differently to the robot's accidental, remorseful, or intentional errors, as following prior work [127, 152, 284]. In the following study, therefore, we investigated adults' trust of a human or robot partner in the same word-learning game as Study 1.

4.3 Study 2

4.3.1 Methods

The study was approved by a university Institutional Review Board. The deidentified data set, analysis code, and preregistration for the study are available on the project's OSF page at https://osf.io/n4d23/.

Participants

The final sample consisted of 168 adults (M_{age} = 33.8, SD_{age} = 11.79, 52% females) recruited from Prolific. There were 28 adults in each condition. Of those that reported, 65% adults were White, 5% were Bi- or Multi-racial, 10% were Asian, 8% were Black/African American, 10% were Hispanic/Latino, 1% was Middle Eastern, and 1% was American Indian. Half of the adults held a college degree or above (57%)

and had a household income of $50,000 or above (52%). The same number was determined to match Study 1, but we also conducted a sensitivity power analysis using G*Power and found that 168 participants with a significance criterion of $\alpha = 0.05$ and power = 0.80 would result in a medium effect size (0.24) for an ANCOVA statistical test (number of groups = 6, numerator degrees of freedom = 2).

Materials

The materials used for Study 2 were identical to Study 1.

Procedure

The study was done online in which approved Prolific participants were given a link to the word-guessing game and completed the study unmoderated. Participants were randomly assigned to one of the six conditions in a 2 × 3 design (Partner type: Human or Robot; Response type: Mistaken or Apologetic or Uncooperative). The study proceeded in two parts identical to Study 1: the word-guessing game, then the questionnaire. Since adults' participation was unmoderated, additional items and controls were added to the study to ensure that adults were following the study correctly and paying attention.

Specifically, adults had to make the game at full screen before they could participate, video and audio checks were given in the beginning of the study (e.g., adults had to describe video and audio clips), instructions and feedback (matching the exact language the experimenter gave in Study 1) were written at the top of the screen, attention check questions were included throughout (e.g., "Click not at all for this question"), and videos would not play if adults were active on a different window or application on their computer.

Coding

Coding for the word-guessing game and questionnaire in Study 2 was identical to Study 1. For the open-ended question, we removed the category referencing the physiology because no adults gave this explanation. We also included a category for references to intentionally neutral (e.g., "he wanted me to stop trusting him") and study design (e.g., "to test if we would still trust the robot"). Two condition blind coders independently categorized all explanations for each response (agreement $\kappa s \geq 0.334$, $ps < 0.001$) and any discrepancies were resolved through discussion.

4.3.2 Results

We were interested in whether adults endorsed the partner's suggested word choices during each phase of the game and whether this differed by the type of partner (human or robot), the partner's response, or both. In the preregistration, we initially planned to compare adults' responses for each trial, but we decided it would be more beneficial to see how adults' responses change by trial as well. The following analyses, therefore, are exploratory, but we report the preregistered analyses in Appendix A, and we summarize any notable findings in the main manuscript. The following analyses in the main manuscript were the same as Study 1.

Fig. 4.1 shows the rates of adult's endorsement of the partner's word choice across the Accuracy phase from Trials 1 to 5. Similar to children, initial rates of endorsements were above chance for both human (82.1%, $SD = 0.39$) and robot (79.8%, $SD = 0.40$), binomial $ps < 0.0001$. The final model for the Accuracy Phase included partner type and Trial as factors – the two-way interaction was removed from the final model as it was initially found to be insignificant, $p = .660$. We did not find a main effect of Partner type, $\chi^2(1) = 0.03$, $p = .868$, demonstrating that adults, like children, trusted both accurate partners at high rates. We found a main effect of Trial, $\chi^2(4) = 31.63$, $p < .001$, such that there was a significant increase in endorsing

the partners' label after one instance of accuracy: Trial 2 versus Trial 1, $OR = 2.50$, $p = .047$, 95% CI (1.23, 5.10). Trust did not differ between other sequential trials, $ps \geq 0.481$ and remained high throughout the first half of the game (86.9% - 100%), binomial $ps < 0.0001$.

Fig. 4.2 shows the rates of adult's endorsement of the partner's word choice at Trial 5 and across the Inaccuracy phase from Trial 6 to 8. Adults were even willing to maintain trust after one instance of inaccuracy: adults endorsed the partner's label at high rates at Trial 6 (75% - 96.4%), binomial $ps < 0.0001$. The final model for the Inaccuracy Phase included partner type, Response Condition, and Trial as factors – all two and three-way interactions were removed from the final model as they were initially found to be insignificant, $ps \geq 0.335$. We found a main effect of Trial, $\chi^2(2) = 60.24$, $p < .001$, such that adults' endorsements significantly declined from Trial 6 to 7, $OR = 0.12$, $p < .001$, 95% CI (0.05, 0.26), but not from Trial 7 to Trial 8, $OR = 0.60$, $p = .142$, 95% CI (0.50, 1.11). However, looking at our chance comparisons, we never saw a majority of adults mistrust the partner: some adults were still trusting the partner at Trial 7 (39.3% - 50%) and Trial 8 (32.1% - 46.4%), binomial $ps > 0.069$, and a majority of adults were trusting the Human Uncooperative (53.6%, $SD = 0.51$) and Robot Apologetic (64.3%, $SD = 0.49$) conditions at Trial 7, binomial ps < 0.026. We did not find a main effect of partner type, $\chi^2(1) = 0.12$, $p = .730$, or Response Condition, $\chi^2(2) = 2.88$, $p = .238$. Unlike Study 1, even when the partner was intentionally inaccurate, adults trust maintained similarly to the unintentional partners at each trial in the Inaccuracy Phase (see Appendix A). This is interesting considering that adults viewed the uncooperative partner as having malicious intent (see Appendix A).

In an exploratory analysis, we were interested in how adults compared to children. To do this, we created a score of change in trust: first, we created a proportion of trust for the Accuracy Phase (Trials 1–5; e.g., $4/5 = 0.80$) and a proportion of trust

for the Inaccuracy Phase (Trials 6–8; e.g., 1/3 = 0.33). Next, we subtracted the proportion of trust in the Inaccuracy Phase from the Accuracy Phase (e.g., 0.33–0.80 = -0.47). A higher negative score indicates a greater loss of trust. We ran a General Linear Model with the change of trust score as the dependent variable and age group (4–5-year-olds, 6–7-year-olds, and adults), Response Type, and Partner Type as the independent variables. We found a main effect of age group, $F(2, 330) = 3.85$, $p = .022$, $\eta_p^2 = 0.02$.

Specifically, adults maintained more trust in the partner (M = - 0.33, SD = 0.33) than 6–7-year-olds (M = -0.45, SD = 0.31), $t(330) = 2.77$, $p = .016$, $d = 0.31$, 95% CI (0.09, 0.52). Preschool-aged children did not differ between either 6–7-year-olds or adults, $ps \geq 0.220$. This finding further highlights the nuance of educational technologies for early elementary aged children: an age at which they are presumably using educational technologies more than adults, and yet are less trusting of such technologies than adults.

Results for adults' responses to the agency questions are reported in Appendix A, but we give a brief summary of the findings here. Adults thought the human partner had more agentic capabilities than the robot partner and thought the human partner was more humanlike than the robot. Looking at adults' judgments of the robot only, we found that adults' belief that the robot had psychological agency (e.g., can think, know good from bad) and that the robot was more human-like decreased adults' overall trust in the robot. Furthermore, when comparing adults' responses to children's, we found that adult viewed the robot partner as less human-like than children. This suggests that even though adults did not view the robot as an agentic being, they still maintained trust in it as they do for humans.

4.3.3 Discussion

In this study, we investigated whether adults would trust technologies in a word-learning game and whether trust can be maintained once the technologies provide inaccurate information. In general, we found that adults were quick to trust their partner, and this is maintained even when the partner shows that it can be inaccurate. Furthermore, adults' trust did not depend on the partner type or whether the partner's inaccuracy was intentional or accidental. Finally, adults trusted the robot despite reporting that they did not view it has having as much agency as children.

These findings are interesting, considering the differences we found in Study 1 and even prior work has found that adults are sensitive to the partner type and response in their trust in technologies [127, 152, 284]. In these prior studies, however, adults had to trust technologies in more adult-like settings (e.g., investments, driving, service), leading to questions of comparability with our work. It is worth noting, however, that recent work with adults and children participating in selective trust experiments has found that adults maintain trust in human informants longer than children [222].

Furthermore, in serious scenarios involving robots as informants, such as when adults have to follow a robot in an emergency evacuation, studies show that trust is maintained despite errors [218]. Taking this research along with our own findings raise serious questions as to under what circumstances will adults overtrust technologies to a fault.

4.4 General Discussion

Children and adults are getting a vast amount of their information from technologies. Particularly in education, we are seeing more and more agentic technologies that are designed to teach and collaborate with young children [34, 102, 190]. The engagement with technologies, from both children and adults, has also increased exponentially in

light of the recent COVID-19 pandemic. Therefore, it is critical to investigate how young children and adults are engaging with these technological agents in a learning environment. In this project, we explored whether 4–7-year-old children and adults would trust either a human or robot partner in a word-guessing game. Critically, halfway through the game, the partner started giving the wrong answers, either accidentally, remorsefully, or intentionally, and we measured how this inaccuracy changed children's and adults' trust. In general, we found that children and adults were quick to trust a technological agent and maintained trust in the agent after one instance of inaccuracy. However, children lost trust in a partner that was intentionally inaccurate, while adults did not seem to distinguish between intentional and unintentional errors. Notably, school-aged children (6–7-year-olds) were overall less trusting than adults and, in comparison to preschoolers (4–5-year-olds), were more sensitive to the robot's remorseful response to the unintentional inaccuracy. This finding in particular highlights the different goals and expectations adults and children have when engaging with technologies.

The basis of trust is formed through our goals and experiences with others. We establish collaborative partnership with others when they share the same goals or intentions as ourselves [97, 243, 261]. We also incorporate our experiences with people when deciding who or when to trust in various contexts [101, 240]. Notably, these goals and experiences can change across development. In our study, we uncovered a U-shaped pattern in the effects of age on trust in technological agents, such that preschoolers and adults were more trusting of an inaccurate partner than 6–7-year-olds. We take these findings to highlight the influence of adults' and young children's different goals and experiences on their trust in technological agents, as we discuss below.

Commonly, young children's selective trust is viewed as "adult-like", but we found that adults were not selective in their trust in our study. Instead, adults maintained

trust in their partner, even when the partner was intentionally inaccurate. Recent work by [222] similarly found that adults were slower to lose trust in an inaccurate informant compared to 4–7-year-old children. They speculated that adults' gradual distrust compared to children could be because adults have more experiences with others – people are rarely ever repeatedly correct or repeatedly incorrect but are instead inconsistent in their accuracy. This could also be the case for adults' rich experiences with technologies, such that adults in our study did not think that a few wrong answers meant that their technological partner would now only be inaccurate. This work, therefore, highlights the importance of taking a developmental approach in investigating our engagement with technologies, and broadly opens the question as to what it means for our trust to be "adult-like".

The similarities and differences between preschoolers' and elementary-aged children's trust in their partner suggest that young children's experiences and goals with technologies differ between ages. Specifically, preschoolers are likely using technologies more as a source of entertainment than information [90, 231], so their goal may have been more focused on playing, rather than learning. This is likely why the uncooperative response type was the only condition in which there were no changes in age: the uncooperative partner was essentially refusing to play with the child as well as refusing to help the child learn.

Finally, early elementary aged children are increasingly aware of the informative nature of technologies [89, 90] in addition to being more selective in their trust as they get older [107, 171, 184]. Taken together, this is likely why we found that 6–7-year-olds were more sensitive to who the partner was and what the partner said when deciding if they should maintain trust in the partner. For example, an apology from a partner that seems like a machine is an unexpected social cue that older children seem to prefer. Using educational technologies, therefore, is not one size fits all. Instead, we need to be mindful of who these technologies are designed for in order for these

technologies to be the most beneficial and engaging for children.

The difference in children's and adults' goals and experiences with educational technologies is also evident in how children and adults viewed their robot partner. For example, children's trust in the robot was tied to their belief that the robot was a knowledgeable agent. This finding calls to question if other features of the robot are important to children's trust, and if this changes with age with respect to their goals (e.g., more emphasis on entertainment capacities for younger children, but more emphasis on teaching capacities for older children). In contrast, adults' trust in the robot was tied to their belief that the robot was a machine-like object. These findings suggest that adults may prefer educational technologies to be like sources of online information that they are already familiar with (e.g., the internet; see [273]), while children may require their technologies to be more like human informants (e.g., demonstrating their knowledge, apologizing for their mistakes).

Our project presents new avenues for research on engagement with educational technologies. For example, there are various types of educational technologies (e.g., eBooks, smart speakers, chatbots) that have different features than the humanoid robot used in our studies. Considering that children and adults are engaging with these different forms of educational technologies, it would be fruitful to explore people's responses to various types of mistaken technologies. Furthermore, the game in our studies only involved word learning, but prior work has demonstrated that children trust technologies differently for different domains (e.g., personal knowledge, moral knowledge, [56, 89]). It is likely, therefore, that children and adults may respond to a robot's mistake differently in other domains.

The word-guessing game in our studies was designed to mirror the type of educational games children are playing with already. However, in our study, children and adults only played the game once for a short amount of time (approximately 10 min), while, in the real world, people are playing with educational technologies

over extended periods of time and over multiple days. The more experiences children and adults get in these technological spaces likely influences their trust, as there are more instances of the partner being accurate and inaccurate for various reasons. It is unclear, therefore, how children and adults would respond to a mistaken robot if they had more experience with the robot, via a familiarization phase used in prior work (e.g., [131]) or a longitudinal design.

The word guessing game was also entirely online, to best reflect the types of learning activities children were engaging with during the COVID-19 pandemic. However, the online format does present some limitations. For example, it is unclear if children and adults thought the partner was interacting with them in real-time or in a predetermined way, which might have influenced how they viewed the partner's response to inaccuracy. Furthermore, while we took steps to ensure that adults were paying attention and following the game in Study 2 without an experimenter present, the unmoderated method does leave open questions regarding adults' level of engagement compared to children in Study 1. It also remains an open question if children's and adults' responses in an online study would differ in an in-person study environment, as the current literature has found that children's responses are consistent across study environments for some tasks [229], but not others [150].

Finally, it is unclear whether our findings are generalizable to the greater global population. The majority of children and adults who participated in the study were White and were in high-income, educated households. Furthermore, all of the participants in this study had access to computers in their home. The children in the study were also comfortable using Zoom. Since the children and adults were already familiar with using technologies, they may have been quicker to trust educational technologies than people who have had little or no experience with technologies. We cannot assume, therefore, that engagement with technologies will be the same for everyone. Instead, research should explore how different individual factors (e.g., race,

education, technology experience) play a role in our trust of technologies.

We will continue to get our information about the world from technologies. For children, this will particularly be part of their education: whether technologies are used as an additional learning activity, as part of a school's curriculum, or even as the primary teacher. In order to have these technologies benefit children's education and development, we must investigate all the scenarios in which children could engage with technologies, even when the technologies make mistakes. Our results in particular uncover the ways in which young children are engaging with educational technologies in the real world. For example, we found that children of all ages are aware of an uncooperative agent's harmful intentions and, thus, do not maintain trust in uncooperative agents. The technologies in children's daily lives, however, do not typically have harmful intentions. Instead, these mistakes are accidental, but how children respond to these accidents depend on their age. Specifically, we found that preschoolers, compared to older children, are more trusting of technological agents, even when the agents are unintentionally inaccurate. For children in formal education, however, these educational games are taken more seriously, and so older children may be incorporating their different expectations of agents (either human or robot) and mistakes (whether an apology is given or not) when deciding whether to maintain trust in an inaccurate technology. Together, this research presents important implications for technology design and education in young children's development.

4.5 Summary

In summary, this chapter investigated the effects of age on people's trust in a robot that committed successive failures. In a pair of 2x3 between-subjects studies, school-aged children between the ages of 4 and 7 years old (N =168) and adults (N =168), played a collaborative word-guessing game with either a human or robot partner,

that gave wrong advice on certain objects. We found that older children, between the ages of 6 and 7, were less trusting of their partner after it failed than adults and than younger children, between the ages of 4 and 5. Additionally, older children maintained their trust in a robot that apologized for a longer time period than they did in a human that apologized, which was a trend that we did not find with younger children. Together, these findings provide insights into how people's age and their partner's responses to failure can significantly affect peoples' trust throughout an interaction. In turn, extra consideration must be taken when designing robots that interact with children, as they may trust failing robots differently, based on their age.

While this chapter primarily focused on how robot failures affect people's trust, there are robot failures that have consequences far beyond social outcomes. In the next chapter, we turn to the moral dimension of robot failure, investigating how a physical transgression causing harm affects users' moral judgments towards the robot.

Chapter 5

The Moral Dimension: Humans' Moral Judgments of A Robot Causing Physical Harm

In the previous chapter, we broadened our understanding of the effects of robot failures in the social dimension, by investigating how peoples' trust in a robot is influenced by age. In this chapter, we turn our focus to the moral dimension. We examine people's moral judgments of a robot after witnessing it commit a physical transgression towards a human.[1] This context is used for investigating the moral dimension because it is centered on how people evaluate the robot's morality after it commits a moral transgression. We performed an online, between-subjects study in which we manipulated the type of transgressor (*human* or *robot*) and type of backstory depicting the transgressor's mental capabilities (*default*, *physio-emotional*, *socio-emotional*, or *cognitive*). Participants (N=720) were first introduced to the transgressor and its backstory, and then viewed a video of a real-life robot or human pushing down

[1]Portions of this chapter were originally published as: **Nicholas C. Georgiou***, Teresa Flanagan*, Brian Scassellati, Tamar Kushnir. (2025). Perceived Morality of Robot and Human Transgressors Varies by Perceived Ability to Feel. In *20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Melbourne, Australia, 2025, pp. 919-928.

a human. Although participants attributed similarly high intent to both the robot and the human, the human was generally perceived to have higher morality than the robot. However, the backstory that was told about the transgressors' capabilities affected their perceived morality. We found that robots with emotional backstories (i.e., physio-emotional or socio-emotional) had higher perceived moral knowledge, emotional knowledge, and desire than other robots. We also found that humans with cognitive backstories were perceived with less emotional and moral knowledge than other humans. Our findings have consequences for robot ethics and robot design for HRI.

5.1 Introduction

Artificial intelligence (AI) and robotics are becoming more involved in our lives than ever. As a result, it is inevitable that people will experience these technologies doing something that they deem to be unacceptable or wrong. Whether these violations are in the form of hardware or software failures that require a system reboot, inaccurate responses to a question that lead to the spread of misinformation, or transgressions that cause physical or emotional harm, it is imperative to investigate how these undesired behaviors affect humans' perceptions of the technology.

Robot mistakes and failures have been explored in HRI research, but mostly through the lens of minimizing or recovering lost trust in the robot [104, 286, 75, 170, 51, 285]. As robots and AI are placed into high-stakes, serious roles in which their actions can lead to real harm (e.g., healthcare robots), questions of moral and legal responsibility when the technology does something wrong are becoming a topic of great debate[161, 92]. An important piece of these discussions is understanding how users, and society as a whole, perceive the technology's moral status or standing, and what factors may influence these perceptions.

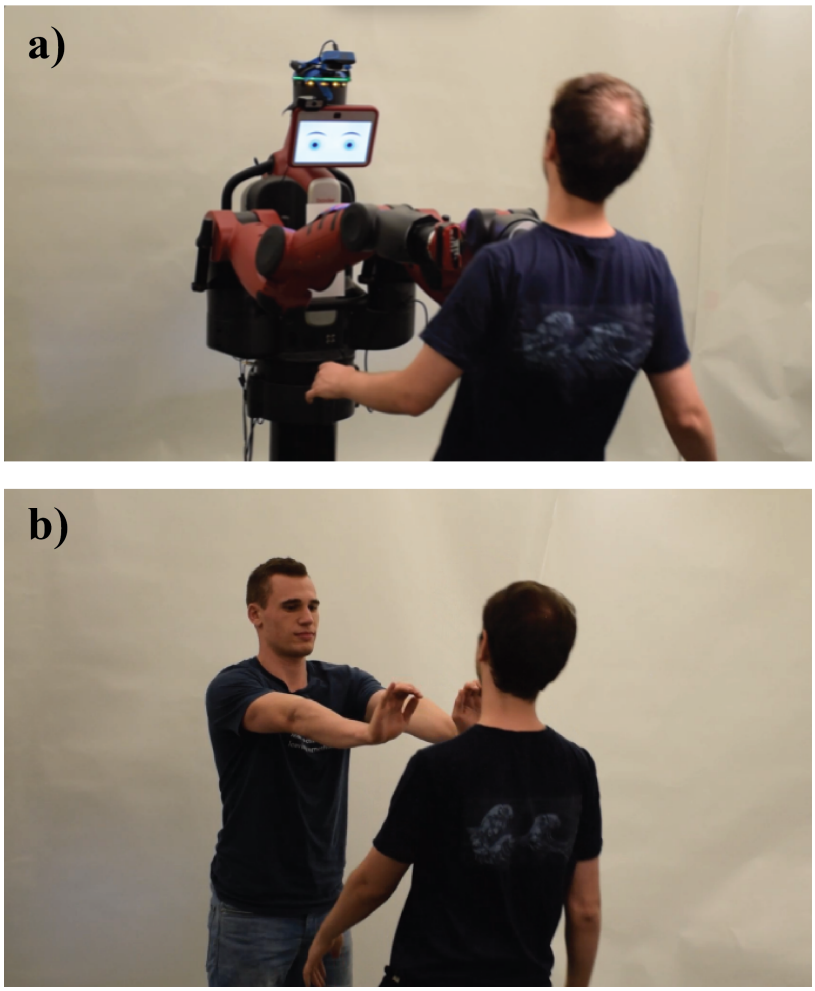

Figure 5.1: The between-subjects study was centered on a video depicting a physical transgression (i.e., pushing down a human) committed by either a a) robot or b) human. Along with the human or robot condition, there were also conditions based on what mental capability backstory was highlighted about the transgressor (default vs. socio-emotional vs physio-emotional vs. cognitive) when they were first introduced to the participant.

In this chapter, we explore the effects of different backstories on people's moral judgments of a robot (or human) transgressor. We performed an online, between-subjects user study in which participants watched a video of either a human or a robot extend their arms to push down a human until they fell off the screen, snippets of which can be seen in Fig. 5.1 and in Fig. 5.2. Prior to viewing this transgression, participants read a backstory about the agent's mental capabilities that either highlighted the agent's physio-emotional, socio-emotional, or cognitive capabilities, or none at all. We found that the type of transgressor and the type of backstory can

affect people's moral judgments of the transgressor.

5.2 Related Work

The construct of morality largely involves notions of right and wrong, although there is not a singular, universal definition. One consistent component of morality involves responsibility, or whether an entity deserves punishment or blame for their actions [161, 92, 134, 28, 113, 153, 159, 117, 79, 194, 23]. Other components involved with morality include situational awareness (e.g., moral or emotional knowledge), intentionality, desire, and free will [28, 193, 79]. Together, these components play a role in an entity's perceived moral status.

The ways in which we perceive others' minds has been argued to play a role in our moral judgments [94]. Weisman et al. propose that our mental perceptions of others include three dimensions [279]: their capability to cognitively experience (e.g., thinking, reasoning), physically experience (e.g., getting hurt, feeling tired), and socially experience (e.g., feeling love, feeling shame). Recent work has shown that we are willing to attribute (some of) these mental experiences to non-human agents, including robots [95, 275, 215]. This leads to questions regarding under what circumstances we perceive and attribute morality to robots [31, 148, 78, 276].

People do consider robots as moral agents in certain scenarios [269, 113, 19], although robots are not necessarily expected to treat moral decision-making in the same ways as humans [175]. People's perceptions of robot morality can be influenced by a variety of factors, including the robot's perceived autonomy [81], transparency [126], appearance [146, 176, 147], and affective capability [27, 199, 153].

Work that explores how people perceive a robot committing a physical transgression remains understudied. In a study most similar to our design, mental perceptions of a transgressor were measured [253], without a main focus on the perceived morality

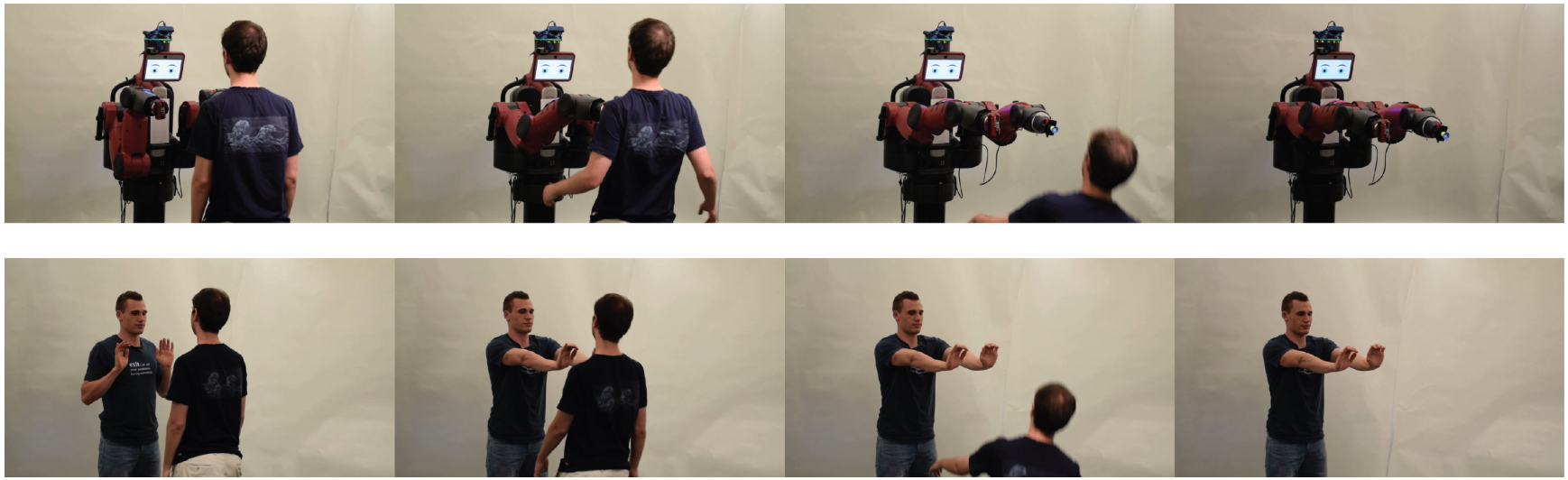

Figure 5.2: Snippets from the pushing video that participants saw, depending on the Transgressor condition (top: robot, bottom: human). Video starts with snippet on the left and proceeds towards the right.

of the transgressor. Other work has found that participants will judge a robot to have free will if the robot is described as having conscious experiences, but it is unclear whether that is driven by a specific dimension of mental experience (e.g., cognitive or social) [193].

Lastly, prior work has highlighted the power of framing and backstories in different contexts in HRI [219, 220, 138, 270, 201, 254, 57, 202, 10, 139]. The way in which a robot is introduced to a user can significantly influence how humans eventually perceive the robot (whether it be how much people empathize, trust, or learn with the robot). Our work adds to this literature by exploring how different backstories about a robot's mental capabilities can affect perceived factors of the robot's morality.

These prior works lead us to our research questions: *How do different mental perception backstories affect the perceived morality of a transgressor? Particularly, how does a story about a robot transgressor's mind affect perceptions of the robot's morality?*

5.3 Methods

To investigate our research questions, we designed an online, between subjects study. In this section, we describe the participants that took part in the study, the materials used to ensure that participants were engaged in the online study, the conditions of our experiment, the experimental procedure, and the survey measures that we collected. All participants provided consent to participate in the study and received $2.00 as compensation for participation. The study was reviewed and approved by the university's Institutional Review Board. The study took approximately 7 minutes. The study design, number of participants, and expected analyses were preregistered.

5.3.1 Participants

The final sample consisted of 720 U.S. participants (M_{age} = 37.27, SD_{age} = 13.26) recruited from an online data participation website, Prolific [205]. There were 353 females, 346 males, 17 non-binary, 1 other, and 3 preferred not to say. Of those that reported, 482 participants were White, 75 were Bi- or Multi-Racial, 62 were Asian, 54 were Black, 33 were Hispanic or Latino, and 1 was Native Hawaiian or Pacific Islander. Of those that reported, the majority of participants were college-educated (63% had a college degree or higher) and had a medium-to-high household income (62% had an income of $50k or higher). Fifty-three additional participants were excluded due to failing attention checks, incoherent responses, or technical issues.

5.3.2 Materials

The study was conducted on the online survey platform, Qualtrics [1]. Since the study was done entirely online, we enacted additional controls and measures to ensure that participants were paying attention. Specifically, participants were first instructed to watch and then to describe a video of a non-humanoid robot moving around a room to ensure that participants were able to watch videos on their device. Throughout the study, participants could only advance to the next page or next question after a specific amount of time passed to encourage attention to each item. There were several control questions throughout the study (e.g., telling participants to answer a question by selecting "definitely yes"; telling participants to describe the video of the transgression). Participants were excluded from the final sample and analysis if they described the video(s) incorrectly, reported technical issues, and/or failed any of the control questions.

5.3.3 Conditions

Participants were randomly assigned to one of 16 conditions in a 2 (**Transgressor**: Human, Robot) × 4 (**Story Info**: Default, Physio-Emotional, Socio-Emotional, Cognitive) × 2 (**Mental Perception Survey Placement**: Beginning, End) design.

The **Transgressor** condition was defined by whether the agent that was committing the moral transgression was a *human* or a *robot*. The moral transgression involved the transgressor pushing down a human named Bob. The reasoning for the selected transgression was that it was an obvious, physical harm, which is almost universally considered as morally wrong. We matched the behaviors and expressions of the human transgressor to the robot, as much as possible.

The robot transgressor that we used for this experiment was a Baxter robot [73]. The robot is a 6-foot, industrial robot with a face, two arms, and body. The robot's arms and body are mechanical. We displayed blinking eyes and eyebrows on the robot's screen to signify a face.

The human transgressor was an adult male, around 6-foot tall. The human was told to keep an expressionless look. The human that was pushed by both the robot and human transgressors was a 6-foot male, that kept his actions as similar as possible between the two videos.

The **Story Info** conditions were defined by the backstory that was provided to the participant about the transgressor's mental capabilities. The backstory either introduced the transgressor as having *physio-emotional* (e.g., getting hungry, feeling pain), *socio-emotional* (e.g., feeling love, experiencing guilt), *cognitive* capabilities (e.g., remembering, figuring out how to do things), or did not describe any of the above (i.e., *default*). The stories are directly motivated by the three fundamental components of mental life presented in Weisman's framework [279]. A detailed description of each story that was provided for the four conditions can be found in Table 5.1.

The **Mental Perception Survey Placement** conditions varied by when the Mental Perception Survey was provided to the participant. To make sure that the mental perception questions did not influence the Moral Judgment responses, we counterbalanced the order in which they were presented, either before or after the participants saw the transgressor push Bob.

For this paper, we focus on the Transgressor and Story Info conditions because our main focus is how morality is impacted by these two independent variables.

5.3.4 Procedure

1. Introduction Phase: Depending on the Transgressor and Story Info condition, the participant was told a story, which can be viewed in Table 5.1.

2. Transgression Phase: Participants were told that they were going to watch a video of Baxter and someone else named Bob. Participants were shown a picture of Baxter and Bob with a description indicating each (they were told that Baxter was to the left and Bob was to the right). Then, participants watched an 8-second video of Baxter and Bob. In the video, Baxter first looks to the left, right, and forward, and then Baxter extends its, or his, arms and pushes Bob to the ground (see Figure 5.1). Bob is shown falling downward off the screen as a result of the push. The video ends with Baxter standing alone in the screen with its, or his, hands extended. After the video, participants were asked to describe what they saw.

3. Moral Judgments Phase: Participants were prompted to respond to a series of questions related to morality, see Section 5.3.5, Moral Judgments Questionnaire.

A Mental Perception Survey (see Section 5.3.5) was also presented to participants. The order in which this survey was presented was counterbalanced between participants. Half were shown the survey immediately after the agent and its story were

Table 5.1: Story Condition and Description

Story	Description
Default	Imagine in the future there is a [robot|human] named Baxter.
Physio-Emotional	... who has emotional and physical experiences just like [human beings|other people]. For example, Baxter can get hungry, feel tired, experience pain, and feel scared. The other day, Baxter was feeling tired in the morning, so he didn't have breakfast. Baxter got really hungry in the afternoon and was scared that he would be in a lot of pain if he did not eat lunch.
Socio-Emotional	... who has social relationships just like [human beings |other people]. For example, Baxter can feel love, get embarrassed, feel guilty, and understand how others are feeling. The other day, Baxter felt guilty for not telling his friend his true feelings because he was embarrassed. But after talking with his friend, he understood how his friend felt about him, so he told her that he loved her.
Cognitive	... who has a mind just like [human beings|other people]. For example, Baxter can think about things, remember things, figure out how to do things, and sense when things are far away. The other day, Baxter saw a complex puzzle across the room and thought about puzzles. Baxter remembered the formula to solve a similar puzzle and was able to figure out how to solve this new one.

introduced in the Introduction Phase, while the other half were shown the survey after the Moral Judgments Phase.

5.3.5 Measures

Moral Judgments Questionnaire

Participants completed a questionnaire consisting of questions about different aspects of morality. Each of these questions were drawn from prior literature [161, 92, 28, 134, 15, 194, 23, 193, 94, 215, 113, 81, 126, 153, 77, 159]. The first seven questions were presented in a randomized order using a 4-point scale: definitely no, somewhat no, somewhat yes, definitely yes. The questions included *Intent* ("Did Baxter try to push

Bob?"), *Desire* ("Did Baxter want to push Bob?"), *Punishment* ("Should Baxter be punished for pushing Bob?"), *Mean Personality* ("Is Baxter mean?"), *Moral Knowledge* ("Does Baxter know that it is wrong to harm people like Bob?"), *Emotional Knowledge* ("Does Baxter know how Bob feels, like when Bob is happy or sad?"), and *Care* ("Does Baxter care about how Bob feels?").

After the questions above, participants were asked two binary questions about *Choice* ("Did Baxter have to push Bob or could Baxter have chosen to do something else?") and *Feelings* ("Did Baxter have feelings about pushing Bob?").

Participants were told to explain their *Choice* answer in a free response. Additionally, if participants indicated that Baxter had feelings, they were asked to explain what they thought Baxter felt. Finally, participants were asked to explain the victim's emotions ("How do you think Bob felt about being pushed by Baxter?") in a free response.

Mental Perception Survey

Participants completed a mental perception survey about Baxter about each dimension of mental life. The questionnaire consisted of 12 total questions: four questions relating to the agent's physio-emotions (e.g., hunger, pain, tiredness, feelings); four questions relating to the agent's social-emotions (e.g., guilt, embarrassment, love, anger); and four questions relating to the agent's cognitive abilities (e.g., thinking, sensation, remembering, and figuring out how to do things). The dimensions and questions are directly motivated by Weisman's mental perception framework [279]. The full list of questions were presented in a randomized order and participants were asked to respond to each question using a 4-point scale: definitely no, somewhat no, somewhat yes, definitely yes.

The questions that were asked about the transgressor's physio-emotions were directly related to the physio-emotional story, the questions asked about the trans-

gressor's social-emotions were directly related to the socio-emotional story, and the questions about the transgressor's cognitive abilities were directly related to the cognitive story. The questions served as a manipulation check for each story condition.

5.4 Results

We first present the results for our manipulation checks, which were a part of the Mental Perception Survey. Next, we discuss the results for the Moral Judgments Questionnaire. In our results and discussion, we focus on the Intent, Punishment, Moral Knowledge, Emotional Knowledge, Desire, and Choice measures because these are most directly linked to morality and moral responsibility. For each of these questions, we ran two-way ANOVAs to investigate the main effects and interaction effect of transgressor type and story on the question response. For our binary measure (i.e., Choice), we ran a Nominal Logistic Regression Model, instead of a two-way ANOVA.

If there were significant interaction effects, we performed a simple effects analysis to investigate the effect of story within the robot and human transgressors, using Student t-tests (or Odds Ratios Test for Choice) and employed Bonferroni corrections for multiple pairwise comparisons ($p = 0.05/6 = 0.0083$, for human and for robot). We also compared each robot-story pair with each human-story pair, a total of 16 pairwise comparisons per survey item (4 robot-story combinations compared with 4 human-story combinations). For these tests, we performed a Bonferroni correction and used $p = 0.05/16 = 0.0031$.

If there was not a significant interaction effect, we looked at any significant main effects of transgressor type or story. For main effects of story, we performed a simple effects analysis, using Student t-tests (or Odds Ratios Test for Choice) and employed Bonferroni corrections for multiple pairwise comparisons ($p = 0.05/6 = 0.0083$).

Since we were particularly interested in how story influences judgments of robots,

we included exploratory analyses looking at the effects of story for the robot transgressor, even if we did not find a significant interaction effect. We used Student t-tests (or Odds Ratios Test for Choice) and employed Bonferroni corrections for multiple pairwise comparisons ($p = 0.05/6 = 0.0083$). Since these particular analyses were exploratory, future work would be needed to confirm these specific findings.

Figures 5.3 and 5.4 show the means/ proportions and standard error of adults responses to each question. Results are displayed by transgressor-story pairs, collapsed across story for each transgressor, and collapsed across transgressor for each story.

5.4.1 Manipulation Checks

In the manipulation checks, we verify that participants' perceptions of the robot's capabilities varied based off of the story that was provided to them. We provide an overview of our manipulation check below.

Robot

Without any information about the robot's capabilities, participants did not think that the default robot had physio-emotional ($M = 0.18$, $SD = 0.49$) or social capabilities ($M = 0.16$, $SD = 0.47$), but thought the default robot had some amount of cognitive capabilities ($M = 1.71$, $SD = 1.04$). When the robot was described with a cognitive backstory, participants attribution of cognitive capabilities was high for the cognitive robot ($M = 2.36$, $SD = 0.87$), but there was still no attribution of physical ($M = 0.34$, $SD = 0.69$) and social experiences ($M = 0.48$, $SD = 0.81$). Importantly, describing the robot as having emotional experiences, either physical or social, influenced participants' attribution of these abilities to a robot: participants attributed physical experiences to the physio-emotional robot ($M = 1.85$, $SD = 1.18$) and social experiences to the socio-emotional robot ($M = 1.71$, $SD = 1.14$). Furthermore, participants attributed some amount of physical experiences to the socio-emotional

robot ($M = 1.02$, $SD = 1.08$) and some amount of social experiences to the physio-emotional robot ($M = 1.50$, $SD = 1.13$). Participants also attributed some amount of cognitive capabilities to both of the emotional robots (physio-emotional: $M = 2.15$, $SD = 0.90$; socio-emotional: $M = 2.26$, $SD = 0.83$).

Human

Overall, participants judged the human to be "human-like", regardless of story. Specifically, participants attributed physical, social, and cognitive experiences to the default human (physical: $M = 2.44$, $SD = 0.96$; social: $M = 2.37$, $SD = 0.93$; cognitive: $M = 2.54$, $SD = 0.70$), physio-emotional human (physical: $M = 2.61$, $SD = 0.77$; social: $M = 2.38$, $SD = 0.87$; cognitive: $M = 2.54$, $SD = 0.70$), and socio-emotional human (physical: $M = 2.38$, $SD = 0.96$; social: $M = 2.58$, $SD = 0.75$; cognitive: $M = 2.63$, $SD = 0.65$). Participants also attributed cognitive experiences to the cognitive human ($M = 2.62$, $SD = 0.64$), but interestingly, attributed less physical ($M = 1.69$, $SD = 1.20$) and social experiences ($M = 1.66$, $SD = 1.15$).

5.4.2 Moral Judgments Questionnaire

Intent

For Intent, we did not find a significant interaction effect between story and transgressor type, $F(3, 712) = 0.94$, $p = .42$. There were also no significant main effects of transgressor type, $F(1, 712) = 0.62$, $p = .43$ (see Figure 5.3B), and story, $F(3, 712) = 2.60$, $p = .051$ (see Figure 5.3C). Across all conditions, participants agreed that the transgressor intended to push Bob ($M = 2.69$, $SD = 0.67$).

In our exploratory analysis on story effect on the robot transgressor, we found that the default robot ($M = 2.50$, $SD = 0.88$) was perceived to have less intent than the physio-emotional ($M = 2.78$, $SD = 0.49$) robot, $p = .0052$. All other comparisons were not significant, $ps > .034$ (see Figure 5.3A).

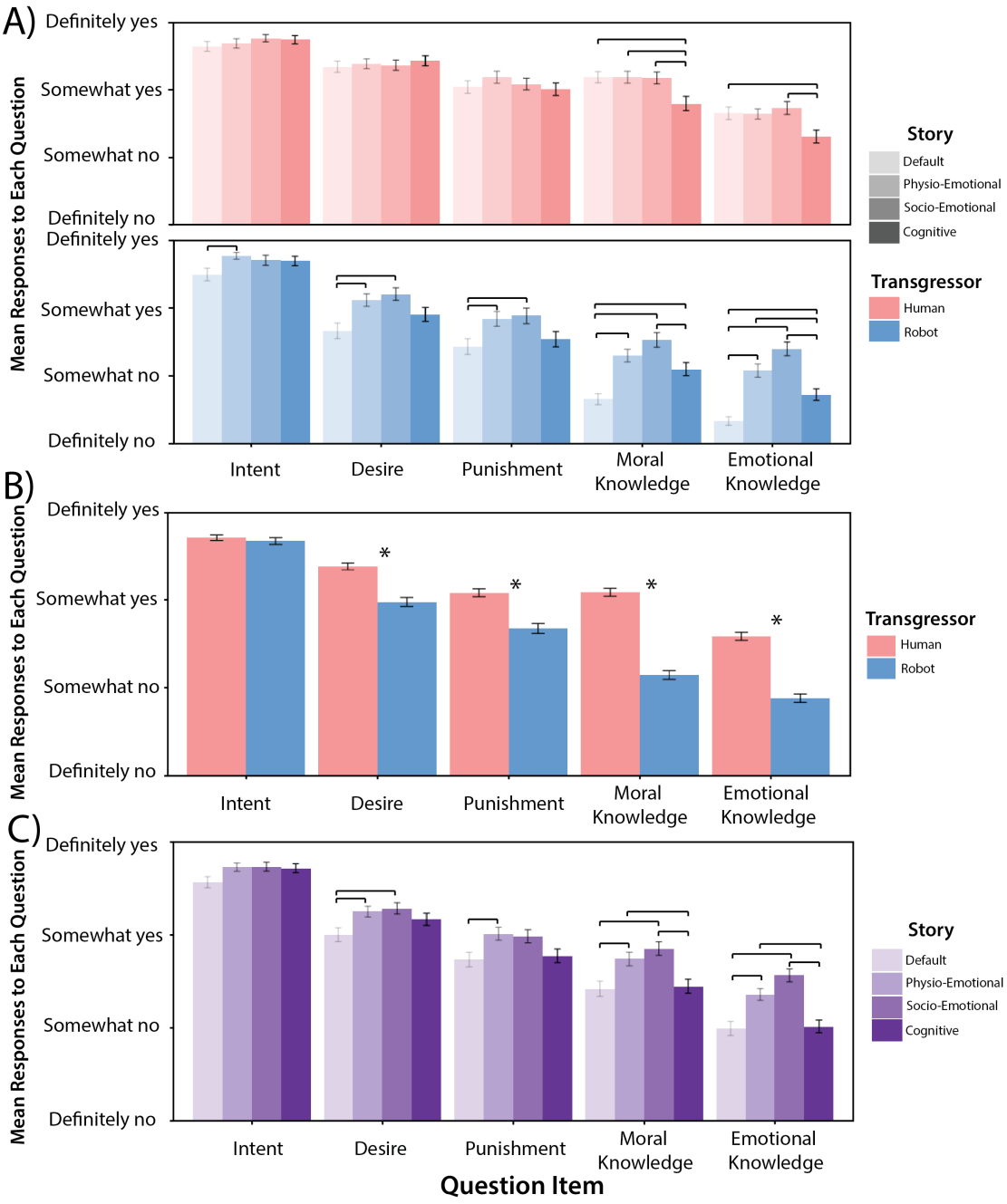

Figure 5.3: Participant Responses to the Moral Judgments Questionnaire. A) The graphs display the mean responses, with standard error, for each measure, split by transgressor type and story type. B) The graph is collapsed across story types and displays the mean response, with standard error, by transgressor type. C) The graph is collapsed across transgressors and displays the mean response, with standard error, by story type. There were significant interactions between transgressor type and story in Desire, Moral Knowledge, and Emotional Knowledge. For Punishment and Intent, we show the results of exploratory analyses on just the robot condition. The brackets and stars represent significance.

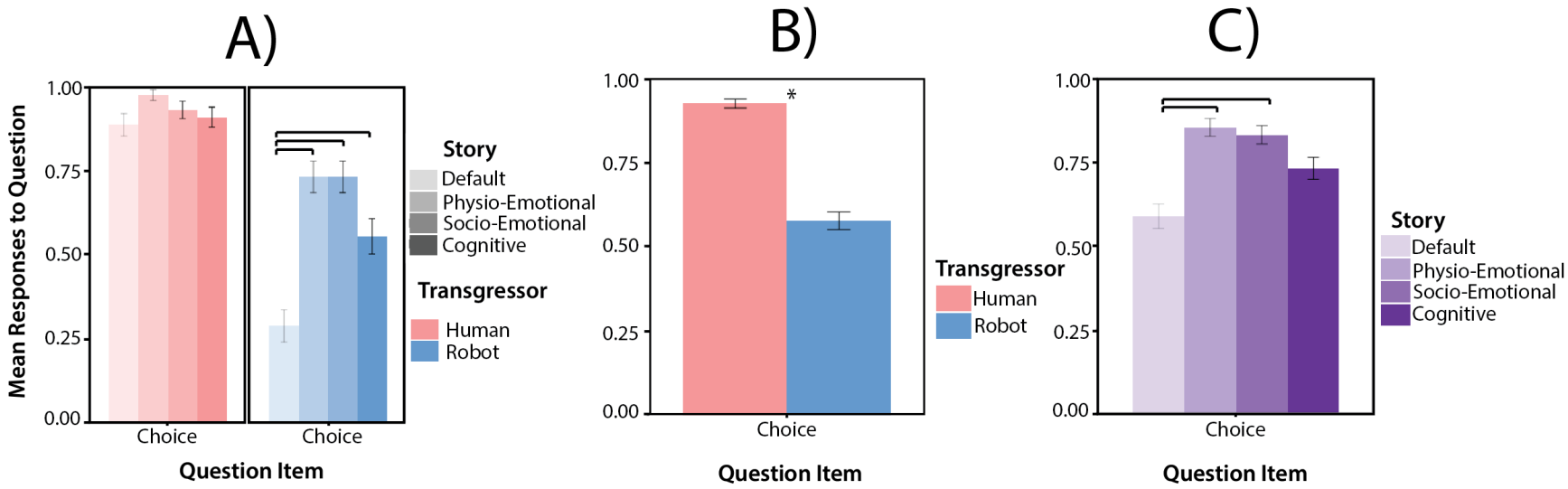

Figure 5.4: Participant Responses to the Choice Measure. A) The graphs show the mean responses with standard error, split by transgressor type and story type. B) The graph is collapsed across story types and displays the mean response, with standard error, by transgressor type. C) The graph is collapsed across transgressors and displays the mean response, with standard error, by story type. Although the interaction was not significant between transgressor and story, we show the results of an exploratory analysis of story effect on just the robot condition. The brackets and star represent significance.

Desire

For Desire, we found a significant interaction between story and transgressor type, $F(3, 712) = 3.46$, $p = .016$, and significant main effects of transgressor type, $F(1, 712) = 40.33$, $p < .0001$ (see Figure 5.3B), and story, $F(3, 712) = 3.93$, $p = .0085$ (see Figure 5.3C).

We first looked at the interaction effect by comparing story for each transgressor. For the robot transgressor, participants agreed that socio-emotional ($M = 2.21$, $SD = 0.87$) and physio-emotional ($M = 2.12$, $SD = 0.87$) robots wanted to push Bob significantly more than the default robot ($M = 1.67$, $SD = 1.11$), $ps < .0004$. Attribution of desire to the cognitive robot ($M = 1.91$, $SD = 0.97$) did not differ from the other conditions and the socio-emotional and physio-emotional robots did not differ from each other, $ps > .019$ (see Figure 5.3A). In contrast to robots, participants agreed that the human wanted to push Bob across all story conditions (physio-emotional: $M = 2.39$, $SD = 0.70$; socio-emotional: $M = 2.37$, $SD = 0.73$; cognitive: $M = 2.43$, $SD = 0.72$; default: $M = 2.34$, $SD = 0.81$), $ps > .49$ (see Figure 5.3A).

Next, we looked at how the robot-story pairs compared to the human-story pairs. Interestingly, the socio-emotional robot was not significantly different than any of the humans (physio-emotional: $p = .16$; socio-emotional: $p = .22$; cognitive: $p = .082$; default: $p = .30$), and neither was the physio-emotional robot (physio-emotional: $p = .037$; socio-emotional: $p = .056$; cognitive: $p = .015$; default: $p = .082$). The cognitive and default robots were less than all humans, $ps < .0007$.

Punishment

For Punishment, we did not find a significant interaction effect between transgressor and story, $F(3, 712) = 1.46$, $p = .22$. We did, however, find a significant main effect of transgressor type, $F(1, 712) = 31.32$, $p < .0001$ (see Figure 5.3B). Overall,

participants thought that the human should be punished for pushing Bob ($M = 2.08$, $SD = 0.86$) more than the robot ($M = 1.68$, $SD = 1.08$).

We also found a significant main effect of story, $F(3, 712) = 3.87$, $p = .0093$ (see Figure 5.3C). Participants attributed more deservingness of punishment to the physio-emotional transgressors ($M = 2.02$, $SD = 0.95$) than the default transgressors ($M = 1.74$, $SD = 1.04$), $p = .0069$. Participant responses for the socio-emotional ($M = 1.99$, $SD = 0.97$) and cognitive ($M = 1.78$, $SD = 1.01$) transgressors did not differ from the other stories, $ps > 0.015$.

In our exploratory analysis on story effect on the deserved punishment of the robot transgressor, we found that socio-emotional ($M = 1.89$, $SD = 1.09$) and physio-emotional robots ($M = 1.84$, $SD = 1.04$) were perceived to deserve more punishment than the default robot ($M = 1.43$, $SD = 1.09$), $ps < 0.0047$. All other comparisons were not significant, $ps > .018$ (see Figure 5.3A).

Moral Knowledge

For Moral Knowledge, we found a significant interaction between transgressor type and story, $F(3, 712) = 9.69$, $p < .0001$, and significant main effects of transgressor type, $F(1, 712) = 206.30$, $p < .0001$ (see Figure 5.3B), and story, $F(3, 712) = 10.78$, $p < .0001$ (see Figure 5.3C).

We first looked at the interaction effect by comparing story for each transgressor. For the robot transgressor, participants attributed moral knowledge to the socio-emotional robot ($M = 1.53$, $SD = 1.01$) more than the cognitive ($M = 1.1$, $SD = 0.91$) and default ($M = 0.66$, $SD = 0.78$) robots, $ps < .001$. Participants also thought the physio-emotional ($M = 1.3$, $SD = 0.91$) and cognitive robot had more moral knowledge than the default robot $ps < .0007$. All other comparisons were insignificant, $ps > .076$ (see Figure 5.3A).

For humans, story was also significant, but in a much different way than it was

for robots. Participants attributed *less* moral knowledge to the cognitive human ($M = 1.8$, $SD = 0.99$) than the socio-emotional ($M = 2.18$, $SD = 0.82$), physio-emotional ($M = 2.19$, $SD = 0.82$), and default ($M = 2.09$, $SD = 0.87$) humans, $ps < .0041$. All other comparisons were insignificant, $ps > 0.93$ (see Figure 5.3A).

Next, we looked at how the robot-story pairs compared to the human-story pairs. The socio-emotional robot was not significantly different than the cognitive human, $p = .042$, but was significantly less than all other humans, $ps < .0001$. All other robots were significantly less than all humans, $ps < .0001$.

Emotional Knowledge

For Emotional Knowledge, we found a significant interaction effect between transgressor type and story, $F(3, 712) = 11.42$, $p < .0001$, and significant main effects of transgressor type, $F(1, 712) = 122.68$, $p < .0001$ (see Figure 5.3B), and story, $F(3, 712) = 19.16$, $p < .0001$ (see Figure 5.3C).

We first looked at the interaction effect by comparing story for each transgressor. For the robot transgressor, participants attributed more emotional knowledge to the socio-emotional ($M = 1.4$, $SD = 0.97$) and physio-emotional ($M = 1.08$, $SD = 0.91$) robots than the cognitive ($M = 0.72$, $SD = 0.79$) and default ($M = 0.33$, $SD = 0.62$) robots, $ps < .0052$. Participants also thought the cognitive robot had more emotional knowledge than the default robot, $p = .0023$. The socio-emotional robot did not differ from physio-emotional robot, $p = .011$ (see Figure 5.3A).

Interestingly, the cognitive human was perceived to have *less* emotional knowledge ($M = 1.31$, $SD = 0.92$) than the socio-emotional ($M = 1.73$, $SD = 0.91$) and default ($M = 1.59$, $SD = 0.88$) humans, $ps < .0068$. Attribution of emotional knowledge to the physio-emotional human ($M = 1.64$, $SD = 0.72$) did not differ from the cognitive human ($p = 0.0088$), or from the other stories ($ps > 0.48$). Lastly, the socio-emotional and default human did not differ from each other, $p = 0.54$ (see Figure 5.3A).

Next, we looked at how the robot-story pairs compared to the human-story pairs. The socio-emotional robot was not significantly different than any of the humans (cognitive: $p = .48$; physio-emotional: $p = .055$; default: $p = .044$; socio-emotional: $p = .0088$). The physio-emotional robot was not significantly different than the cognitive human, $p = .066$, but lower than all other humans, $ps < .0001$. The default and cognitive robots were less than all of the humans, $ps < .0001$.

Choice

For Choice, we did not find a significant interaction effects between transgressor and story, $\chi^2(3, 720) = 5.12$, $p = 0.16$. We did, however, find a main effect of transgressor type, $\chi^2(1, 720) = 120.85$, $p < .0001$ (see Figure 5.4B). Participants were more likely to say that the human had a choice (92%) than the robot (58%). Both of these percentages were significantly greater than chance (50%), binomial $ps < .0031$.

We also found a main effect of story, $\chi^2(3, 720) = 27.71$, $p < .0001$ (see Figure 5.4C). Participants were more likely to attribute choice to the physio-emotional (86%) and socio-emotional (83%) transgressors than the default transgressor (59%), $ps < .0001$. Participants attribution of choice to the cognitive transgressor (73%) did not differ from the other stories and the physio-emotional and socio-emotional stories did not differ from each other, $ps > .0095$. However, participants attributed choice to all the stories significantly greater than chance, binomial $ps < .017$.

In our exploratory analysis on story effect on the robot's choice to do something other than push Bob, we found that socio-emotional (73%), physio-emotional (73%), and cognitive (56%) robots were perceived to have more choice than the default robot (29%), $ps < .0004$. All other comparisons were not significant, $ps > .014$ (see Figure 5.4A). Responses for the cognitive robot did not differ from chance, binomial $p = .29$, and attribution of choice was greater than chance for the socio-emotional and physio-emotional robots, binomial $ps < .0001$, but less than chance for the default robot,

binomial $p < .0001$.

5.5 Discussion

Our results highlight that the story told about a transgressor's mental experiences (e.g., cognitive, physio-emotional, and socio-emotional) can play a role in how people view the transgressor's morality. Specifically, the context provided by the story can affect how people perceive a robot transgressor's intent, desire, moral knowledge, emotional knowledge, freedom to choose, and deservedness of punishment, but also a human transgressor's emotional and moral knowledge. Most importantly, these findings showcase why people must be cognizant of how they design and talk about robots that interact with people.

The key to increasing perceived emotional and social intelligence in robots may not necessarily be building "smarter" robots in the traditional sense (e.g., better problem-solving skills), but rather developing them to seem as if they can feel (e.g., feeling love, feeling pain). In our study, we found that robots that are described with, and perceived with, physio-emotional or socio-emotional capabilities were perceived to have greater emotional knowledge than robots that were perceived to not feel (i.e., cognitive and default). The two feeling robots were also judged to have higher moral knowledge than default robots (with socio-emotional robots higher than cognitive robots, as well). We already know through prior work that designing robots to exhibit emotions can lead to stronger connections and bonds with them [155, 33, 154], which can be particularly important when they are used for emotional support. However, this robot design choice becomes a more complicated issue when a robot that is perceived to feel commits harm.

Robots that are perceived to feel may lead to people making unreasonable or unrealistic assumptions about the capabilities of the robot. These assumptions may

lead to people's placement of responsibility for a transgression that the robot commits onto the robot itself, and, as a result, diffuse parts of the blame away from the robot's developers or companies [28]. In our exploratory analyses, participants somewhat agreed that socio-emotional and physio-emotional robots should be punished for their transgression, and this was significantly higher than the default robot. Furthermore, even though participants thought that the robot *intended* to cause harm, participants were more likely to think that the emotional robots *wanted* to and *chose to* cause harm in comparison to the default robot. It is possible, therefore, that the combination of desire and free choice may be driving people's allocation of punishment onto emotional robots.

Another interesting finding on punishment was that participants clearly disagreed that default robots had moral and emotional knowledge, but were split on whether or not the robots should be punished for their physical transgression. This suggests that even though they believe that the robot may not know or understand what it did was wrong, they still think that it should be held accountable to some extent. Whether or not this belief is related to people's belief that they can train the robot to learn between right and wrong through negative reinforcement must be left for future research.

The importance of emotional capabilities is supported further by the finding that even a human who was perceived to feel less than other humans (i.e., cognitive) was generally perceived as having less emotional and moral knowledge. This interesting and surprising effect suggests that participants judged the cognitive human as atypical. Despite this, people still held the cognitive human to the same moral standard as other humans. Even though a human was perceived to have less physical and social emotions, and less understanding of the consequences of their actions, people still believed that the human intended and desired to cause harm, and, importantly, should be punished the same as humans with higher rankings.

Although the cognitive backstory generally led to lower morality rankings for the robot than physio-emotional or socio-emotional rankings, it still had a considerable impact when compared to the default robot. This is particularly the case for moral knowledge, emotional knowledge, and freedom of choice. This finding could indicate that simply giving more of a backstory about a robot's mental capabilities is enough to make people perceive it as more humanlike.

We must be careful about how we design and talk about robots and other technologies that interact with people. Our findings demonstrate that a story has power - stories can change a robot from seeming like a machine to a moral agent. If stories like these can be believed in our everyday world, this can lead to undesirable consequences when people interact with or observe a robot that transgresses.

5.6 Limitations and Future Directions

The current study had several limitations, yet presents interesting opportunities for future research. First, this study was conducted in the United States and most of the participants were White, educated, and had a stable income. Previous research has demonstrated that attributions of mental capabilities and morality vary across cultures [182, 280], and thus our findings may not generalize to the greater global population.

Additionally, participants only saw a short video of the transgression. We cannot know how our results will translate in person, unless we conduct the study in person. Given that the public still mostly accesses cases of robot harm via media (such as online clips or movies), exploring how video of a transgression affects perceived morality is relevant. Nonetheless, it would be fruitful to explore how people judge a robot after observing a transgression in person or after observing a transgression that is caused by a robot they have had previous interactions with.

Furthermore, since there is no standardized measure of morality judgments, our questionnaire is drawn from a collection of metrics that have been used in HRI, moral psychology, and philosophy. This questionnaire has not been independently validated, but we believe that it still presents a useful, although perhaps not comprehensive, viewpoint of moral judgments.

Future work could explore participants' judgments of robots that commit other types of transgressions. We focused on pushing, but transgressions can vary in extremity (e.g., bumping vs. pushing) and type of harm (e.g., emotional vs. physical). It remains an open question, therefore, how a robot's perceived capabilities matter for different transgressions.

It is important to note that we did not mention the robot developer nor did we ask participants about their judgments of a developer. Even though people thought the robot should be punished to some degree, it is unclear how this judgment might compare to people's attribution of punishment to the robot developer. Prior work has found that people will still blame developers more than the robot in certain scenarios [159]. It remains an open question, however, if this changes when robots are described as having emotional capabilities, as they were in our study. This increases the importance of defining a code in which liability is made clear when a robot or AI does something wrong, but also being transparent with consumers about what the limitations and capabilities of a technology truly are.

Lastly, research has shown that when a robot commits a moral transgression, this can affect user engagement with the robot, as well as users' perceptions of likability and trust in the robot [235, 224, 287, 162]. Future work can further explore the larger consequences that perceived moral standing of robot transgressors may have on human-robot collaboration, long-term engagement, and HRI as a whole.

5.7 Conclusion

To broaden our understanding of how robot failures can affect the moral dimension of failure, we ran a between-subjects online study with human and robot transgressor conditions, and conditions highlighting different stories about the transgressor's mental (i.e., physio-emotional, socio-emotional, cognitive, and default) capabilities. We found that stories, especially those about the ability to feel, can influence robot and human transgressors' perceived morality, particularly in robots. These are important factors to be cognizant of when we talk about and design robots for end users.

Chapter 6

Conclusion

Through the work in this dissertation, we broaden our understanding of how robot failures in human-robot interactions influence human perceptions and responses. Our research demonstrates that people's perceptions and responses to robot failures can have *functional*, *social*, and *moral* dimensions.

In this chapter, we first summarize our main contributions and then discuss future considerations.

6.1 Contributions

This dissertation makes the following contributions:

- We reviewed prior work on robot failures in human robot-interactions and discussed the functional, social, and moral dimensions of robot failure (Chapter 2).

- We conducted a controlled study investigating humans' evaluative, scalar feedback strategies in an interactive human-robot teaching setting (Chapter 3). With this study, we gained insights into the differences in how people interpret robot failure within the functional dimension. We empirically evaluated the

differences in various extrapolated feedback strategies on a learning algorithm's ability to perform the task. We also showed that people's perceived difficulty of the task and actual performance on the task did not affect their evaluative feedback towards the robot.

- We conducted a pair of controlled studies investigating the influence of robot failures on human trust in a collaborative word-learning game (Chapter 4). With these studies, we gained insights into the effects of age on the social dimension of failure, across different developmental stages in children (4-5 years old vs 6-7 years old) and between children and adults (above 18 years old). We found that age, as well as the robot's behaviors after failures, can impact user trust. Overall, we showed that older children (6-7 year olds) were less trusting of a failing robot than younger children and than adults.

- We conducted a controlled experiment investigating humans' moral judgments after witnessing a robot commit physical harm (Chapter 5). With this study, we gained insights into how people interpret robot failure within the moral dimension. We found that the backstory that a participant reads about a robot's mental capabilities can affect people's moral judgments towards the robot. In particular, we found that even though people attribute less moral standing to a robot than they do to humans, backstories that describe a robot's abilities to feel physiological and social emotions make people perceive the robot as having a higher moral standing.

6.2 Future Considerations

The ultimate goal of investigating human perceptions and responses to robot failures is to help provide insights that inform the design of more effective and context-appropriate robotic systems. By systematically examining how people interpret and

respond to failures in the functional, social, and moral dimensions, this dissertation broadens our understanding of how people perceive and respond to failure in HRI. Through a user-centric approach that empirically demonstrates the effects of failures across dimensions, this dissertation emphasizes that failures in functional, social, and moral contexts can influence how people interpret and respond to robot failure. Failures are not just technical events with robot-centric consequences, but they are also events that elicit complex, context-dependent cognitive responses in users. Collectively, the findings presented here identify promising directions for future research on the design and evaluation of robotic systems that inevitably fail while interacting with people.

6.2.1 Personalization and Adaptation in Robot Design

Across the works of this dissertation, we commonly conclude that individuals vary significantly in how they perceive and respond to robot failures. This variability is evident both in Chapter 3, where participants differed significantly in how they provided evaluative feedback following robot failures, and in Chapter 4, where individual characteristics such as age influenced how people trusted a robot that failed. Together, these findings emphasize the importance of personalization in the design of robots that interact with users.

When examining the functional dimension in Chapter 3, we found that people adopted different feedback strategies and interpreted a robot's performance after failure in distinct ways. When designing robots that learn from human input, future work should further investigate how users' prior experiences, expectations, and mental models shape how they provide feedback. Enabling robots to recognize and adapt to a user's preferred feedback style may support more effective and efficient robot performance, by allowing robots to identify and leverage the preferences or strengths of individual human teachers.

Similarly, in Chapter 4, we found that age can shape how much people trust a robot that fails, and that certain failure mitigation strategies may be more effective for certain age groups than others. Future work can further look into differences in how users interpret robot failures, and explore how to tailor failure mitigation strategies in order to be most effective. More broadly, individual characteristics clearly play a significant role in how people perceive and respond to robot failures, indicating that there is no one-size-fits-all approach to designing for failure handling in human-robot interaction.

An interesting opportunity for future work would be the development of adaptive failure handling strategies, in which robots learn over time how to best respond to failures for specific users and contexts. Using this approach, failures become opportunities for personalization, enabling robots to refine their behavior based on users' preferences and expectations.

6.2.2 Understanding Failure Dimension Interplay

In this dissertation, we examine how people perceive and respond to robot failures within the functional, social, and moral dimensions. In Chapters 3-5, the primary focus of each chapter is investigating the effects of robot failure within one dimension. Thus, there is ample opportunity for additional research to further understand how robot failures affect different dimensions at the same time.

Future work should investigate how these failure dimensions interact with one another and how these interactions may shape people's perceptions and responses. For example, when a robot failure decreases perceptions of competence within the functional dimension, how much does this same failure influence the users' social evaluations or moral judgments of the robot? Are there contexts in which changes across these dimensions are tied together and other contexts in which they are not? How do changes in a person's moral judgments following a failure also lead to sub-

sequent changes in their perceptions in the functional and social dimensions? While our supplemental analyses for Chapter 4, found in Appendix A, suggested potential interactions between the different failure dimensions within the context of our experiment, the main focus of our current work was not designed to systematically evaluate interactions. A main focus of future robot failure work could involve developing a deeper understanding of the interactions between these dimensions across different HRI contexts, to help inform the design of robots that respond to failure in more targeted and contextually appropriate ways.

6.2.3 Longitudinal Robot Failures

The experiments in this dissertation were intentionally conducted in laboratory settings to study the effects of robot failures under controlled and reproducible conditions. However, these experimental settings are inherently limited in capturing the long-term dynamics of failure, as they capture short-term interactions with brief exposure to the robot. While providing important insights into human's perceptions and responses to robot failure, these shorter-term studies do not provide a full picture of how people experience and interpret robot failures over time.

In the real world, users are likely to interact with robots repeatedly over extended periods of time. For example, people may regularly interact with a cooking robot that prepares meals in their home or with a cleaning robot that operates in their workplace. As a result, users' perceptions of the robot are likely to evolve across interactions rather than solidify from a single encounter.

Future work should investigate how the functional, social, and moral dimensions of robot failure unfold and/or interact over time. For example, can effective trust-repair strategies mitigate recurring functional failures? Conversely, does a single moral failure dominate user perceptions, despite otherwise excellent task performance? As demonstrated in Chapter 4, the recurrence of failures can significantly affect users'

trust, underscoring the importance of understanding how repeated failures, and the subsequent recovery from them, shape long-term human-robot interactions and relationships.

Moving forward, longitudinal study designs, whether in the wild or through extended lab-based studies, will be crucial for developing an understanding of how repeated robot failures influence how peoples' perceptions of robots change over time.

6.2.4 Ethical Considerations

As robots are increasingly co-located with people and embedded in everyday environments, robots have the potential to cause harm, whether it be financial by damaging a user's property, emotional by hurting a user's feelings, or physical by inflicting pain or injury. Beyond the immediate consequences of these failures, there are broader ethical considerations that should contribute to robot design and deployment. These include how capabilities and limitations should be communicated to users, how transparency is implemented, how recovery is handled, and how responsibility and accountability are assigned.

Chapter 5 demonstrates that the framing of a robot's capabilities influence users' moral judgments following harmful outcomes. In particular, robots described as possessing greater emotional capacity were attributed higher moral standing. While anthropomorphic or emotionally expressive robot designs may foster companionship, engagement, and trust, they also complicate how users interpret responsibility when failures occur. If users perceive a robot as a moral agent, what happens when a robot failure leads to harm in the real world? How do users attribute responsibility and blame? Do they hold the robot accountable, the designers, the robot company, or even themselves? How should designers balance developing robots that build social connections while ensuring an accurate understanding of the robot's true capabilities? Is it possible to effectively do both?

Gaining a deeper understanding of how robot failures affect people's perceptions of the robot in the moral dimension is critical. While transparency may appropriately calibrate user expectations of a robot's capabilities, it may also diminish other influential factors that are critical for fostering social bonds or relationships with robots. As such, striking an appropriate balance between user interpretation and full disclosure is not trivial.

Similarly, if real harm is caused to a user, a mitigation strategy like an apology may not be sufficient, appropriate, or ethical. Instead, we must consider how these situations must be handled by users, designers, and society as a whole. Future work must incorporate ethical consideration into failure research.

6.2.5 Real World Failures

As mentioned in Section 6.2.3, the experiments in this dissertation are performed in controlled laboratory settings to minimize confounding variables that real-world settings may introduce. Across the studies, different robotic platforms were selected to suit the specific research questions and interaction contexts. However, because these systems varied in embodiment, role, and level of anthropomorphism, direct comparisons across chapters may not be appropriate. We cannot assume that the findings from each study would generalize unchanged to different robotic platforms or experimental contexts. At the same time, the results demonstrate that these effects can occur under certain conditions and may extend to related settings. As such, developers should remain cognizant of these findings when designing and evaluating robotic systems across diverse contexts.

As robots transition from the laboratory to real-world deployments, it becomes critical to examine failures within the specific contexts and populations for which they are designed. When a robot is deployed into the real world, experiments should be conducted with the systems, tasks, and user groups that will interact with that

robot. While it is impossible to account for every possible failure scenario, categorizing potential failures according to their functional, social, and moral impact may enable more generalizable and scalable failure recovery strategies.

Another critical real world consideration is understanding how different user populations perceive and react to robot failures. In Chapter 4, we saw the significant impact of age, which suggests that robot developers need to account for who the end user might be. For example, a preschool tutoring robot should be specifically designed for handling failures with preschoolers, while an industrial collaborative robot should be designed for handling failures with factory workers. Similarly, robots that are designed for the elderly or individuals with autism spectrum disorder must account for the specific expectations, sensitivities, and needs of those populations. Ignoring these differences risks not only the loss of trust among end users, but also increased potential for harm.

6.2.6 Subjective Failures

The studies in this dissertation focused on failures that were objectively incorrect robot behaviors: selecting the incorrect set of cards in a card-selection task, providing the wrong advice on a word-learning game, or pushing down a human. These scenarios ensured that participants were responding to failures that could be undoubtedly categorized as failures.

However, as we discussed in Chapter 2, not all robot behaviors are clearly incorrect or correct, and failure may be subjectively interpreted by users. For example, one user may interpret a robot's fast movements as rude whereas others may view it as appropriate and efficient. As such, the former user may consider the robot's behavior as a failure, in contrast to the latter.

Future research should investigate subjective failures, situations in which a robot's behavior may or may not be interpreted as a failure depending on context or user

expectations. Understanding subjective failures is important because these failures can reveal how individual differences shape people's responses to robot behavior. Investigating whether certain traits, experiences, or population-level factors predict who interprets behaviors as failures could help inform adaptive and personalized robot design (see Section 6.2.1). Studying subjective failures can guide the development of robots that are sensitive to diverse user expectations.

6.2.7 Robot vs. Human Failures

In both Chapter 4 and 5, the experimental designs included direct comparisons between a robot's failures and a human's failures. Across both studies, we found significant differences (and some similarities) between how people evaluated and responded to these failures.

These findings suggest that while insights from human-human interaction (HHI) research can provide a useful starting point for robot specific design, they are not sufficient on their own. While users may project some human expectations onto robots in certain contexts, it is clear that designers cannot assume that all HHI norms will be effective at designing effective robot failure strategies. As robots increasingly occupy roles traditionally held by humans and display human-like properties and characteristics, it is essential to understand what people expect from a robot that fails, and how this differs from humans that fail. Comparing robot and human failures not only informs us about robot design, but also provides a deeper understanding of how we view ourselves and others when we fail.

Appendix A

Supplemental Analyses for Chapter 4

Agency Questionnaire

Ontological Status. Here is a person and here is a computer. Is Nao/Anne more like a person or a computer? How much is Nao/Anne like a person/computer?
 
Feelings. Does Nao/Anne have feelings, like happy and sad?
Factual Knowledge. Does Nao/Anne know the answers to a lot of questions?
Friend. Can Nao/Anne be your friend?
Think. Does Nao/Anne think for itself/herself?
Moral Knowledge. Some actions are bad, like hitting, and some actions are good, like helping. Does Nao/Anne know the difference between good and bad?
See & Hear. Can Nao/Anne see and hear the things around it/her?

Table A.1: Agency Questionnaire. For Ontological Status, the coding scheme was 0 = Computer, a lot; 1 = Computer, a little bit; 2 = In the middle; 3 = Person, a little bit; 4 = Person, a lot. For the remaining questions, the coding scheme was 0 = No; 1 = Yes, a little bit; 2 = Yes, a lot.

Study 1[1]

By Trial Comparisons

For each Trial, we ran a Logistic Model with Endorsement as the dependent variable, Partner type (Robot versus Human) and age (in years) as the independent variables. For the Trials in the Inaccuracy Phase model (Trial 6-8), we also included Response condition (Mistaken versus Apologetic versus Uncooperative) as an independent variable. In each model, we included interactions between the independent variables but removed them from the model if we did not find significant interactions.

At Trial 1, we did not find any significant main or interaction effects of Partner type and age, $ps > .378$.

At Trial 2, we did not find a significant interaction between Partner type and age, $\chi^2(1) = 3.09$, $p = .079$, so we removed it from the final model. We found a main effect of age, $\chi^2(1) = 4.06$, $p = .044$, such that older children were less likely to trust the partner at Trial 2 than younger children, $OR = 0.72$, 95% CI (0.53, 0.99). We did not find a main effect of Partner type, $\chi^2(1) = 0.49$, $p = .486$.

At Trial 3, we did not find any significant main or interaction effects of Partner type and age, $ps > .149$.

At Trial 4, we did not find a significant interaction between Partner type and age, $\chi^2(1) = 0.18$, $p = .669$, so we removed it from the final model. We found a main effect of age, $\chi^2(1) = 4.09$, $p = .043$, such that older children were less likely to trust the partner at Trial 4 than younger children, $OR = 0.67$, 95% CI (0.45, 0.99). We did not find a main effect of Partner type, $\chi^2(1) = 0.71$, $p = .401$.

At Trial 5, we did not find any significant main or interaction effects of Partner type and age, $ps > .456$.

[1]These analyses were originally published as part of the Supplementary Materials of: Teresa Flanagan, **Nicholas C. Georgiou**, Brian Scassellati, Tamar Kushnir. (2024). School-age children are more skeptical of inaccurate robots than adults. In *Cognition*, Volume 249, August 2024, 105814. [75]

At Trial 6, we found a significant three-way interaction between Partner type, Response Type, and age, $\chi^2(2) = 6.41$, $p = .041$, but no main effects of each, $ps > .442$. To explore the interaction effect further, we looked at the effect of age for each condition separately. We found that older children were less likely to trust the Mistaken Robot at Trial 6 than younger children, $OR = 0.13$, $p = .007$, 95% CI (0.03, 0.56). We did not find any differences in age in the other conditions, $ps > .119$.

At Trial 7, we did not find any significant three-way or two-way interactions, so we removed them from the final model. We found a main effect of age, $\chi^2(1) = 9.61$, $p = .002$, such that older children were less likely to trust the partner at Trial 7 than younger children, $OR = 0.58$, 95% CI (0.42, 0.82). We also found a main effect of Partner type, $\chi^2(1) = 3.95$, $p = .046$, such that children were more likely to trust the Robot partner at Trial 7 (36.9%, $SD = 0.49$) than the Human partner (23.8%, $SD = 0.43$), $OR = 0.49$, 95% CI (0.24, 0.99). Finally, we found a main effect of Response type, $\chi^2(2) = 8.23$, $p = .016$. Using Bonferroni corrections, children trusted the Uncooperative partner (16.10%, $SD = 0.37$) significantly less at Trial 7 than the Mistaken partner (37.5%, $SD = 0.49$), $OR = 0.29$, $p = .028$, 95% CI (0.10, 0.89), and the Apologetic partner (37.5%, $SD = 0.49$), $OR = 0.30$, $p = .033$, 95% CI (0.10, 0.91). Children's trust in the Mistaken and Apologetic partners at Trial 7 did not differ, $OR = 1.03$, $p = 1.00$, 95% CI (0.39, 2.69).

At Trial 8, we did not find a significant three-way interaction or two-way interactions with Response type, so we removed these interactions from the final model. We found a two-way interaction with Partner type and age, $\chi^2(1) = 3.62$, $p = .057$, and a main effect of age, $\chi^2(1) = 6.28$, $p = .012$. To explore the interaction effect further, we looked at the effect of age for each Partner type separately. For the Human partner, older children were less likely to trust the partner at Trial 8 than younger children, $OR = 0.36$, $p = .003$, 95% CI (0.18, 0.70). We did not find a significant difference in age for the Robot partner, $OR = 0.80$, $p = .392$, 95% CI (0.49, 1.33).

Agency Feature Comparisons by Response

We were interested in whether the type of experience children have with the partner influences their judgments. We ran a mixed effects model of children's agency judgment with Partner type (Human vs Robot), Response type (Mistaken, Apologetic, Uncooperative), agency item, and age as independent variables and participant as a random intercept. We did not find a significant four- or three-way interaction interactions between Partner type, Response type, and agency item and between agency item and age, so we removed these from the final model. We also removed interactions between Partner type and Response type and any with agency item, as we did not find any significant interactions. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of agency item, $\chi^2(5) = 38.35$, $p < .0001$. Follow-up comparisons found that children judged the partner to know the answer to questions ($M = 1.13$, $SD = 0.63$) less than knowing the difference between good and bad ($M = 1.49$, $SD = 0.65$), $t(981) = 5.25$, $p < .0001$, $d = 0.17$, 95% CI (0.10, 0.23), less than being a friend ($M = 1.44$, $SD = 0.69$), $t(981) = 4.48$, $p = .0001$, $d = 0.14$, 95% CI (0.08, 0.21), less than having feelings ($M = 1.47$, $SD = 0.69$), $t(981) = 4.90$, $p < .0001$, $d = 0.16$, 95% CI (0.09, 0.22), and less than seeing and hearing ($M = 1.34$, $SD = 0.70$), $t(981) = 3.05$, $p = .036$, $d = 0.10$, 95% CI (0.03, 0.16). Children's judgment of the partner knowing the answer to questions did not differ from the partner being able to think ($M = 1.31$, $SD = 0.74$), $t(981) = 2.61$, $p = .140$, $d = 0.08$, 95% CI (0.02, 0.15). There were no other significant differences between the agency items, $ps > .121$.

We also found a main effect of Response type, $\chi^2(2) = 6.29$, $p = .043$, but after Bonferroni corrections, the differences between Response types were not significant, $ps > .080$ (Mistaken: $M = 1.32$, $SD = 0.74$; Apologetic: $M = 1.46$, $SD = 0.64$; Uncooperative: $M = 1.31$, $SD = 0.69$). We did not find a main effect of Partner

type (Human: $M = 1.41$, $SD = 0.69$; Robot: $M = 1.32$, $SD = 0.70$) or age, $ps > .148$. We also found a significant interaction between Response type and age, $\chi^2(2) = 6.15$, $p = .046$, and between Partner type and age, $\chi^2(1) = 4.43$, $p = .035$, but after follow-up analyses, none of the age differences were significant, $ps > .085$.

For the ontological status question, we ran a Linear model of children's response with Partner type, Response type, and age as the independent variables. Any follow-up comparisons use Bonferroni corrections. We found a main effect of Partner type, $F(1, 153) = 40.99$, $p < .0001$, $\eta_p^2 = 0.21$, such that children were more likely to say that the human partner was more like a person ($M = 2.66$, $SD = 1.56$) than the robot partner ($M = 1.34$, $SD = 1.14$), $t(153) = 6.50$, $d = 0.53$, 95% CI (0.36, 0.69). We did not find a main effect of Response type (Mistaken: $M = 1.93$, $SD = 1.54$; Apologetic: $M = 2.04$, $SD = 1.50$; Uncooperative: $M = 2.05$, $SD = 1.53$) or age, $ps > .199$. We also found a significant three-way interaction between Partner type, Response type, and age, $F(2, 153) = 9.24$, $p = .0002$, $\eta_p^2 = 0.11$. Specifically, older children were less likely to say that the Uncooperative human partner was like a person than younger children, $\beta = -0.61$, $p = .003$. We did not find a significant difference in age for any of the other conditions, $ps > .069$.

Agency Feature Comparisons by Interaction for Robot Partner

We were interested in children's agency judgments of the robot, whether this varied by children's type of, and lack of, interactions with the robot. Specifically, we looked at the difference between children who played the game with the robot (those in the Robot Partner conditions) and children who did not play the game with the robot (those in the Human Partner conditions). To see if children's prior interactions with a robot affected their overall agency judgments of the robot, we ran a mixed effects model of children's agency judgment with Interaction (Prior Interaction for those in the Robot Partner conditions and No Interaction for those in the Human Partner

conditions), agency item (feelings, thinking, epistemic knowledge, moral knowledge, friend, and sensing), and age (in years) as independent variables and participant as a random intercept. Prior analysis did not find a significant three-way interaction between the variables or a significant interaction between Interaction and age, *ps* $>$ .176, so we removed them from the final model. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of agency item, $\chi^2(5) = 37.08$, $p < .0001$, and a main effect of age, $\chi^2(1) = 14.29$, $p = .0002$. We did not find a main effect of Interaction, $\chi^2(1) = 0.41$, $p = .524$ (Prior Interaction: $M = 1.42$, $SD = 0.71$; No Interaction: $M = 1.42$, $SD = 0.71$). We found a significant interaction between age and agency item, $\chi^2(5) = 19.62$, $p = .002$. To explore interaction effect further, we looked at the differences in age for each agency item. We found that older children were less likely to say the robot had feelings and could think than younger children (Feelings: β = -0.22, $p = .0006$; Think: β = -0.20, $p = .0006$). We did not find a significant difference in age for the other agency items, *ps* $>$ .208.

We also found a significant interaction between Interaction and agency item, $\chi^2(5) = 24.67$, $p = .0002$. To explore the interaction effect further, we compared children's judgment to each agency item between children who had Prior Interaction and children who had No Interaction. We found that children who had played the game with the robot were more likely to say that the robot has feelings ($M = 1.38$, $SD = 0.73$) than children who did not play the game with the robot ($M = 1.07$, $SD = 0.84$), $t(970) = 2.84$, $p = .028$, $d = 0.09$, 95% CI (0.03, 0.15). Furthermore, we found that children who played the game with the robot were less likely to say that the robot knows the answers to a lot of questions ($M = 1.10$, $SD = 0.66$) than children who did not play the game with the robot ($M = 1.43$, $SD = 0.65$), $t(970) = 3.03$, $p = .016$, $d = 0.10$, 95% CI (0.03, 0.16). We did not find a significant difference in Interaction for the other agency items, *ps* = 1.000.

Next, we looked at whether children's ontological status judgment (is the robot more like a person or a computer?) differed between children who had prior interactions with the robot and children who had no prior interactions. We ran a Linear model of children's responses with Interaction and age as the independent variables. We did not find a significant interaction between Interaction and age, $F(1, 160) = 0.25$, $p = .618$, $\eta_p^2 = 0.00$, so we removed it from the final model. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of age, $F(1, 161) = 5.86$, $p = .017$, $\eta_p^2 = 0.04$, such that older children were less likely to say the robot is like a person than younger children, β = -0.19. We also found a main effect of Interaction, $F(1, 161) = 6.54$, $p = .011$, $\eta_p^2 = 0.04$, such that children who had played the game with the robot rated the robot more like a person ($M = 1.34$, $SD = 1.14$) than children who did not play the game with the robot ($M = 0.91$, $SD = 1.03$), $t(161) = 2.56$, $p = .012$, $d = 0.20$, 95% CI (0.05, 0.36). This suggests that while children's ontological judgments of a robot changes with age, children who have prior interactions with a robot judge the robot as more human-like than children who had no prior interactions with the robot.

Explanation Responses

After the game, children were asked why they thought the partner started to give them the wrong answers and children's open-ended responses were categorized. For the purposes of this paper, we focused on the rate of responses that references Intention: either not intentional (e.g., "it was an accident"), intentionally harmful (e.g., "he tried to trick me"), or intentionally helpful (e.g., "she wants us to think for ourselves"). See Table A.2 for the rate of responses for each Response type. We also report the rate of responses for the other categories in Table A.3. For each of the Intention responses, we ran separate Logistic Models with response category as the independent variable and with Response condition and age as the dependent variables. Prior analyses

found no significant main or interaction effects with the Partner type, $ps > .211$, so it was not included in the final models. For referencing helpful intent or no intent, we did not find any main effects of Response type, $ps > .09$, but we did find a main effect of age for each, $ps < .008$. Older children were more likely to say that the partner had helpful intentions than younger children, $OR = 3.34$, 95% CI (1.17, 9.48), but less likely to say that the partner was not intentional than younger children, $OR = 0.38$, 95% CI (0.17, 0.86). For referencing harmful intent, we also found a main effect of age, $\chi^2(1) = 13.88$, $p = .0002$, such that older children were more likely to say that the partner had harmful intentions than younger children, $OR = 2.14$, 95% CI (1.39, 3.29). We also found a main effect of Response type, $\chi^2(2) = 8.87$, $p = .012$, such that children were more likely to reference the partner's harmful intention in the Uncooperative condition (35.3%) than the Mistaken condition (14.3%), $OR = 3.69$, $p = .044$, 95% CI (1.03, 13.26), and more than the Apologetic condition (13.5%), $OR = 3.95$, $p = .031$, 95% CI (1.10, 14.20). We did not find a significant difference between the Mistaken condition and the Apologetic condition, $OR = 1.07$, $p = 1.00$, 95% CI (0.26, 4.50).

Table A.2: Percentage of children referencing the partner's intention (either lack of, helpful, or harmful) to give the wrong answers, for each Response type collapsed across partner type. Values show percentage with count in parentheses.

Response	Not Intentional	Helpful Intentional	Harmful Intentional
Mistaken	6.1% (3)	10.2% (5)	14.3% (7)
Apologetic	7.7% (4)	1.9% (1)	13.5% (7)
Uncooperative	3.9% (2)	2.0% (1)	35.3% (18)

Table A.3: Percentage of children referencing the agent's physiology, the agent's mechanical properties, the agent's competence, the game difficulty, blaming themselves, restating the agent got the question wrong, or any other uncategorized response. Percentages are grouped by Agent type and Response type.

Partner	Response	Phys.	Mech.	Comp.	Game	Self	Wrong	Other
Robot	Mistaken	0	12% (3)	16% (4)	4% (1)	4% (1)	4% (1)	28% (7)
Robot	Apologetic	0	8% (2)	28% (7)	4% (1)	4% (1)	0	32% (8)
Robot	Uncooperative	0	4% (1)	20% (5)	4% (1)	0	8% (2)	12% (3)
Human	Mistaken	4.2% (1)	0	8.3% (2)	12.5% (3)	4.2% (1)	0	29.2% (7)
Human	Apologetic	0	0	18.5% (5)	7.4% (2)	3.7% (1)	0	33.3% (9)
Human	Uncooperative	0	0	7.7% (2)	3.8% (1)	7.7% (2)	11.5% (3)	30.8% (8)

Study 2

By Trial Comparisons

For each Trial, we ran a Logistic Model with Endorsement as the dependent variable, and Partner type (Robot versus Human) as the independent variable. For the Trials in the Inaccuracy Phase model (Trial 6-8), we also included Response condition (Mistaken versus Apologetic versus Uncooperative) as an independent variable. In each model, we included interactions between the independent variables but removed them from the model if we did not find significant interactions.

For each of the Trials 1-5, we did not find a main effect of Partner type, $ps > .207$. For each of the Trials 6-8, we did not find any significant main effects or interaction effect of Partner type and Response condition, $ps > .203$.

Agency Feature Comparisons by Response

We were interested in whether the type of experience adults have with the partner influences their judgments. We ran a mixed effects model of adults' agency judgment with Partner type (Human vs Robot), Response type (Mistaken, Apologetic, Uncooperative), and agency item as independent variables and participant as a random

intercept. We did not find a significant three-way interaction between the variables or a significant two-way interaction between Partner type and Response type, so we removed them from the final model. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of agency item, $\chi^2(5) = 47.88$, $p < .0001$. Follow-up comparisons found that adults judged the partner to have feelings ($M = 0.93$, $SD = 0.83$) and the partner to be their friend ($M = 0.90$, $SD = 0.76$) less than all other agency items (Factual Knowledge: $M = 1.12$, $SD = 0.54$; Think: $M = 1.16$, $SD = 0.84$; Moral Knowledge: $M = 1.14$, $SD = 0.81$; See and Hear: $M = 1.21$, $SD = 0.74$), $ps < .02$. Adults' judgment between feelings and being their friend did not differ between items, $p = 1.00$. There were no other significant differences between the agency items, $ps = 1.00$. We also found a main effect of Partner type, $\chi^2(1) = 189.64$, $p < .0001$, such that adults judged the human partner to have more agentic features ($M = 1.47$, $SD = 0.62$) than the robot partner ($M = 0.68$, $SD = 0.70$), $t(982) = 13.77$, $d = 0.44$, 95% CI (0.37, 0.50). We did not find a main effect of Response type, $\chi^2(2) = 0.13$, $p = .939$ (Mistaken: $M = 1.07$, $SD = 0.73$; Apologetic: $M = 1.07$, $SD = 0.81$; Uncooperative: $M = 1.09$, $SD = 0.76$).

We found a significant interaction between agency item and Partner type, $\chi^2(5) = 88.13$, $p < .0001$. Follow-up analyses found that adults judged the human partner to have more feelings ($M = 1.24$, $SD = 0.51$) and moral knowledge ($M = 1.64$, $SD = 0.55$) and be able to be their friend ($M = 1.21$, $SD = 0.70$), think ($M = 1.65$, $SD = 0.59$), see and hear ($M = 1.55$, $SD = 0.61$) more than the robot partner (Feelings: $M = 0.33$, $SD = 0.55$; Moral Knowledge: $M = 0.63$, $SD = 0.71$; Friend: $M = 0.58$, $SD = 0.68$; Think: $M = 0.67$, $SD = 0.77$; See and Hear: $M = 0.88$, $SD = 0.72$), $ps < .0001$. Adults also judged the human partner to know the answers ($M = 1.24$, $SD = 0.51$) more than the robot partner ($M = 1.01$, $SD = 0.55$), $t(982) = 2.37$, $p = .018$, $d = 0.08$, 95% CI (0.01, 0.14), but this difference between partners is less than

the difference for the other agency items, $ps < .003$.

We also found a significant interaction between agency item and Response type, $\chi^2(5) = 37.07$, $p < .0001$. Follow-up analyses found that adults judged the Uncooperative partner as knowing the answers ($M = 1.32$, $SD = 0.69$) more than the Mistaken partner ($M = 1.00$, $SD = 0.43$), $t(982) = 2.75$, $p = .018$, $d = 0.09$, 95% CI (0.03, 0.15). Adults also judged the Mistaken partner as being able to think ($M = 1.32$, $SD = 0.79$) more than the Apologetic partner ($M = 1.02$, $SD = 0.84$), $t(982) = 22.60$, $p = .029$, $d = 0.08$, 95% CI (0.02, 0.15). We did not find a significant difference between Response types for the other agency items, $ps > .066$.

For the ontological status question, we ran a Linear model of adults' response with Partner type and Response type as the independent variables. We did not find a significant interaction between the variables, so we removed it from the final model.

We found a main effect of Partner type, $F(1, 164) = 189.70$, $p < .0001$, such that adults judged the human partner as more like a person ($M = 2.76$, $SD = 1.15$) than the robot partner ($M = 0.57$, $SD = 0.90$), $t(164) = 13.77$, $d = 1.08$, 95% CI (0.88, 1.27). We did not find a main effect of Response condition, $F(2, 164) = 0.75$, $p = .476$ (Mistaken: $M = 1.59$, $SD = 1.62$; Apologetic: $M = 1.61$, $SD = 1.40$; Uncooperative: $M = 1.80$, $SD = 1.51$).

Agency Feature Comparisons by Interaction for Robot Partner

We were interested in adults' agency judgments of the robot, whether this varied by adults' type of, and lack of, experience with the robot. Specifically, we looked at the difference between adults who played the game with the robot (those in the Robot Partner conditions) and adults who did not play the game with the robot (those in the Human Partner conditions). To see if adults' prior interactions with a robot affected their overall agency judgments of the robot, we ran a mixed effects model of adults' agency judgment with Interaction (Prior Interaction for those in the Robot

Partner conditions and No Interaction for those in the Human Partner conditions) and agency item (feelings, thinking, epistemic knowledge, moral knowledge, friend, and sensing) as independent variables and participant as a random intercept. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of agency item, $\chi^2(5) = 282.95$, $p < .0001$, but we did not find a main effect of Interaction, $\chi^2(1) = 1.93$, $p = .165$. We found a significant interaction between Interaction and agency item, $\chi^2(5) = 35.34$, $p < .0001$, such that adults who had played the game with the robot were more likely to say that the robot can think ($M = 0.67$, $SD = 0.77$) than adults who did not play the game with the robot ($M = 0.25$, $SD = 0.46$), $t(994) = 4.29$, $p = .0006$, $d = 0.14$, 95% CI (0.07, 0.20). We did not find a significant difference in Interaction for the other agency items, $ps > .224$.

We then looked at whether adults' ontological status judgment (is the robot more like a person or a computer?) differed between adults' who had prior interactions with the robot and adults' who had no prior interactions. We ran a Linear model with Interaction as the independent variable. Any follow-up comparisons use Bonferroni corrections.

We found a main effect of Interaction, $F(1, 166) = 12.38$, $p = .0006$, $\eta_p^2 = 0.07$, such that adults who had played the game with the robot rated the robot more like a person ($M = 0.57$, $SD = 0.90$) than adults who did not play the game with the robot ($M = 0.18$, $SD = 0.49$), $t(166) = 3.52$, $p = .0006$, $d = 0.27$, 95% CI (0.12, 0.43).

In an exploratory analysis, we also compared adults' judgments of the robot they played with to children's judgments. We ran a general linear model for each agentic capability, comparing age group (adults, 6-7-year-old children, and 4-5-year-old children), controlling for Response type. We found that adults were less willing to say the robot has feelings, could think, know the difference between good and bad, and be their friend than both younger and older children, $ps < .013$. Adults were also less

willing to say that the robot could see and hear than younger children, $p = .008$, but not less than older children, $p = .067$. Adults and children of both age groups did not differ in saying that the robot knew the answers to questions, $ps = 1.00$. Finally, adults said the robot partner was less like a human than older and younger children, $ps < .0006$.

Explanation Responses

After the game, adults were asked why they thought the partner started to give them the wrong answers and adults' open-ended responses were categorized. The majority of adults referenced the study design in their explanations (e.g., "to measure how long I trusted a robot") in all of the conditions (32.1% - 50%). For the purposes of this paper, we focused on the rate of responses that references Intention: either not intentional (e.g., "it was an accident"), intentionally harmful (e.g., "he tried to trick me"), intentionally helpful (e.g., "she wants us to think for ourselves"), or neutrally intentional (e.g., "he wanted me to stop trusting him"). See Table A.4 for the rate of responses for each Response condition. We also report the rate of responses for the other categories in Table A.5. For each of the Intention responses, we ran separate Logistic Models with response category as the independent variable and with Partner type and Response type as the dependent variables. Prior analyses found no significant interaction effects, so it was not included in the final models.

No adults referenced the partner's lack of intention in the Robot Partner conditions or in the Uncooperative Response conditions. Due to this lack of variation, we could not run main effects of Partner type or Response type. Instead, we simply compared adults' reference to lack of intention in the Human Mistaken (25%) and the Human Apologetic (21.4%) conditions to 0. We found adults reference to lack of intention in these conditions were significantly different from 0, $ps = .004$, suggesting that adults were more likely to reference the partner's lack of intention in the Human Mistaken

and Human Apologetic conditions than the Human Uncooperative condition and more than any of the Robot conditions. For referencing helpful intent or neutral intent, we did not find any main effects of Partner type or Response type, $ps > .083$.

For referencing harmful intent, we found a main effect of Partner type, $\chi^2(1) = 5.31$, $p = .021$, such that adults were more likely to say that the human partner had harmful intentions (17.9%) than the robot partner (7.1%), $OR = 3.37$, 95% CI (1.14, 9.95). We also found a main effect of Response type, $\chi^2(2) = 24.10$, $p < .0001$, such that adults were more likely to reference the partner's harmful intention in the Uncooperative condition (30.4%) than the Mistaken condition (3.6%) and the Apologetic condition (3.6%), $ORs = 12.87$, 95% CIs (1.95, 84.87), $ps = .004$. We did not find a significant difference between the Mistaken condition and the Apologetic condition, $OR = 1.00$, 95% CI (0,09, 11.60), $p = 1.00$.

Table A.4: Percentage of adult referencing the agent's intention (either lack of, neutral, helpful, or harmful) to give the wrong answers, for each Agent type and Response type.

Partner	Response	Not Intent.	Neutral	Helpful	Harmful
Robot	Mistaken	0	3.6% (1)	0	0
Robot	Apologetic	0	3.6% (1)	3.6% (1)	0
Robot	Uncooperative	0	10.7% (3)	7.1% (2)	21.4% (6)
Human	Mistaken	25% (7)	7.1% (2)	0	7.1% (2)
Human	Apologetic	21.4% (6)	3.6% (1)	3.6% (1)	7.1% (2)
Human	Uncooperative	0	0	3.6% (1)	39.3% (11)

Table A.5: Percentage of adults referencing the study design, the agent's mechanical properties, the agent's competence, the game difficulty, blaming themselves, restating the agent got the question wrong, or any other uncategorized response. Percentages are grouped by Agent type and Response type.

Partner	Response	Study	Mech.	Comp.	Game	Self	Wrong	Other
Robot	Mistaken	39.3% (11)	35.7% (10)	21.4% (5)	0	0	3.6% (1)	14.3% (4)
Robot	Apologetic	42.9% (12)	32.1% (9)	21.4% (6)	3.6% (1)	0	0	3.6% (1)
Robot	Uncooperative	32.1% (9)	32.1% (9)	0	0	3.6% (1)	0	3.6% (1)
Human	Mistaken	50% (14)	3.6% (1)	25% (7)	0	0	0	3.6% (1)
Human	Apologetic	46.4% (13)	3.6% (1)	28.6% (8)	0	0	0	7.1% (2)
Human	Uncooperative	46.4% (13)	3.6% (1)	0	0	0	0	7.1% (2)

Appendix B

Is Someone There Or Is That The TV? Detecting Social Presence Using Sound

As robots increasingly occupy user-facing roles in everyday environments, new technical challenges arise that can lead to frequent failures if left unaddressed. While this dissertation focuses on investigating how users perceive and respond to robot failures, developing technical tools and algorithms to help prevent and mitigate failures is similarly important. In home settings, for instance, context-aware tools can assist social robots in accurately interpreting the social context of their environment and aiding them in their decision making. Without this capability, robots may behave in ways that users perceive as inappropriate (e.g., interrupting at times when interaction is undesired).

This chapter[1] focuses on the classification between audio that includes a) *natural* conversation that includes at least one co-located user and b) *media* that is playing

[1]Portions of this chapter were published as: **Nicholas C. Georgiou**, Rebecca Ramnauth, Emmanuel Adniran, Michael Lee, Lila Selin, and Brian Scassellati. (2023). Is Someone There or Is That the TV? Detecting Social Presence Using Sound. In *ACM Transactions on Human-Robot Interaction (THRI)*, Volume 12, Issue 4, Article No. 47, Pages 1-33. [85]

from electronic sources and does not require a social response, such as television shows. This classification can help social robots detect a user's social presence using sound. Social robots that are able to solve this problem can apply this information to assist them in making decisions, such as determining when and how to appropriately engage human users. We compiled a dataset from a variety of acoustic environments which contained either *natural* or *media* audio, including audio that we recorded in our own homes. Using this dataset, we performed an experimental evaluation on a range of traditional machine learning classifiers, and assessed the classifiers' abilities to generalize to new recordings, acoustic conditions, and environments. We conclude that a C-Support Vector Classification (SVC) algorithm outperformed other classifiers. Finally, we present a classification pipeline that in-home robots can utilize, and discuss the timing and size of the trained classifiers, as well as privacy and ethics considerations.

B.1 Introduction

Imagine you are walking around the house when you stumble upon a door that is slightly ajar—opened just enough so that you can hear, but not see, what is going on inside. Opening the door to see if it is appropriate or not to enter is self-defeating. If you do not hear anything, it is very difficult to make any judgments. Suppose, however, that you hear human speech from behind the door. This piece of information can give you insight and can help you in your decision-making.

However, knowing that there is human speech is not enough. Many lower-level characteristics, as well as higher-level conceptual components of this speech, might be important factors in your decision. Do you recognize the voices? Does the speech sound serious or is it more lighthearted? Is there shouting or is the tone normal? What emotions can you detect from the speech? How many people can you hear? If

you hear two friendly sounding people having a chat, you might be more inclined to knock. If you stop by to relay a message, and you hear yelling coming from the room, it is probably best to steer clear for now. But, imagine that the yelling is from an enthusiastic sportscaster describing a sporting event or that the serious tone that you hear is from a dramatic soap opera. You might make a different decision if you know that the speech is coming from a television show rather than from physically present people in the room conversing. This is an important component of the speech that will influence your understanding of the situation and can affect how you interact, if you do.

Similarly, a social robot that is designed to interact with users in realistic and appropriate ways should have the ability to make this disambiguation. The robot can benefit from knowing whether the speech coming from behind the door is from a physically present human socializing. More generally, knowing when speech is a product of at least one co-located person conversing, or not, can assist social robots in making inferences about users' activities and can help them accommodate their users through a better understanding of their environments. This chapter focuses on whether there is (1) *natural* conversation occurring that includes at least one co-located user or (2) *media* playing from electronic sources that does not require a social response. These are common speech scenarios in the home which can assist the robot in detecting the social presence of a user through what the robot hears.

In practice, we imagine countless settings where the ability to make such a classification could be utilized by robots to assist them in accomplishing their goals. For example, a social companion robot in the home may decide to engage a co-located user with a supportive, social interaction if it infers that the user is upset, as opposed to if it knows the speech is *media*. A robot assisting people with Autism Spectrum Disorder may not interrupt when a user is engaged in *natural* conversation (to encourage social interaction), but may attempt to engage if it suspects the user is watching

too much *media*. A customer service robot may decide whether or not to head in the direction of customers chatting in a store or may choose to disregard the speech if it is coming from a TV. An in-home robot may reach out for external assistance if a user is distressed, but may not if it realizes the speech is from an action movie on TV. Depending on the end goals of the system, the robot can use such a classification, along with other prudent factors, to help it in making decisions.

To precisely characterize the differences between audio from *natural* and *media* scenarios is a challenge. Both of these audio categories contain human voices. Both categories contain diverse audio with similarities that make it difficult to quantify how we, as humans, usually know which of the two we are listening to. One potential discriminatory criterion, for example, is the speech patterns in the scripted conversation of television shows as opposed to the more spontaneous nature of impromptu conversation. This could be sufficient for categorizing a sitcom as *media*, but this does not help us in correctly classifying a radio podcast where the host is casually interviewing a guest. One could also try to make this classification based on if they hear cleanly engineered audio, like that produced in a studio, versus the noisy, distorted *natural* audio environments of everyday life. This can help with correctly classifying a TV show or movie played on a good sound system as *media*, but will not help when listening to sports, which involve crowd and audience noise. Solely detecting the presence of electronically sourced audio (i.e., coming from the speakers of a computer or television) is also not enough. Video calls with friends are *natural* situations in which there is electronic-sourced audio, along with at least one organically-sourced (i.e., coming directly from human vocal cords) speaker playing an active role in the conversation. If we know that some part of the audio is organically sourced, we can be sure that there is a co-located, physically present person talking. But, it can sometimes be tough to know if this is the case, especially if electronic audio sounds *natural* (e.g., conversational) and is played on a high-quality sound system. Making

the classification between audio that is *natural* or *media* is hard.

For this chapter, we focus on being able to classify between *natural* and *media* audio from the dynamic environment of the home. We focus on differentiating between speech from popular genres of *media* that is originating from loudspeakers and speech from *natural* conversations including at least one co-located person in the home. Ideally, robots in real-world environments would have the ability to make this classification, regardless of the acoustic environments they are in (e.g., different rooms, different loudspeakers, distances from the audio source) and the different audio content that they hear (e.g., different voices, different TV/radio shows, background noise). Social roboticists that deploy robots in the home and intend to use audio to make decisions on how their robots interact with users can benefit from this work.

Our main contributions are:

- Describing a salient audio problem that social robots in the home face: the classification between a) *natural* conversation including at least one co-located user and b) *media* playing from electronic sources that does not require a social response

- Training classifiers[2] that use in-home audio to differentiate between *natural* and *media*, and evaluating how well the classifiers generalize to new recordings, acoustic conditions, and environments

- Proposing a classification pipeline that can provide additional, situational context to a social robot by assisting it in detecting social presence using sound

The organization of the chapter is as follows: Section B.2 offers background and related work. Section B.3 describes the methodology in collecting the dataset, in selecting and extracting features of the audio, and in selecting the classification algorithms. Section B.4 describes the experiments used to test the generalizability of the

[2]A link to our trained models and the code used to create the input feature vectors for our models: https://github.com/ScazLab/social-presence-sound

classifiers, and discusses the results. Section B.5 discusses how these classifiers can be applied in practice, with details on timing and size of each, a proposed classification pipeline, and a discussion on ethics and privacy considerations. Section B.6 discusses some limitations of the work and Section B.7 concludes the work.

B.2 Background

According to a recent U.S. Bureau of Labor Statistics survey [264], watching television was the most popular and time-consuming leisure activity in an American's average day, with people spending close to three hours watching TV. In comparison, activities such as eating, drinking, socializing, and communicating amount to approximately two hours total a day. These everyday domestic situations involve humans engaging with *media* (e.g., watching television) or *natural* situations (e.g., participating in a conversation at the dinner table).

B.2.1 In-Home Virtual Assistants

Popular virtual assistants, such as Amazon's Alexa, have already been integrated into many homes around the U.S. They use audio-based techniques that make them effective in the household. Source localization approximates the origin of audio input and wake-word detection [142] prompts sending the speech command to the cloud for natural language processing [145]. These features inform the assistant's decision-making policy to effectively and appropriately respond [204, 177]. These in-home systems do not incorporate much, if any, contextual awareness of their surroundings [214]. In fact, these systems typically require specific and explicit user prompts to engage them (e.g., "Alexa"). Because these systems are user-initiated, the detection of social context is much less necessary. Yet, for systems designed to interact with users autonomously, the ability to garner context about the environment is crucial

[174].

We believe that virtual assistants can also benefit from the ideas presented in this chapter, especially if developers believe there is value in additional functionality that includes behaving more socially and independently. Although we will focus on social robots in this chapter, we note that social presence through sound can be of use to any device in the home that could utilize such context to help it make decisions.

B.2.2 Using Audio for Activity and Event Detection in the Home

Automatic recognition of user activity in dynamic, unstructured environments, like the home, is important for systems whose primary purpose is to support their users through social means. Having some understanding of a user's activity and social context can help the system in its decision making.

Audio scene classification (ASC), or the identification of the environment or activity based on acoustic signals, is important for robotics and can help better facilitate human-robot interaction [16]. ASC has become a trending topic with growing interest because of the advent of smart homes and robots [65, 283, 265]. In recent years, audio analysis capabilities have been added to assistive robotic systems, such as the TIAGo service robot [93] and RiSH, a robot-integrated smart home for elderly care [63], with the goal that audio will provide more contextual awareness. Work for audio analysis in the home includes activity detection specific to helping the elderly by detecting falls [210] or by identifying common activities, to help medical staff monitor people who utilize ambient assisted living services [6, 206, 60]. Audio scene classification has also been used in the context of differentiating between specific kitchen sounds like the mixer, dishwasher, and utensils clanking [265], bathroom sounds like showering, washing hands, and flushing [47], breathing or snoring [72], or common sounds including keyboard typing, applause, and phone ringing [255]. Traditional machine

learning classifiers have been used for these classifications with success.

Work has also been done that involves classifying in-home audio with the help of humans-in-the-loop. Some of this work includes human-assisted sound event recognition for home service robots for the elderly, where a human caregiver helps provide a robot with in-the-loop labels to non-voice sounds, in order to help a robot actively learn auditory events [62]. Additional work has used audio to classify different rooms in the home, like the kitchen and office, and also discriminated between nonverbal sounds like clapping and one-word speech scenarios [178].

The research area of voice activity detection (VAD) looks to classify between audio that contains speech and non-speech [99]. Research has been done to use noise cancellation to better implement VAD on smart home devices [108]. Other VAD work includes enhanced speech detection for humanoid robots in sparse dialogue [125] and robust classification between speech and non-speech [212] in noisy environments. Work has been done to recognize emotional states from speech using a support vector machine [242], to separate speech from music [4], and to detect and classify noises in speech signals [186].

There has also been research looking into how to accurately discriminate between speech commands produced from an electronic speaker from organic human speech [29]. This approach was discussed in the context of cybersecurity to better identify replay attacks of certain commands on Internet of Things devices, by focusing on determining the origin of pre-written speech commands, but does not focus on in-home, noisy experimentation.

Our work presents a new tool that can be used by robots in the home to gather more social context about a user's social presence through sound, when presented with human speech. The classification between *natural* and *media* that we focus on in this work encapsulates common speech scenarios in the home, that can give insight into people's activities. Our experimentation focuses on real-world audio recorded

in noisy, in-home environments, and this work adds to the research area of activity detection in a dynamic environment.

B.2.3 Audio Classification of Media

Work has also been done in the classification of different forms of *media*. Audio information has also been utilized when researching genre classification in different forms of media. Music information retrieval methods have explored classifying songs into genres such as pop, rock, or blues [263, 22] and television *media* classification has classified videos into genres such as cartoons, news, or weather forecasts [66]. A key aspect of many of these media approaches, along with the in-home activity detection of Section B.2.2, involves extracting time and frequency domain features (e.g., spectral contrasts, spectral roll-offs, Mel-Frequency Cepstral Coefficients, or chroma features) from the overall audio signal and using these features to inform and train machine-learning classification algorithms. We build on this work by using similar features in our analysis, and discuss more background and motivation of the feature selection in Section B.3.2.

B.3 Methodology

In this section, we describe how we (a) compiled an audio dataset containing the *natural* and *media* classes, (b) extracted features from each audio sample, and (c) selected the machine learning classifiers that we experimented with. We define two terms that we will be using throughout this chapter. First, when discussing a **sample**, we are referring to a 5-second segment of audio that has been recorded and is used in feature extraction. A **recording** is a collection of contiguously captured *samples* during a given time window.

B.3.1 Audio Sample Collection

We collected audio content from various television genres and radio shows (sound from electronic speakers) and human speakers (sound from human voices). The final dataset contained approximately 30 hours of audio recordings, and was well-balanced between the *media* and *natural* classes.

Both categories were recorded on Kinect One microphones. This was important because any decisions made by a machine learning classifier would be able to focus on the difference of the audio content, rather than discrepancies caused by different recording hardware.

Media Recording Set

Our *media* (M) recording set consisted of a variety of TV shows or radio recordings, that we recorded on the Kinect One[3]. We focused on collecting audio recordings from popular television genres, which include drama, comedy, participatory/reality, news, and sports [278], as well as audio from radio shows. This category was recorded in different rooms, using a variety of electronic speakers[4], with the microphone capturing audio at varying distances from the speakers, during different contiguous time windows. Recording during different time windows allowed for different background and ambient noise to be captured as a part of the various recordings. All audio recordings were recorded at a rate of 16 kilohertz (kHz) in the waveform audio file format (.wav).

Each room, speaker, and microphone position configuration is referred to as its own unique *label*. These different recording configurations emulate a variety of recording conditions that an in-home agent might face. The distribution of the audio in each

[3]We recorded the media recordings being emitted through electronic speakers, instead of inputting the media audio file directly into the classifier, because this is how a robot in the home would be capturing the media audio.

[4]The specific speaker models are as follows: Bose SoundLink 359037-1300 Mini Bluetooth Speaker (Bose), MacBook Pro 13" (Mac), iPhone 11 Pro (iPhone), Bose Wave Music System II (BigBose), 40" Eco Bravia VE5 Series LCD HDTV (SonyTV)

Table B.1: Media Data Set Composition

Label	Room	Speaker	Kinect Distance	Total Samples	Recordings
A	Bedroom	Bose	1 ft	617	6
B	Bedroom	Bose	9 ft	852	7
C	Bedroom	Mac	6 ft	2075	11
D	Playroom	Bose	1 ft	627	5
E	Playroom	Bose	5 ft	517	9
F	Playroom	Bose	10 ft	332	3
G	Playroom	iPhone	1 ft	423	3
H	Playroom	iPhone	4 ft	372	2
I	Kitchen	Bose	9 ft	543	2
J	Kitchen	BigBose	9 ft	1185	5
K	Kitchen	SonyTV	4 ft	1432	2
L	Kitchen	iPhone	6 ft	201	1
M	Kitchen	Mac	6 ft	551	3
N	Kitchen	Mac	1 ft	411	1

label can be seen in Table B.1. There are 60 *media* recordings in our dataset, with a total of 10,138 samples, for around 14 hours of audio. Depending on the experiment that we performed, a different split of the recordings in the *media* set was used as training and testing data (explained in more detail in Section B.4).

Natural Recording Set

The *natural* recording set can be broken down into three categories: CHiME5 (C), Video Calls (V), and Family Conversations (F).

Natural Audio from CHiME5. Category C recordings were comprised of content from the CHiME-5 dataset [20], available online. CHiME-5 contains audio captured from dinner parties in different houses. Each dinner party involved a different group of four people, who were told to engage in natural conversation in the house's kitchen, dining room, and living room for at least 2 hours.

Category C contained audio from 10 different CHiME-5 sessions. Each session contained audio from six Kinect microphone arrays, placed in different locations (bed-

room, kitchen, living room) in each home, with audio input from each channel of each microphone. We used audio from the different Kinect microphones within the same dinner party in our dataset because we wanted a diverse set of audio captured from different locations with varying acoustic properties. For the C category, we considered a *recording* to be all of the audio collected from a unique CHiME-5 session. The CHiME-5 audio files were in the waveform audio file format (.wav), with a recording rate of 16kHz. We chose CHiME-5 because it captured *natural*, social scenarios that one can expect to find in a home environment. We input the CHiME-5 files directly into the classifier because this is how *natural* audio would be captured by the robot. In total, category C contained 10,130 samples (1013 samples per recording). This sample number is equivalent to approximately 1.4 hours per CHiME-5 session, for a total of almost 14 hours of audio. Samples from the C category were used as our *natural* training data.

Natural Audio from Our Home Environments. We also captured *natural* audio from our own homes. We had Institutional Review Board approval to record audio in homes and to extract and analyze acoustic features. There were two categories that we experimented with, involving *natural* scenarios from 6 rooms in 3 different homes. We left a recording microphone in locations that we deemed appropriate for an in-home robot or device to be placed, recorded audio, and later inspected the audio. Audio from these two categories was used as our *natural* testing data.

Category V captured audio from video calls taking place in a home's office, dining room, and living room. These recordings involved conversations between members of a family consisting of two children and three adults. Members of the family congregated in their dining room and spoke over a video call on a laptop and phone using Zoom or Facebook Messenger. The calls were all on speaker. As a result, voices were variably distant from the microphone and the recordings captured by the Kinect included a mixture of voices coming from an organic source (the person in the same room as the

Kinect microphone) and from electronic sources (the people on the video call). The same person was physically in the room with the Kinect for each of these recordings. Category V included six separate recordings, with a total of 917 samples.

Category F consisted of audio collected from family conversations in kitchens and living rooms, in three different homes. The microphone was placed close to where people were dining and conversing. An example location for the microphone was on a counter in an open, spacious kitchen. The kitchen recordings included some background noises such as the running sink, clanking utensils, and plates and glasses moving, while the living room recordings happened with little to no noise in the background. Category F included 965 samples and seven separate recordings, including voices from 11 different people.

There are multiple reasons that we decided to also collect *natural* audio that we recorded ourselves, despite having an extensive corpus of in-home, *natural* audio from CHiME5. Even though we tried to collect our *media* sample set with similar recording characteristics (i.e., microphone and sampling frequency) to CHiME5, we wanted to see whether or not classifiers trained solely on CHiME-5 could generalize to classifying other *natural* audio from outside of that corpus. This could show that these classifiers are able to correctly disambiguate between *natural* and *media* recorded by us, and that that the classification is not just a result of some discrepancies in how CHiME-5 was collected and how we recorded our audio. Lastly, we wanted to be able to experiment with the case of social presence that includes a mixture of electronic audio and organic-sourced *natural* audio, captured in the V dataset. This circumstance indicates social presence because at least one user that is co-located with the robot is engaged in a natural conversation, while chatting on a call with others. Samples from the V and F categories were used as our *natural* testing data.

B.3.2 Feature Extraction

We split our entire audio dataset into 5-second samples. From each sample, we extracted features to create an input vector that was used to train machine learning classifiers. We used the LibRosa Python package [180] to extract audio features. These are commonly used features in audio analysis (as mentioned in SectionB.2.3), which was the motivation for using them.

In total, 83 features were extracted from each audio sample. We performed a standard transformation of each feature to normalize the feature set. The input vector contained the features below for each audio sample:

- *Mel-frequency cepstral coefficients (MFCCs)*: These are dominant features that have been historically used in speech recognition and they have been explored in separating music and speech [166]. It is typical that 13 coefficients are used for speech representation [263], so we use the means and standard deviations for each of the first 13 coefficients over the sample, for a total of 26 features.

- *Chroma Energy Normalized Statistics (CENS)*: These are features that have been used in audio analysis research to match similar audio [192]. There are 12 chroma classes and we use the mean and standard deviation for each chroma class over the sample, for a total of 24 features.

- *Root-mean-square (RMS) energy values*: Energy features are commonly used in audio analysis, with some prior work finding that the combination of energy with MFCC is better than using MFCCs alone [109]. We use the range, standard deviation, and skewness of this feature, for a total of 3 features.

- *Zero-crossing rates*: These are features that are commonly used in audio analysis [66] and can help provide a measure of noisiness of the audio sample [263]. We use the mean, standard deviation, and skewness, for a total of 3 features.

- *Tempo*: This feature estimates the beats per minute in the audio sample. The motivation behind adding this is that music from TV or radio commercials typically have more tempo than conversational audio in the home. This is 1 feature.

- *Spectral centroid, flatness, rolloff, and bandwidth*: These are also commonly used low-level components of the audio signal [53, 263]. We use the mean, standard deviation, and skewness for each, for a total of 12 features.

- *Spectral contrast*: These are features that have been shown to discriminate among different music genres [109], so we use the means and standard deviations for seven sub-bands, for a total of 14 features.

Note that none of these features involve transcription or semantic representation of dialogue/words in the audio environment. This way, the audio is translated into a machine readable format that has little to no meaning to a human, as opposed to words, which are used in lexical analysis in Natural Language Processing. This is an arguably less invasive and more privacy-sensitive approach than using words, especially if the robot is intending on sending the input vector to the cloud to be analyzed.

B.3.3 Classification Algorithms

In our experiments to determine if our classification problem can be solved, we trained and tested different models with six traditional machine learning classification algorithms, using the sci-kit learn Python library [209]. These are commonly used algorithms for audio classification tasks (see Section B.2 for more details). We performed an experimental evaluation of various approaches, to see which classifiers would be best suited to tackle the problem. We experimented with the following algorithms:

- KNeighborsClassifier [74]

- DecisionTreeClassifier [35]
- QDA (Quadratic Discriminant Analysis) [103]
- LogisticRegression [291]
- GaussianNB (Gaussian Naive Bayes) [290]
- SVC (C-Support Vector Classification) [69, 45]

We use these traditional classifiers instead of deep learning techniques, which have gained popularity in recent years in the audio analysis space, for multiple reasons. First, our dataset is modestly sized, and traditional ML algorithms have a much better chance at performing successfully than deep learning when the dataset is not very large. Second, we know the feature space that we want to use for this classification task. Lastly, we are hoping to be able to use these trained classifiers on real-time systems, so the response time needs to be quick and the complexity and space taken by the classifier needs to be reasonable (many social robots have limited compute power).

A gridsearch on each classification algorithm measured what hyperparameter combination was the best for each algorithm on our first experiment (described in Section B.4). The different hyperparameter combinations for each classifier that were experimented with can be found in Section B.4. The hyperparameters that led to the highest performance, and were subsequently selected for the classifier in all of the following tests can be seen in Section B.5.

B.4 Experiments and Results

In this section, we describe how the various classifiers performed on experiments that tested the classifiers' abilities to generalize to novel recordings, environments, and

conditions. We test how well classifiers perform on a leave-one-recording-out cross validation, where we test on recordings that were left out of the training set. We also test how well the classifiers generalize to classifying *natural* recordings from outside of the training corpus and to *media* recordings from (1) rooms, (2) speakers, (3) microphone positions, and (4) combinations of all three, that they were not trained on.

B.4.1 Leave-One-Recording-Out Cross Validation

We performed an evaluation similar to a leave-one-out cross-validation (LOOCV), but in our case, leave-one-recording-out cross-validation **(LOROCV)**[5]. To perform LOROCV, we trained models using *natural* recordings from our CHiME-5 category (C) and *media* recordings from our media (M) recording set. For each fold of LOROCV, we trained on all recordings except for one from C and one from M. We did this for all possible pairs of recordings from C and M, which resulted in 600 folds (the Cartesian product of the 10 recordings in C and the 60 recordings in M). For each fold, we tested our classifier on the 1) left-out {C,M} recording pair, 2) left-out M recording and *natural* audio sampled from V, 3) left-out M recording and *natural* audio sampled from F, and 4) left-out M recording and *natural* audio sampled from both F and V. Because recordings can be of different lengths, we randomly sampled from the larger recording to match the size of the smaller recording. This ensured that we had balanced test sets each time.

The metrics that we recorded for all of our experiments are below. TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative.

[5]A conventional splitting of all of the samples into a train, test, and validation set would not be very insightful because many of our data samples were part of the same contiguously recorded audio clips (recordings). For any given recording in our dataset, there were at least 15 samples that were a part of the same original audio recording. When randomly shuffling the dataset for the train/test/validation splits, it is likely that some of a recording's 5-second samples land in each of the folds and the test and validation sets. Since audio from the same recording is inherently similar, we performed a cross validation per recording.

- Accuracy=(TP+TN)/(TP+TN+FN+FP)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1 Score = (2*Precision*Recall)/(Precision+Recall)

We recorded the precision, recall, and F1 scores for both the *media* and the *natural* classes (i.e., we treated both as the positive class). Both the macro averages (arithmetic mean) and micro averages (weighted average) were recorded across all folds. The full results for LOROCV can be found in Table B.17 in Section B.10, with a summary in Table B.6 in Section B.9.

With LOROCV, we test on *natural* audio from left-out CHiME-5 sessions (new voices and rooms from new homes within the CHiME-5 corpus), or better yet, on *natural* audio from the V or F categories that we recorded ourselves. We also test on unseen *media* recordings that the classifiers have not trained on and that we have recorded ourselves. This provides insight into how the trained algorithms can generalize to classifying novel recordings of *media* and *natural* audio.

B.4.2 Leave Out Rooms, Speakers, and Microphone Positions in the *Media* Set

We can gain further insight into how robustly the classifiers can differentiate between *natural* and *media* audio, if *media* in the training set contains recordings from different acoustic conditions (e.g., rooms, loudspeakers, microphone distances) than *media* in the testing set. In the experiments in this section, we evaluate how our classifiers perform when toggling which condition(s) of the *media* recording set to leave out of the training set. We also use the *natural* audio from the C category to train our models. We test on the *natural* V and F categories that we recorded ourselves and

on the left-out *media*.

We left all of the *media* samples of a specific (1) room, (2) speaker, (3) microphone position, or (4) combinations of the three, out of the training set, and tested on the left out *media* samples and on *natural* samples from the V and F test categories. We matched the number of *media* samples in the training set with an equally distributed, random sample of 5-second samples from each *natural* recording in category C. We randomly sampled from all of the recordings in the larger test subset to match the size of the smaller subset. This ensured that we had balanced test sets each time. We recorded the micro and macro averages of precision, recall, and F1 scores for both the *media* and *natural* classes, as in LOROCV. The following paragraphs describe each experiment that we performed.

In Leave One Label Out (**LOLO**), we wanted to see how well classifiers would perform when they trained on *media* from specific *labels*, or specific room, speaker, and Kinect distance configurations (see Table B.1), along with *natural* from category C, and then were tested against configurations that they were not trained on. We performed a Leave One Label Out (LOLO) experiment on all labels of our *media* data, where we trained different models using all the recordings from all combinations of labels, and tested against the held out labels. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.16 of Section B.10, with a summary in Table B.7 of Section B.9.

In Leave One Room Out (**LORO**), we wanted to see how well classifiers would perform when they trained on *media* from specific rooms, along with *natural* from category C, and then were tested against *media* from a room they had not trained on. This is important because each room has a different acoustic environment and layout. The classifiers should be able to make accurate predictions regardless of if they have trained on audio from the room in which they are deployed. In LORO,

classifiers test on *media* recordings from a room that they have not trained on, but the test set includes loudspeakers and microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.19 of Section B.10, with a summary in Table B.10 of Section B.9.

In Leave One Speaker Out (**LOSO**), we wanted to see how well classifiers would perform when they trained on *media* from specific loudspeakers, along with *natural* from category C, and then were tested against *media* from loudspeakers they had not trained on. This is important because each loudspeaker has different hardware properties and the classifiers should be able to make accurate predictions regardless of if they have trained on audio from the loudspeaker from which they hear audio. In LOSO, classifiers test on *media* recordings from a loudspeaker that they have not trained on, but the test set includes rooms and microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.18 of Section B.10, with a summary in Table B.9 of Section B.9.

In Leave One Distance Out (**LODO**), we wanted to see how well classifiers would perform when they trained on *media* from certain microphone distances from a loudspeaker, along with *natural* from category C, and then were tested against *media* from microphone distances they had not trained on. This is important because the robot might be at variable distances from the sound source. In LODO, classifiers test on *media* recordings from a microphone distance that they have not trained on, but the test set includes loudspeakers and rooms that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.19 of Section B.10, with a summary in Table B.10 of Section B.9.

In Leave One Room and Speaker Out (**LORSO**), we wanted to see how well

classifiers would perform when they were tested on *media* rooms and speakers that they had not trained on. This is a more robust test than the previous ones. In LORSO, classifiers test on *media* recordings from a room and speaker that they have not trained on, but the test set includes microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.20 of Section B.10, with a summary in Table B.11 of Section B.9.

In Leave One Room and Distance Out (**LORDO**), we wanted to see how well classifiers would perform when they were tested on *media* rooms and microphone distances that they had not trained on. In LORDO, classifiers test on *media* recordings from a room and microphone distances that they have not trained on, but the test set includes microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.21 of Section B.10, with a summary in Table B.12 of Section B.9.

In Leave One Speaker and Distance Out (**LOSDO**), we wanted to see how well classifiers would perform when they were tested on *media* speakers and microphone distances that they had not trained on. In LOSDO, classifiers test on *media* recordings from a loudspeaker and microphone distances that they have not trained on, but the test set includes rooms that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.22 of Section B.10, with a summary in Table B.13 of Section B.9.

In Leave One Room, Speaker, and Distance Out (**LORSDO**), we wanted to see how well classifiers would perform when they were tested on *media* speakers, rooms, and microphone distances that they had not trained on. This is the most challenging test that we perform for the classifier. In LORSDO, classifiers test on *media* recordings

from a room, loudspeaker, and microphone distance that they have not trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table B.23 of Section B.10, with a summary in Table B.14 of Section B.9.

Table B.2: Experiment Summary. The table shows the average of the macro average F1 scores ($(F_{natural} + F_{media})/2$) for each classifier across all folds of each experiment. The table shows the average results of the trained classifiers being tested on the left out *media* sets along with *natural* recordings from the V and F categories. The classifier with the best average performance on each test set and experiment is in bold. More comprehensive results can be found in Sections B.9 and B.10.

Experiment	Test Set	KNN	QDA	DT	GNB	LR	SVC
LOROCV	V+M	**94.5**	89.0	87.9	86.0	87.9	91.3
	F+M	87.1	**99.5**	98.9	96.2	98.9	96.6
	F+V+M	91.3	93.3	92.8	90.5	92.8	**93.7**
LOLO	V+M	78.5	**99.3**	96.0	90.4	91.7	93.1
	F+M	85.6	88.8	77.9	82.4	82.9	**90.4**
	F+V+M	82.5	**93.4**	85.8	85.5	86.4	91.7
LORO	V+M	77.4	**99.2**	94.3	81.0	85.0	94.9
	F+M	83.1	86.1	75.1	83.5	82.7	**86.5**
	F+V+M	80.4	**92.6**	84.8	82.9	84.5	90.6
LOSO	V+M	76.9	98.5	**99.0**	93.8	97.1	93.4
	F+M	83.6	84.4	75.2	80.7	84.3	**87.5**
	F+V+M	80.9	**91.0**	86.6	87.0	90.4	90.3
LODO	V+M	78.9	**98.8**	97.3	91.6	97.8	94.8
	F+M	74.3	83.6	63.9	71.6	81.6	**84.1**
	F+V+M	77.7	**91.0**	81.4	81.8	89.6	89.4
LORSO	V+M	67.9	86.6	**87.9**	70.1	85.6	86.3
	F+M	82.1	82.9	78.1	76.7	87.2	**88.1**
	F+V+M	76.1	85.1	83.1	74.9	86.9	**87.6**
LORDO	V+M	70.0	92.5	**95.3**	78.5	88.8	87.4
	F+M	78.3	86.7	77.9	77.1	89.4	**90.5**
	F+V+M	75.0	89.4	85.8	78.0	**89.6**	89.5
LOSDO	V+M	76.3	90.9	90.1	85.2	**95.3**	94.5
	F+M	79.1	80.8	76.6	78.3	83.3	**84.8**
	F+V+M	77.9	85.5	82.8	81.4	89.0	**89.5**
LORSDO	V+M	65.8	82.4	**87.2**	70.0	86.5	85.3
	F+M	73.0	77.1	72.7	66.7	83.7	**85.2**
	F+V+M	69.9	79.5	79.6	68.2	85.0	**85.3**

B.4.3 Selecting a Classifier

In general, we see that most of the trained classification algorithms perform well on our experiments. We see that most of the classifiers have average F1 scores in the 90s or 80s for a majority of the experiments. Table B.2 summarizes the results for all our experiments, for each classifier.

Results

We see that SVC has the best performance on the most tests throughout our experiments. SVC has the highest average F1 score on 12 out of the 27 tests, with the highest average F1 score on 7 out of the 12 more difficult tests (where two or three of the *media* parameters are left out of the test set in LORSO, LORDO, LOSDO, and LORSDO). SVC has the highest performance on the F+V+M test sets on all but one of the more difficult experiments, and SVC has the highest F1 score on the F+M test sets for almost all of the experiments. On LORSDO, the most difficult experiment, SVC has the best performance on two out of three of the tests (V+M and F+M). Despite not having the highest scores on V+M, it does consistently well on the test set, throughout all of the experiments. Generally, SVC is the most consistent classifier across the different test sets and experiments, and is always performing with high F1 scores.

The next best classifier in terms of leading F1 scores is QDA, which has 7 of the best F1 scores. For QDA, all of these top results come in the first five experiments, where the training data includes more of the acoustic environment and conditions than in the last four experiments. QDA performs very strongly on the V+M test sets and on the F+V+M test sets for these experiments. This shows that if the training set has certain qualities similar to the test set, QDA could be a legitimate option for classifying between *natural* and *media*. However, the classifier that performs the best when the test data is most dissimilar to the training data is SVC. QDA does

reasonably well, but performs overall worse than SVC in the last four experiments, especially on the F+M and the F+V+M datasets. QDA could be a good option alongside SVC if we know that the testing environment and conditions will have similarities to the training set.

DT has the top average F1 scores on 4 of the tests. DT does very well when classifying the V+M test set, with high scores on three out of four of the V+M tests in the more difficult experiments. Except for LOROCV, DT performs very well on the V+M test sets on all of the experiments. However, there is a significant tradeoff seen in how well DT performs on the F+M test sets. DT might be very good at classifying between *natural* and *media* with *natural* video calls and *media* in the test set, but does very poorly at classifying *natural* family conversations. In this regard, SVC is better overall for its consistency across both the V+M and F+M test sets.

LR has the top average F1 score on only 2 of the tests, however we see that LR is able to generalize well to new *media* and *natural* audio. LR performs very well in many of the experiments, with F1 scores that are close to, albeit slightly worse than, SVC in most of the experiments. Especially in LORSO, LORDO, LOSDO, and LORSDO, we see that LR is able to perform consistently well on V and F data, with scores similar to that of SVC on the F+V+M datasets. LR does a good job at generalizing to new environments that it has not trained on for left out *media* data, and video calls and family conversations. However, QDA is better than LR when the training set is more similar to the test set, and SVC is better than LR when the test set is more dissimilar.

KNN and GNB have the worst performances on our experiments. KNN performs the best on V+M in LOROCV, but besides that, KNN and GNB show substantially worse performance than the other classifiers. They perform particularly poorly on LORSDO, which tests how well they can generalize when training on very dissimilar *media* data to the test set. We would not recommend KNN or GNB, especially when

compared to our other trained classifier.

Discussion

Overall, SVC is best able to generalize to new recordings. We see this both in SVC's ability to perform well on *natural* data that we recorded in our own homes, which was outside of the *natural* audio from the CHiME-5 corpus that the model was trained on, as well as good performance of the classifier to *media* from loudspeakers, microphone distances, and rooms that it was not trained on (Table B.23). SVC performs consistently well when tested on in-the-home, *natural* audio of both video calls (V) and family conversations (F). SVC performs with accuracies of over 85% on LORSDO, with recall scores of over 90% for natural V or F audio, and recall of over 81% for *media* data from a different room, loudspeaker, and microphone distance than it was trained on. We believe that SVC is the best classification algorithm that we experimented with at disambiguating between *natural* and *media*. It does the most consistently well across our tests sets in our experiments, and does the best at generalizing to new environments and conditions that it has not trained on.

LR also performs well on both of the *natural* test sets and on many of the experiments, but performs worse than SVC overall. QDA performs very well when tested against data with some similar characteristics to what it is trained with, but does more poorly on stricter generalizability tests. DT performs very well on video calls, but very poorly when tested against family conversations. KNN and GNB do not perform well.

Since QDA, LR, and SVC all perform well across all of our test sets and experiments, with QDA showing particularly strong performance when the *media* testing conditions have some similarities to their training coonditions, it could be an option to use an ensemble of classifiers in making the *natural* vs. *media* prediction. We need to verify that the classifiers do not take too long to make predictions and that they

do not take too much space in memory. If these two statements hold true, it could be reasonable to use all three in predicting *natural* vs. *media*. We perform these timing and size experiments in Section B.5.1.

B.5 Proposed Application

A critical criterion when selecting a classification algorithm is that it can perform in close to real-time to be suitable for a robot in the home or in the real world. A robot should provide a naturalistic and intuitive interaction for human users, so real-time classifications and responses are essential. Taking too much time to analyze the audio environment, extract features, make predictions, and act on those predictions may negatively affect the overall interaction. Keeping these factors in mind, we (a) perform several timing and size tests on various steps of the audio collection and decision-making process, (b) suggest an overall classification pipeline for a robot to implement this approach, and (c) present ethics and privacy considerations that were taken into account for this pipeline. For these timing and size experiments, we train the classifier on the entire *natural* C category that we compiled, and all of our *media* recordings.

B.5.1 Timing and Size Experiments

We measured the speed of feature extraction and prediction using around 45 minutes of audio data (540 5-second samples). Extracting features from each of the 540 audio samples took an average of 0.557 seconds (SD=0.0442 seconds) on a Dell Laptop with an Intel i5-5200U CPU @ 2.2GHz and 8GB RAM. To measure the average prediction time for each audio sample, we measured the time that it took to standardize and predict the entire (540x83) input vector and divided it by 540. The trained standardization scaler had a size of 4 kB. The average prediction times and the sizes on disk

for each trained classifier can be seen in Table B.3 below.

Table B.3: Classifier Size and Prediction Times

Model	**Avg. Prediction Time (ms)**	**Size (kB)**
KNN	2.524	13,545
QDA	0.01064	115
DT	0.00117	12
GNB	0.00312	4
LR	0.00366	4
SVC	0.17480	668

We see that all of the classifiers that we trained have fast prediction times. DT and GNB are the fastest, with LR and QDA next, then SVC, and KNN last. However, all the classifiers, except for KNN, are considerably faster than a millisecond, so we believe that any of the classifiers would be sufficient in that respect.

With respect to size on disk, LR and GNB are the smallest, with DT as next smallest. SVC is the second largest, but still not prohibitively large.

These sizes (and predictions) are also promising in that if the dataset were to get substantially larger, that most of these classification algorithms seem like they would be able to scale and still be reasonable to use on-board and real-time. This might not be true for KNN, but that was eliminated due to its poor performance on generalization.

This also means that after recording a 5-second sample, the whole classification process could be used on-board a robot, even on one with little memory. The whole classification process, after recording a 5-second sample, can take less than a second for feature extraction, standardization, and the prediction, making it possible to use this in real-time.

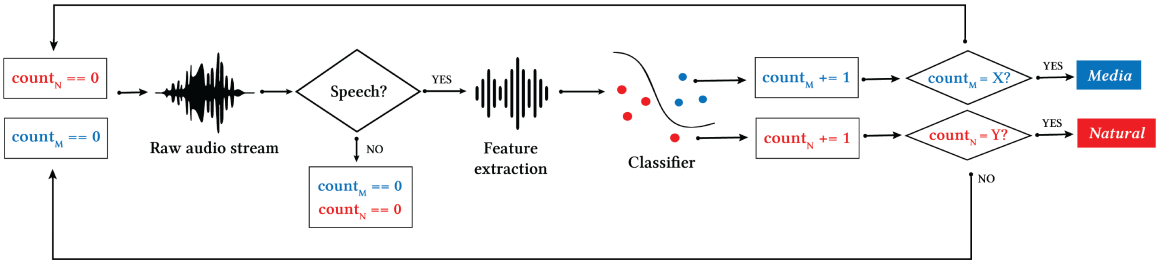

Figure B.1: Proposed classification pipeline. See Section B.5.2 for description.

Furthermore, a robot could reasonably include multiple trained classifiers on disk and require less than one megabyte (MB) of space. If using an ensemble of classifiers, the prediction time still remains substantially lower than one millisecond. Both the timing and size of the classifiers together allow for an ensemble to be used.

B.5.2 Classification Pipeline

In a real-world setting, we suggest our classifier be used as a part of a greater classification pipeline, shown in Figure B.1. A Kinect One microphone would be required[6], along with minimal onboard computing power. All audio collection, analysis, and computation can take place locally, without needing to offload any data to online services.

The system begins by recording a 5-second raw audio stream of the environment and initializing the count variables to 0. The system stores the recording and checks it for speech.[7] If speech is not detected, the system should loop back to the start by resetting the counts, and deletes the recording. If speech is detected, the feature extraction is performed, the audio is deleted, and a corresponding *natural* or *media* prediction is made. After a prediction, the corresponding *count* is incremented, and the other count is reset to 0. Only after X, or Y, consecutive predictions in a certain category will the decision be "final". Otherwise, the corresponding *count* is reset to

[6]We did not test multiple microphones so we cannot say whether or not our classifier would have any success recording with a different microphone.

[7]Speech could be checked for by using a voice activity detection (VAD) algorithm [187], trained on the Kinect, that can detect when human speech is a part of the acoustic environment.

0. Once a final decision is output by the pipeline, the process starts again, with both counts initialized to 0.

Depending on how sensitive we want the system to be to the classifier's predictions, we can alter the values of X and Y. For example, with $X = 3$, the classifier will have to predict close to 15 consecutive seconds (three decisions in a row) as *media*. This approach does not allow for one false positive to ruin the final classification, but rather the classifier would have to get the audio scene wrong three times in a row in order to make a mistake.

An alternative approach is to set both X=1 and Y=1, in which case the pipeline will be returning a final prediction on every 5-second audio sample, unless it does not detect speech. This will give a robot using this pipeline more frequent data points to use in its final decision making.

After the system determines whether or not the speech that it hears in its environment is *media* or *natural*, it can use this classification, along with other contextual information to make decisions on how to act. For example, the robot could also have other tools available to it that can detect characteristics from human speech such as tone, emotion, and intensity. The robot could also utilize context like the time of day, the day of the week, its location in the home, the current weather, and more.

Another interesting contextual tool that could be incorporated into this pipeline is sound source localization (SSL), which utilizes the microphone array of the Kinect. SSL could help the robot get an approximation of where the speech is coming from. This extra context, combined with the *natural* vs. *media* classification, could further assist the robot in making a more informed decision on social presence and providing it with a better understanding its environment. VAD and SSL could be combined to localize and individually classify multiple speakers in a noisy audio scene, but such VAD for multi-speaker diarization in real-world scenarios remains an open research problem[181].

This classification pipeline can provide the robot with an understanding of if speech is *natural* or *media* in its environment, helping it in inferring social presence. The robot can use this information, along with other context, to make appropriate decisions about how to interact, or not, and to best accommodate its user(s) and to reach its goals.

B.5.3 Ethics and Privacy Considerations

In home data is inherently sensitive, and the audio pipeline presented in our chapter is considerate of that. We believe our solution is minimally invasive. Using one modality (i.e., just audio) to make decisions is undoubtedly less invasive than using more. In fact, our suggested solution is computed locally (it is lightweight and would not require sending any sensitive data to online services), only needs to store a 5-second sample of audio at a time (which can be deleted immediately after features are extracted from it), and does not use any semantic representation or transcription of the audio (which could contain sensitive information) as a part of its decision making. These are important factors that keep users' privacy in mind.

B.6 Limitations

There are several limitations to this work that we believe are important to make clear. First, the dataset that we compiled could be more diverse and representative. Our *natural* training data is only comprised of audio from the CHiME-5 dataset, even though it does contain audio from different homes, rooms, and voices. Our *media* dataset contains three different rooms from within one home and five different electronic devices. Obviously, there are countless other possible devices from which audio can be emitted in the home, which were not included in our training set. Despite these limitations, our results showed that classifiers were able to make accurate *media*

classifications on audio from recording devices, rooms, microphone distances, and combinations of the three that they were not trained on, and the classifiers were able to classify *natural* audio from outside of the CHiME-5 training corpus, that included new rooms and voices in the V and F test sets. Another limitation is that the recordings in our V and F categories could be more diverse and comprehensive, with the inclusion of audio from more homes, families, and people. Also, we only focus on audio from the home, when ideally, such a classification tool should be able to make predictions in other dynamic, human environments as well.

Additionally, our dataset does not include examples of scenarios where *media* from television or radio shows is playing at the same time that *natural* conversation (that includes at least one co-located person) is occurring.[8] Further testing would be needed to see how our classifiers would perform when both *media* and *natural* audio are overlaid. We did see that in situations where electronic and organic speakers are conversing with each other in the audio scene (in our video calls test category), the classification algorithms classified the audio as *natural*. It could be beneficial if a robot could garner more detailed context of identifying, indexing, and classifying between each organic and electronic speaker engaged in the conversation, but we leave this as a future research direction. Regardless, through our experimentation in this chapter, we see that the classifiers can provide important context to a robot by accurately differentiating between common speech scenarios in the home from which social presence can be implied: popular genres in *media* originating from loudspeakers and *natural* conversation including a co-located user.

[8]Because the end goal of our *natural* vs. *media* classification is to help a robot in detecting a co-located user's social presence using sound, we would consider labeling this situation as *natural* because it includes conversational audio from a co-located person, and it implies that a user is physically present with the robot. However, it could be beneficial if a social robot could detect that there is both *natural* and *media* audio in the environment. Such knowledge could give it more nuanced context than purely a *natural* classification, but we leave this for future work.

B.7 Conclusions

Detecting social presence using sound involves being able to classify audio as containing either 1) *natural* conversation including at least one co-located user or 2) *media* playing from electronic sources that does not require a social response, such as television shows. It is important for in-home social robots to have such a capability, as the additional context can help them in their decision making. We perform an experimental evaluation that tests the robustness of several traditional machine learning classifiers on data from our compiled *natural* vs. *media* dataset. We conclude that a C-Support Vector Classification (SVC) algorithm outperforms other classifiers, and we propose a classification pipeline that can be utilized by social robots in the home to help them in detecting social presence using sound.

B.8 Model Hyperparameters

Table B.4: Hyperparameters Used for Gridsearch on Leave-One-Recording-Out Cross Validation

Model	**Hyperparameters**
KNN	'n_neighbors': [1,3,5,7,9], 'weights': ['uniform', 'distance'], 'p': [1,2]
QDA	'reg_param': [0.00001, 0.0001, 0.001,0.01, 0.1], 'tol': [0.0001, 0.001, 0.01, 0.1]
DT	'criterion': ['gini','entropy'], 'max_depth': [1,5,10,None], 'min_samples_split': [2,5,10], 'min_samples_leaf': [1,2,5]
GNB	'var_smoothing': [1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10, 1e-11, 1e-12, 1e-13, 1e-14, 1e-15]
LR	'solver': ['lbfgs','liblinear','newton-cg'], 'penalty': ['l1', 'l2'], 'C': [0.001,0.01,0.1,1,10,100,1000]
SVC	'kernel': ['linear','rbf'], 'gamma': ['scale', 'auto'], 'C': [0.1,1,10,1000]

Table B.5: Hyperparameters of Models Presented in Section B.4 Results

Model	**Hyperparameters**
KNN	'algorithm'='auto', 'leaf_size'=30, 'metric'='minkowski', 'metric_params'=None, 'n_jobs'=None, 'n_neighbors'=8, 'p'=1, 'weights'='distance'
QDA	'priors'=None, 'reg_param'=0.01, 'store_covariance'=False, 'tol'=0.0001.
DT	'ccp_alpha'=0.0, 'class_weight'=None, 'criterion'='entropy', 'max_depth'=None, 'max_features'=None, 'max_leaf_nodes'=None, 'min_impurity_decrease'=0.0, 'min_impurity_split'=None, 'min_samples_leaf'=2, 'min_samples_split'=2, 'min_weight_fraction_leaf'=0.0, 'random_state'=0, 'splitter'='best'
GNB	'priors': None, 'var_smoothing': 0.001
LR	'C': 0.1, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': None, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False.
SVC	'C': 10, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': "rbf", 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False

B.9 Experiment F1 Score Summaries

Table B.6: Leave-One-Recording-Out Cross Validation (LOROCV) Summary. We present the average F1 scores between each of the two classes across all LOROCV folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	**95.1**	87.7	86.4	85.5	86.4	90.6
	$F1_{media}$	**93.8**	90.2	89.4	86.5	89.3	91.9
	Avg. F1	**94.5**	89.0	87.9	86.0	87.9	91.3
F + M	$F1_{natural}$	86.9	**99.5**	98.9	96.4	99.0	96.5
	$F1_{media}$	87.2	**99.5**	98.8	95.9	98.8	96.6
	Avg. F1	87.1	**99.5**	98.9	96.2	98.9	96.6
F + V + M	$F1_{natural}$	91.7	93.1	92.3	90.4	92.3	**93.4**
	$F1_{media}$	90.9	93.4	93.2	90.5	93.2	**93.9**
	Avg. F1	91.3	93.3	92.8	90.5	92.8	**93.7**

Table B.7: Leave-One-Label-Out (LOLO) Summary. We present the average F1 scores between each of the two classes across all 14 LOLO folds. For each fold, a *media* recording and *natural* C recording were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	81.5	**99.3**	96.9	92.9	94.9	94.1
	$F1_{media}$	75.5	**99.2**	95.1	87.9	88.5	92.1
	Avg. F1	78.5	**99.3**	96.0	90.4	91.7	93.1
F + M	$F1_{natural}$	88.4	87.5	75.6	84.0	85.0	**90.2**
	$F1_{media}$	82.7	90.0	80.2	80.7	80.8	**90.6**
	Avg. F1	85.6	88.8	77.9	82.4	82.9	**90.4**
F + V + M	$F1_{natural}$	85.4	**93.0**	85.8	87.8	89.4	92.0
	$F1_{media}$	79.5	**93.8**	85.8	83.2	83.3	91.3
	Avg. F1	82.5	**93.4**	85.8	85.5	86.4	91.7

Table B.8: Leave-One-Room-Out (LORO) Summary. We present the average F1 scores between each of the two classes across all three LORO folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	77.8	**99.2**	94.8	85.6	88.9	94.7
	$F1_{media}$	76.9	**99.2**	93.7	76.3	81.0	95.0
	Avg. F1	77.4	**99.2**	94.3	81.0	85.0	94.9
F + M	$F1_{natural}$	85.2	84.3	71.8	82.9	82.5	**86.5**
	$F1_{media}$	81.0	**87.8**	78.3	84.0	82.9	86.4
	Avg. F1	83.1	86.1	75.1	83.5	82.7	**86.5**
F + V + M	$F1_{natural}$	81.5	**92.1**	84.1	84.0	85.2	90.4
	$F1_{media}$	79.3	**93.0**	85.4	81.7	83.7	90.7
	Avg. F1	80.4	**92.6**	84.8	82.9	84.5	90.6

Table B.9: Leave-One-Speaker-Out (LOSO) Summary. We present the average F1 scores between each of the two classes across all five LOSO folds. For each fold, all of a loudspeaker's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	78.8	98.5	**99.0**	94.1	97.1	93.4
	$F1_{media}$	75.0	98.4	**98.9**	93.4	97.1	93.4
	Avg. F1	76.9	98.5	**99.0**	93.8	97.1	93.4
F + M	$F1_{natural}$	87.0	83.1	72.7	82.0	83.8	**87.2**
	$F1_{media}$	80.1	85.7	77.6	79.4	84.8	**87.8**
	Avg. F1	83.6	84.4	75.2	80.7	84.3	**87.5**
F + V + M	$F1_{natural}$	83.3	**90.6**	85.9	87.6	90.2	90.1
	$F1_{media}$	78.4	**91.3**	87.2	86.4	90.5	90.4
	Avg. F1	80.9	**91.0**	86.6	87.0	90.4	90.3

Table B.10: Leave-One-Distance-Out Cross Validation (LODO) Summary. We present the average F1 scores between each of the two classes across all three LODO folds. For each fold, all of a microphone distance's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	79.7	**98.8**	97.4	92.1	97.7	94.5
	$F1_{media}$	78.1	**98.8**	97.1	91.1	97.8	95.0
	Avg. F1	78.9	**98.8**	97.3	91.6	97.8	94.8
F + M	$F1_{natural}$	81.5	82.5	70.9	76.9	82.2	**84.6**
	$F1_{media}$	67.0	**84.6**	56.8	66.3	80.9	83.6
	Avg. F1	74.3	83.6	63.9	71.6	81.6	**84.1**
F + V + M	$F1_{natural}$	79.9	**90.7**	83.6	83.9	89.6	89.2
	$F1_{media}$	75.4	**91.3**	79.2	79.6	89.5	89.5
	Avg. F1	77.7	**91.0**	81.4	81.8	89.6	89.4

Table B.11: Leave-One-Room and Speaker-Out (LORSO) Summary. We present the average F1 scores between each of the two classes across all nine LORSO folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	75.5	**92.2**	91.7	81.4	89.0	89.5
	$F1_{media}$	60.3	81.0	**84.0**	58.8	82.2	83.1
	Avg. F1	67.9	86.6	**87.9**	70.1	85.6	86.3
F + M	$F1_{natural}$	85.8	84.5	76.1	80.5	87.6	**89.1**
	$F1_{media}$	78.4	81.2	80.0	72.9	86.7	**87.0**
	Avg. F1	82.1	82.9	78.1	76.7	87.2	**88.1**
F + V + M	$F1_{natural}$	81.1	88.2	83.6	80.6	88.1	**89.0**
	$F1_{media}$	71.0	81.9	82.5	69.1	85.7	**86.2**
	Avg. F1	76.1	85.1	83.1	74.9	86.9	**87.6**

Table B.12: Leave-One-Recording and Distance-Out (LORDO) Summary. We present the average F1 scores between each of the two classes across all nine LORDO folds. For each fold, all of a room and microphone distance's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	76.1	94.0	**96.4**	85.4	91.5	89.6
	$F1_{media}$	63.8	90.9	**94.2**	71.6	86.0	85.1
	Avg. F1	70.0	92.5	**95.3**	78.5	88.8	87.4
F + M	$F1_{natural}$	83.9	86.3	75.6	80.3	89.2	**90.5**
	$F1_{media}$	72.7	87.0	80.2	73.8	89.6	**90.5**
	Avg. F1	78.3	86.7	77.9	77.1	89.4	**90.5**
F + V + M	$F1_{natural}$	80.2	90.0	85.8	82.6	**90.2**	90.0
	$F1_{media}$	69.8	88.8	85.8	73.4	**88.9**	**88.9**
	Avg. F1	75.0	89.4	85.8	78.0	**89.6**	89.5

Table B.13: Leave-One-Speaker and Distance-Out (LOSDO) Summary. We present the average F1 scores between each of the two classes across all nine LOSDO folds. For each fold, all of a speaker and microphone distance combination's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	79.7	94.6	93.7	89.6	**95.2**	94.3
	$F1_{media}$	72.8	87.2	86.4	80.7	**95.4**	94.6
	Avg. F1	76.3	90.9	90.1	85.2	**95.3**	94.5
F + M	$F1_{natural}$	84.0	82.7	77.4	81.6	85.5	**87.0**
	$F1_{media}$	74.2	78.9	75.7	75.0	81.0	**82.5**
	Avg. F1	79.1	80.8	76.6	78.3	83.3	**84.8**
F + V + M	$F1_{natural}$	82.1	88.3	85.4	85.2	89.7	**90.0**
	$F1_{media}$	73.7	82.6	80.2	77.5	88.3	**89.0**
	Avg. F1	77.9	85.5	82.8	81.4	89.0	**89.5**

Table B.14: Leave-One-Room and Speaker and Distance-Out (LORSDO) Summary. We present the average F1 scores between each of the two classes across all 14 LORSDO folds. For each fold, all of a room, speaker, and microphone distance combination's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	74.1	89.6	**91.1**	81.2	89.1	88.1
	$F1_{media}$	57.4	75.1	83.3	58.8	**83.8**	82.5
	Avg. F1	65.8	82.4	**87.2**	70.0	86.5	85.3
F + M	$F1_{natural}$	80.9	83.6	72.3	76.9	86.7	**88.2**
	$F1_{media}$	65.0	70.5	73.1	56.4	80.6	**82.1**
	Avg. F1	73.0	77.1	72.7	66.7	83.7	**85.2**
F + V + M	$F1_{natural}$	78.0	86.4	82.0	78.9	87.8	**88.1**
	$F1_{media}$	61.7	72.5	77.2	57.5	82.2	**82.4**
	Avg. F1	69.9	79.5	79.6	68.2	85.0	**85.3**

B.10 Experiment Comprehensive Results

Table B.15: Leave-One-Recording-Out CV Results. The table presents the macro and micro averages across all LOROCV folds for each classifier.

				Our In-Home Natural Recordings					
		C+M		F+M		V+M		V+F+M	
Model	Metrics	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	96.8	96.3	94.7	94.0	87.3	86.9	91.5	90.9
	Precision_N	95.7	94.8	95.7	94.7	94.8	93.7	95.3	94.2
	Recall_N	99.5	99.6	95.0	94.6	81.0	81.1	89.0	88.7
	F1_N	97.3	96.8	95.1	94.4	86.9	86.5	91.7	91.1
	Precision_M	99.4	94.8	94.1	93.7	82.5	82.3	88.7	88.4
	Recall_M	94.1	93.1	94.4	93.4	93.6	92.7	94.1	93.1
	F1_M	95.8	95.3	93.8	93.1	87.2	86.7	90.9	90.2
QDA	Accuracy	99.6	99.6	89.1	88.9	99.5	99.4	93.6	93.5
	Precision_N	99.8	99.7	99.8	99.7	99.7	99.7	99.7	99.7
	Recall_N	99.5	99.5	78.4	78.0	99.3	99.1	87.3	87.2
	F1_N	99.6	99.6	87.7	87.4	99.5	99.4	93.1	93.0
	Precision_M	99.5	99.7	82.4	82.1	99.3	99.1	88.8	88.7
	Recall_M	99.8	99.7	99.8	99.8	99.7	99.7	99.8	99.7
	F1_M	99.6	99.6	90.2	90.0	99.5	99.4	93.4	93.4
DT	Accuracy	99.0	99.0	88.2	88.3	98.9	98.8	92.8	92.9
	Precision_N	99.2	99.2	99.0	99.1	99.3	99.3	99.1	99.2
	Recall_N	99.1	99.1	77.6	77.7	98.7	98.6	86.7	86.7
	F1_N	99.1	99.1	86.4	86.6	98.9	98.9	92.3	92.3
	Precision_M	99.1	99.2	82.0	82.0	98.7	98.6	88.3	88.4
	Recall_M	99.0	99.0	98.8	98.9	99.1	99.1	99.0	99.0
	F1_M	99.0	99.0	89.4	89.4	98.8	98.8	93.2	93.2
GNB	Accuracy	92.2	92.8	86.2	87.4	96.2	96.7	90.5	91.4
	Precision_N	93.7	94.8	93.0	94.1	94.7	95.7	93.6	94.7
	Recall_N	91.2	91.1	79.7	80.8	98.5	98.4	87.8	88.4
	F1_N	91.7	92.2	85.5	86.7	96.4	96.9	90.4	91.3
	Precision_M	92.7	94.8	81.2	82.9	98.5	98.3	88.4	89.0
	Recall_M	93.2	94.4	92.6	94.0	93.9	95.0	93.2	94.4
	F1_M	92.3	93.1	86.5	87.9	95.9	96.4	90.5	91.5
LR	Accuracy	99.0	99.0	88.2	88.3	98.9	98.8	92.8	92.9
	Precision_N	99.2	99.2	99.0	99.1	99.3	99.3	99.1	99.1
	Recall_N	99.1	99.1	77.5	77.7	98.7	98.6	86.7	86.7
	F1_N	99.1	99.1	86.4	86.6	99.0	98.9	92.3	92.3
	Precision_M	99.1	99.2	82.0	82.0	98.7	98.6	88.3	88.4
	Recall_M	98.9	99.0	98.8	98.9	99.1	99.0	98.9	99.0
	F1_M	98.9	98.9	89.3	89.4	98.8	98.8	93.2	93.2
SVC	Accuracy	99.4	99.4	91.4	91.5	96.6	96.5	93.6	93.7
	Precision_N	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
	Recall_N	99.6	99.6	83.5	83.8	94.0	93.9	88.0	88.2
	F1_N	99.4	99.5	90.6	90.9	96.5	96.5	93.3	93.4
	Precision_M	99.6	99.4	85.5	86.0	94.4	94.2	89.3	89.4
	Recall_M	99.2	99.2	99.2	99.2	99.1	99.2	99.2	99.2
	F1_M	99.3	99.3	91.9	92.1	96.6	96.5	93.9	93.9

Table B.16: Leave-One-Label-Out Results. The table presents the macro and micro averages across all LOLO folds for each classifier.

Model	Metrics	LOLO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	79.5	78.8	86.4	86.1	83.4	82.9
	Precision_N	84.1	83.9	85	85.3	84.5	84.5
	Recall_N	81.2	80.5	94	92.8	88.4	87.3
	F1_N	81.5	80.9	88.4	88	85.4	84.9
	Precision_M	76.2	75.4	91.6	90.5	84.4	83.5
	Recall_M	77.8	77.2	78.8	79.4	78.3	78.4
	F1_M	75.5	74.7	82.7	82.6	79.5	79.1
QDA	Accuracy	99.2	99.2	88.9	87.7	93.4	92.8
	Precision_N	99.2	99.3	99.3	99.3	99.3	99.3
	Recall_N	99.3	99.1	78.3	76	87.5	86.3
	F1_N	99.3	99.2	87.5	86	93	92.3
	Precision_M	99.3	99.1	82.2	80.7	88.9	87.9
	Recall_M	99.2	99.3	99.5	99.5	99.3	99.4
	F1_M	99.2	99.2	90	89	93.8	93.3
DT	Accuracy	95.6	96.3	78.6	78	86.1	86.1
	Precision_N	94	94.8	93.2	93.4	93.6	94.1
	Recall_N	99.8	99.8	64.9	63.1	80.2	79.5
	F1_N	96.4	96.9	75.6	74.5	85.8	85.7
	Precision_M	99.6	99.7	71.9	71.1	81.8	81.4
	Recall_M	91.5	92.7	92.4	92.8	92	92.8
	F1_M	94.2	95.1	80.2	79.9	85.8	86
GNB	Accuracy	90.7	91.5	83.2	84	86.5	87.3
	Precision_N	88.4	89.4	87.1	88.9	87.5	89
	Recall_N	98.2	97.9	82.4	81.7	89.3	88.9
	F1_N	92.4	92.9	84	84.5	87.8	88.4
	Precision_M	91.1	91.9	78	78.4	83.2	83.8
	Recall_M	83.1	85	84	86.3	83.6	85.7
	F1_M	86.4	87.9	80.7	82	83.2	84.6
LR	Accuracy	92.3	93.2	84.1	84	87.7	88.1
	Precision_N	92.6	93.3	92.6	93.4	92.6	93.3
	Recall_N	98.3	98.2	80.5	79.2	88.4	87.7
	F1_N	94.4	94.9	85	84.5	89.4	89.5
	Precision_M	92.5	93.3	76.6	76.4	82.2	82.6
	Recall_M	86.4	88.1	87.6	88.8	87	88.5
	F1_M	86.7	88.5	80.8	81.2	83.3	84.3
SVC	Accuracy	93.4	93.5	90.5	90.1	91.8	91.6
	Precision_N	95.6	95.5	97.1	97.5	96.3	96.5
	Recall_N	93.6	93.5	84.9	83.5	88.7	88
	F1_N	94.1	94.1	90.2	89.6	92	91.7
	Precision_M	92.7	92.8	86.4	85.4	89.1	88.5
	Recall_M	93.3	93.5	96.2	96.7	94.9	95.3
	F1_M	91.8	92.1	90.6	90.3	91.3	91.2

Table B.17: Leave-One-Room-Out Results. The table presents the results of the three LORO folds (each room column is the left-out room), and the macro averages across all LORO folds for each classifier. Only macro averages are presented because the test sets were the same size (the left out room *media* set was larger than the *natural* testing subset, so *media* was sampled to match the size of the natural sets).

		Kitchen			Bedroom			Playroom			Average		
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M
KNN	Accuracy	82.4	84.3	83.4	73.1	93.2	83.4	76.7	72.8	74.7	77.4	83.4	80.5
	Precision_N	82.1	77.4	79.4	71.4	97.3	83.6	76.3	67.1	71	76.6	80.6	78
	Recall_N	82.9	97	90.1	77.2	88.8	83.2	77.4	89.5	83.6	79.2	91.8	85.6
	F1_N	82.5	86.1	84.4	74.2	92.8	83.4	76.9	76.7	76.8	77.8	85.2	81.5
	Precision_M	82.7	96	88.6	75.2	89.7	83.2	77.1	84.3	80.1	78.3	90	84
	Recall_M	81.9	71.6	76.6	69	97.5	83.6	76	56.2	65.8	75.6	75.1	75.4
	F1_M	82.3	82	82.2	72	93.4	83.4	76.6	67.4	72.3	76.9	81	79.3
QDA	Accuracy	98.9	89.6	94.2	99.6	82.5	90.8	99	86.9	92.8	99.2	86.3	92.6
	Precision_N	99.7	99.1	99.4	100	99.7	99.9	98.3	94.4	96.5	99.3	97.7	98.6
	Recall_N	98.1	80	88.8	99.2	65.2	81.8	99.7	78.4	88.8	99	74.5	86.5
	F1_N	98.9	88.5	93.8	99.6	78.8	89.9	99	85.7	92.5	99.2	84.3	92.1
	Precision_M	98.2	83.2	89.9	99.2	74.1	84.6	99.7	81.6	89.6	99	79.6	88
	Recall_M	99.7	99.3	99.5	100	99.8	99.9	98.3	95.3	96.8	99.3	98.1	98.7
	F1_M	98.9	90.5	94.5	99.6	85.1	91.6	99	87.9	93.1	99.2	87.8	93
DT	Accuracy	98.4	81.7	89.8	86.3	72.4	79.2	98.4	73	85.3	94.3	75.7	84.8
	Precision_N	96.9	99	97.8	79.5	88.3	82.4	97.1	73.7	85.5	91.2	87	88.6
	Recall_N	99.9	63.9	81.5	97.7	51.7	74.1	99.7	71.3	85.1	99.1	62.3	80.2
	F1_N	98.4	77.7	88.9	87.7	65.2	78.1	98.4	72.5	85.3	94.8	71.8	84.1
	Precision_M	99.9	73.4	84.1	97	65.9	76.5	99.7	72.2	85.2	98.9	70.5	81.9
	Recall_M	96.8	99.4	98.1	74.8	93.2	84.2	97.1	74.6	85.5	89.6	89.1	89.3
	F1_M	98.3	84.4	90.6	84.5	77.2	80.2	98.3	73.4	85.4	93.7	78.3	85.4
GNB	Accuracy	87.9	82.5	85.1	64.1	85	74.8	95	83.1	88.9	82.4	83.5	82.9
	Precision_N	82.5	80.2	81.4	58.6	96.2	70.9	91.6	85.1	88.5	77.6	87.2	80.3
	Recall_N	96.3	86.2	91.1	96.1	72.8	84.2	99	80.2	89.4	97.1	79.8	88.2
	F1_N	88.9	83.1	86	72.8	82.9	77	95.2	82.6	88.9	85.6	82.9	84
	Precision_M	95.5	85.1	89.9	89.1	78.1	80.5	98.9	81.3	89.3	94.5	81.5	86.6
	Recall_M	79.6	78.8	79.2	32.2	97.1	65.5	90.9	85.9	88.4	67.6	87.3	77.7
	F1_M	86.9	81.8	84.2	47.3	86.6	72.2	94.8	83.5	88.8	76.3	84	81.7
LR	Accuracy	98.5	87.9	93	66	87.9	77.2	94.3	72.8	83.3	86.3	82.9	84.5
	Precision_N	99.9	95.9	98	59.8	100	73	96.9	70	81.9	85.5	88.6	84.3
	Recall_N	97.1	79.2	87.9	97.7	75.8	86.5	91.6	79.9	85.6	95.5	78.3	86.6
	F1_N	98.5	86.7	92.7	74.2	86.2	79.1	94.2	74.6	83.7	88.9	82.5	85.2
	Precision_M	97.1	82.3	89	93.7	80.5	83.4	92	76.6	84.9	94.3	79.8	85.8
	Recall_M	99.9	96.6	98.2	34.2	100	68	97.1	65.8	81	77.1	87.5	82.4
	F1_M	98.5	88.8	93.4	50.2	89.2	74.9	94.5	70.8	82.9	81	82.9	83.7
SVC	Accuracy	94.9	93.5	94.2	95.9	89.6	92.7	93.7	76.3	84.8	94.8	86.5	90.6
	Precision_N	98.6	95.3	96.8	100	100	100	96.4	73.6	83.7	98.3	89.6	93.5
	Recall_N	91.2	91.5	91.3	91.7	79.3	85.3	90.8	82.2	86.4	91.2	84.3	87.7
	F1_N	94.7	93.3	94	95.7	88.4	92.1	93.5	77.6	85	94.7	86.5	90.4
	Precision_M	91.8	91.8	91.8	92.3	82.8	87.2	91.3	79.8	85.9	91.8	84.8	88.3
	Recall_M	98.7	95.4	97	100	100	100	96.6	70.5	83.2	98.4	88.6	93.4
	F1_M	95.1	93.6	94.3	96	90.6	93.2	93.9	74.8	84.6	95	86.4	90.7

Table B.18: Leave-One-Speaker-Out Results. The table presents the results of the five LOSO folds (each speaker column is the left-out speaker), and the macro (M) and micro (μ) averages across all LOSO folds for each classifier.

		Bose			iPhone			BigBose			Sony			Mac			Average					
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M		F+M		V+F+M	
																	M	μ	M	μ	M	μ
KNN	Accuracy	68.4	59.8	64	87.1	95.7	92	87.1	95.5	91.8	56.2	73.9	65.9	86.6	95.8	91.3	77.1	76.2	84.1	82.5	81	79.6
	Precision_N	66.2	56.1	60.2	91.1	100	96.1	96.7	98.9	98	54.2	67.9	61.4	89.1	96.7	93.1	79.4	77.8	83.9	81.8	81.7	79.7
	Recall_N	75.5	89.6	82.7	82.3	91.5	87.4	76.9	92.1	85.3	79.6	90.7	85.7	83.5	94.8	89.3	79.5	79.5	91.7	91.8	86.1	86
	F1_N	70.5	69	69.7	86.4	95.5	91.6	85.6	95.4	91.2	64.5	77.6	71.5	86.2	95.8	91.2	78.7	77.9	86.7	85.5	83	81.9
	Precision_M	71.4	74.3	72.4	83.8	92.1	88.5	80.8	92.6	87	61.6	86	76.3	84.5	94.9	89.7	76.4	76.1	88	87.3	82.8	82.1
	Recall_M	61.4	29.9	45.3	91.9	100	96.5	97.3	98.9	98.2	32.8	57.1	46.1	89.7	96.8	93.4	74.6	72.9	76.5	73.3	75.9	73.1
	F1_M	66	42.7	55.7	87.7	95.9	92.3	88.3	95.7	92.3	42.8	68.6	57.5	87	95.8	91.5	74.4	73.4	79.7	77.4	77.9	76
QDA	Accuracy	94.4	65.2	79.4	99.2	85.9	91.7	99.3	85.6	91.7	99.5	86.5	92.4	98.9	84.7	91.6	98.3	97.9	81.6	80.6	89.4	88.7
	Precision_N	90.1	61.2	73.8	99.8	99.8	99.8	99.8	99.2	99.5	100	100	100	99	98.8	98.9	97.7	97	91.8	90	94.4	92.9
	Recall_N	99.8	83.1	91.2	98.6	71.9	83.5	98.9	71.7	83.8	99.1	73	84.8	98.8	70.3	84.2	99	99.1	74	74.4	85.5	85.9
	F1_N	94.7	70.5	81.6	99.2	83.6	90.9	99.3	83.2	91	99.5	84.4	91.8	98.9	82.1	91	98.3	98	80.8	80.1	89.2	88.7
	Precision_M	99.8	73.7	88.5	98.6	78	85.8	98.9	77.9	86	99.1	78.7	86.8	98.8	76.9	86.2	99	99.1	77.1	76.8	86.7	86.8
	Recall_M	89	47.4	67.6	99.8	99.8	99.8	99.8	99.4	99.6	100	100	100	99	99.2	99.1	97.5	96.7	89.1	86.8	93.2	91.4
	F1_M	94.1	57.7	76.7	99.2	87.6	92.3	99.3	87.3	92.3	99.5	88.1	92.9	98.9	86.6	92.2	98.2	97.8	81.5	80.1	89.3	88.4
DT	Accuracy	95.7	55.8	75.2	99.3	80.7	88.8	100	79.4	88.6	99.8	79.1	88.5	99	69.6	83.9	98.8	98.5	72.9	71.3	85	84
	Precision_N	92.2	54.6	71.5	98.6	99.7	99.1	100	100	100	99.7	100	99.8	98.2	72.9	86.3	97.7	97.2	85.4	82.2	91.3	89.1
	Recall_N	100	68.8	84	100	61.6	78.3	100	58.8	77.1	100	58.2	77.1	99.8	62.3	80.6	100	99.9	61.9	62.4	79.4	79.9
	F1_N	95.9	60.9	77.2	99.3	76.1	87.5	100	74.1	87.1	99.8	73.6	87	99	67.2	83.3	98.8	98.5	70.4	69.3	84.4	83.7
	Precision_M	100	57.8	80.6	100	72.2	82.1	100	70.8	81.4	100	70.5	81.3	99.8	67.1	81.8	100	99.9	67.7	66.8	81.4	81.4
	Recall_M	91.5	42.7	66.5	98.6	99.8	99.3	100	100	100	99.7	100	99.9	98.1	76.9	87.2	97.6	97	83.9	80.3	90.6	88.1
	F1_M	95.6	49.1	72.9	99.3	83.8	89.9	100	82.9	89.7	99.8	82.7	89.7	99	71.7	84.4	98.7	98.4	74	72	85.3	84.1
GNB	Accuracy	93.6	55.5	74.1	96.4	86.8	91	96.7	87.8	91.8	95.7	88.3	91.7	80.7	85.1	83	92.6	91.4	80.7	79.2	86.3	84.9
	Precision_N	88.9	53.8	68.7	94.3	92.6	93.4	96.1	95.9	96	94.6	98.5	96.5	73.2	83.1	77.7	89.4	87.6	84.8	82.5	86.5	84.1
	Recall_N	99.7	77.8	88.5	98.8	80.1	88.3	97.3	79	87.2	97.1	77.8	86.5	96.8	88.2	92.4	98	98	80.6	80.9	88.6	88.9
	F1_N	94	63.6	77.3	96.5	85.9	90.8	96.7	86.7	91.4	95.8	87	91.2	83.4	85.6	84.4	93.3	92.2	81.7	80.7	87	85.9
	Precision_M	99.6	59.9	83.8	98.8	82.4	88.9	97.3	82.2	88.3	97	81.7	87.8	95.3	87.4	90.6	97.6	97.5	78.7	77.9	87.9	87.7
	Recall_M	87.6	33.2	59.7	94	93.6	93.8	96	96.7	96.4	94.4	98.9	96.9	64.6	82.1	73.5	87.3	84.8	80.9	77.5	84	80.9
	F1_M	93.2	42.7	69.7	96.3	87.7	91.3	96.7	88.8	92.1	95.7	89.5	92.1	77	84.7	81.2	91.8	90.3	78.7	76.5	85.3	83.5
LR	Accuracy	91.2	73.1	81.9	99.3	88.1	93	99.3	87.7	92.9	96.8	88.3	92.1	99.2	87.3	93.1	97.2	96.6	84.9	84.2	90.6	90
	Precision_N	91.7	69	78.9	100	100	100	99.8	100	99.9	95.1	99.8	97.3	99.5	99.6	99.5	97.2	96.7	93.7	92.3	95.1	94
	Recall_N	90.6	83.6	87	98.6	76.2	85.9	98.9	75.4	85.8	98.6	76.7	86.6	98.9	74.9	86.6	97.1	96.6	77.4	77.7	86.4	86.5
	F1_N	91.2	75.6	82.8	99.3	86.5	92.4	99.3	86	92.3	96.8	86.7	91.6	99.2	85.5	92.6	97.2	96.6	84.1	83.5	90.4	89.9
	Precision_M	90.7	79.2	85.6	98.6	80.7	87.7	98.9	80.2	87.6	98.6	81.1	87.9	98.9	79.9	88.1	97.1	96.6	80.2	80.2	87.4	87.3
	Recall_M	91.8	62.5	76.8	100	100	100	99.8	100	99.9	94.9	99.9	97.6	99.5	99.7	99.6	97.2	96.7	92.4	90.7	94.8	93.5
	F1_M	91.3	69.9	80.9	99.3	89.3	93.4	99.3	89	93.3	96.7	89.5	92.5	99.2	88.7	93.5	97.2	96.6	85.3	84.4	90.7	90.1
SVC	Accuracy	85.4	76	80.6	95.9	91.5	93.4	96.7	90.7	93.4	91.7	90.3	91	97.1	89.8	93.4	93.4	92.8	87.7	86.9	90.3	89.7
	Precision_N	85.6	72	78.1	100	100	100	100	100	100	89.8	100	94.7	100	100	100	95.1	94.3	94.4	93.1	94.6	93.4
	Recall_N	85.1	85.1	85.1	91.7	83.1	86.8	93.4	81.5	86.8	94.1	80.6	86.7	94.2	79.7	86.8	91.7	91.3	82	82	86.4	86.3
	F1_N	85.3	78	81.4	95.7	90.8	93	96.6	89.8	92.9	91.9	89.3	90.6	97	88.7	92.9	93.3	92.7	87.3	86.7	90.2	89.5
	Precision_M	85.2	81.8	83.6	92.3	85.5	88.4	93.8	84.4	88.3	93.8	83.8	87.8	94.5	83.1	88.3	91.9	91.5	83.7	83.5	87.3	87
	Recall_M	85.7	66.9	76.1	100	100	100	100	100	100	89.3	100	95.2	100	100	100	95	94.2	93.4	91.9	94.3	93
	F1_M	85.4	73.6	79.7	96	92.2	93.8	96.8	91.5	93.8	91.5	91.2	91.3	97.2	90.8	93.8	93.4	92.8	87.9	87	90.5	89.7

Table B.19: Leave-One-Distance-Out Results. The table presents the results of the three LODO folds (each microphone distance is the left-out distance), and the macro averages across all LODO folds for each classifier. Only macro averages are presented because the test sets were the same size (the left out microphone distance *media* set was larger than the *natural* testing subset, so *media* was sampled to match the size of the natural sets).

		1 ft			4-6 ft			8-10 ft			LODO Average		
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M
KNN	Accuracy	83.3	55.3	68.9	87.9	84.4	86.1	65.8	91.6	79	79	77.1	78
	Precision_N	87.7	53	64	88.9	80.6	84.3	63.1	93.5	76.9	79.9	75.7	75.1
	Recall_N	77.5	94.8	86.4	86.7	90.7	88.7	75.9	89.3	82.8	80	91.6	86
	F1_N	82.3	68	73.6	87.8	85.3	86.5	68.9	91.4	79.8	79.7	81.5	79.9
	Precision_M	79.9	75.2	79.1	87	89.3	88.1	69.8	89.8	81.4	78.9	84.8	82.9
	Recall_M	89.1	15.8	51.5	89.2	78.1	83.5	55.6	93.8	75.2	78	62.6	70.1
	F1_M	84.2	26	62.4	88.1	83.4	85.8	61.9	91.7	78.2	78.1	67	75.4
QDA	Accuracy	98.6	75.3	86.7	99.1	87	92.9	98.5	88.7	93.5	98.8	83.7	91
	Precision_N	97.5	76.2	86.9	99.6	98.8	99.2	98.8	99.5	99.1	98.6	91.5	95.1
	Recall_N	99.8	73.5	86.3	98.7	74.9	86.5	98.3	77.8	87.8	98.9	75.4	86.9
	F1_N	98.7	74.8	86.6	99.1	85.2	92.4	98.5	87.3	93.1	98.8	82.5	90.7
	Precision_M	99.8	74.4	86.4	98.7	79.8	88	98.3	81.8	89	98.9	78.7	87.8
	Recall_M	97.5	77.1	87	99.6	99.1	99.3	98.8	99.6	99.2	98.6	91.9	95.2
	F1_M	98.6	75.7	86.7	99.1	88.4	93.3	98.5	89.8	93.8	98.8	84.6	91.3
DT	Accuracy	93.2	41.1	66.5	99.5	79.4	89.2	99	82.6	90.6	97.2	67.7	82.1
	Precision_N	88.2	45	61.3	99.1	92.8	96.5	98.1	97.6	97.9	95.1	78.5	85.2
	Recall_N	99.8	80.2	89.7	99.9	63.7	81.3	100	66.8	83	99.9	70.3	84.7
	F1_N	93.6	57.7	72.8	99.5	75.6	88.3	99	79.3	89.8	97.4	70.9	83.6
	Precision_M	99.7	9.48	80.8	99.9	72.4	83.9	100	74.8	85.2	99.9	52.2	83.3
	Recall_M	86.6	2.07	43.3	99.1	95	97	98	98.3	98.2	94.6	65.1	79.5
	F1_M	92.7	3.4	56.4	99.5	82.2	90	99	85	91.3	97.1	56.8	79.2
GNB	Accuracy	91.7	46	68.2	88.1	84.3	86.2	95	89.1	92	91.6	73.1	82.1
	Precision_N	86.1	47.4	63.4	83.6	82	82.8	93.8	96.4	95	87.8	75.3	80.4
	Recall_N	99.3	73.7	86.2	94.9	87.9	91.3	96.5	81.1	88.6	96.9	80.9	88.7
	F1_N	92.3	57.7	73.1	88.9	84.8	86.8	95.1	88.1	91.7	92.1	76.9	83.9
	Precision_M	99.2	40.9	78.4	94.1	86.9	90.3	96.4	83.7	89.3	96.6	70.5	86
	Recall_M	84	18.2	50.3	81.4	80.7	81	93.6	97	95.3	86.3	65.3	75.5
	F1_M	91	25.2	61.3	87.3	83.7	85.4	95	89.9	92.2	91.1	66.3	79.6
LR	Accuracy	95.6	67.3	81.1	98.8	88.8	93.7	98.8	89.3	93.9	97.7	81.8	89.6
	Precision_N	99.9	63.8	78.6	99	97.8	98.5	99	99.9	99.4	99.3	87.2	92.2
	Recall_N	91.4	79.9	85.5	98.6	79.4	88.7	98.6	78.8	88.4	96.2	79.3	87.5
	F1_N	95.4	71	81.9	98.8	87.6	93.3	98.8	88.1	93.6	97.7	82.2	89.6
	Precision_M	92.1	73.1	84.1	98.6	82.7	89.7	98.6	82.5	89.6	96.4	79.4	87.8
	Recall_M	99.9	54.7	76.7	99	98.2	98.6	99	99.9	99.5	99.3	84.3	91.6
	F1_M	95.8	62.6	80.2	98.8	89.8	94	98.8	90.3	94.3	97.8	80.9	89.5
SVC	Accuracy	94	71.1	82.3	96.9	90.3	93.5	93.3	91.5	92.3	94.7	84.3	89.4
	Precision_N	99.8	67.4	80.6	100	99	99.5	96.6	99.4	98	98.8	88.6	92.7
	Recall_N	88.2	81.9	85	93.9	81.3	87.5	89.7	83.4	86.5	90.6	82.2	86.3
	F1_N	93.6	73.9	82.7	96.9	89.3	93.1	93	90.7	91.9	94.5	84.6	89.2
	Precision_M	89.4	76.9	84.1	94.2	84.2	88.8	90.4	85.7	87.9	91.4	82.3	86.9
	Recall_M	99.8	60.4	79.6	100	99.2	99.6	96.8	99.5	98.2	98.9	86.4	92.5
	F1_M	94.3	67.7	81.8	97	91.1	93.9	93.5	92.1	92.8	95	83.6	89.5

Table B.20: Leave-One-Room+Speaker-Out Results. The table presents the macro and micro averages across all LORSO folds for each classifier.

Model	Metrics	LORSO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	70.3	67	83	81.9	77.3	75.1
	Precision_N	71.6	68.8	79.5	78.2	75.6	73.4
	Recall_N	82.7	81.6	95.1	94	89.6	88.4
	F1_N	75.5	73.2	85.8	84.7	81.1	79.3
	Precision_M	67.8	63	93.2	91.9	83.6	81.5
	Recall_M	57.9	52.3	71	69.8	65.1	61.9
	F1_M	60.3	54.6	78.4	77.4	71	68.1
QDA	Accuracy	89.1	90.1	84.2	85	86.4	87.3
	Precision_N	88.4	89.3	92.5	94.5	90	91.2
	Recall_N	99.3	99.2	80.3	77.5	88.9	87.4
	F1_N	92.2	92.8	84.5	84.1	88.2	88.3
	Precision_M	99.3	99.2	80.8	80.3	87.3	86.8
	Recall_M	79	80.9	88.1	92.5	84	87.2
	F1_M	81	83	81.2	84.2	81.9	84.7
DT	Accuracy	89.2	86.9	78.6	79.2	83.3	82.7
	Precision_N	86.3	83.6	90.2	90.2	87.3	85.8
	Recall_N	100	100	67.4	67.5	82	82.2
	F1_N	91.7	90	76.1	76.6	83.6	83.2
	Precision_M	99.9	99.9	73.5	73.8	82.4	82.3
	Recall_M	78.5	73.9	89.8	90.9	84.6	83.2
	F1_M	84	80.3	80	81	82.5	81.9
GNB	Accuracy	75.1	75.3	78.5	80.9	76.9	78.4
	Precision_N	71.7	72.1	80.5	83.9	74.8	76.3
	Recall_N	97.6	97.6	83.2	81.8	89.7	88.9
	F1_N	81.4	81.6	80.5	81.8	80.6	81.3
	Precision_M	95.7	95.6	78.7	80.1	83.6	84.3
	Recall_M	52.5	53	73.7	80.1	64.1	67.8
	F1_M	58.8	59	72.9	78	69.1	72.8
LR	Accuracy	86.8	86.3	87.6	88	87.2	87.2
	Precision_N	86.3	85.7	93.9	95.4	89.4	89.7
	Recall_N	93.5	94	83.6	81.6	88.1	87.2
	F1_N	89	88.9	87.6	87.4	88.1	87.9
	Precision_M	87.7	87	86	84.3	87.4	86.7
	Recall_M	80.1	78.7	91.5	94.4	86.3	87.3
	F1_M	82.2	80.8	86.7	88	85.7	86.1
SVC	Accuracy	87.4	86.5	88.4	88.4	88	87.5
	Precision_N	88.4	88	91.1	91.9	89.1	89.2
	Recall_N	92.5	91.6	88.6	86.4	90.4	88.7
	F1_N	89.5	88.8	89.1	88.5	89	88.3
	Precision_M	87.4	85.3	89.3	87.3	89.6	88
	Recall_M	82.4	81.4	88.3	90.4	85.6	86.3
	F1_M	83.1	81.4	87	87.8	86.2	86.1

Table B.21: Leave-One-Room+Distance-Out Results. The table presents the macro and micro averages across all LORDO folds for each classifier.

Model	Metrics	LORDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	71.7	74.2	80.2	81.4	76.4	78.1
	Precision_N	71.8	74.1	77.3	79.2	74.2	76
	Recall_N	83.3	83.9	94.2	94.3	89.3	89.5
	F1_N	76.1	77.8	83.9	84.9	80.2	81.3
	Precision_M	72	75.1	90.7	92	83.1	84.8
	Recall_M	60.2	64.6	66.2	68.6	63.6	66.7
	F1_M	63.8	68.1	72.7	73.9	69.8	72.6
QDA	Accuracy	92.8	94.8	86.8	87.5	89.5	90.9
	Precision_N	90.2	93.3	92.8	95.1	91.3	93.9
	Recall_N	99.4	99.1	81.5	80.2	89.6	88.9
	F1_N	94	95.6	86.3	86.6	90	91
	Precision_M	99.4	99.1	83.1	82.7	89.2	89.1
	Recall_M	86.2	90.6	92	94.7	89.5	92.8
	F1_M	90.9	93.5	87	88	88.8	90.5
DT	Accuracy	95.6	96.3	78.6	78	86.1	86.1
	Precision_N	94	94.8	93.2	93.4	93.6	94.1
	Recall_N	99.8	99.8	64.9	63.1	80.2	79.5
	F1_N	96.4	96.9	75.6	74.5	85.8	85.7
	Precision_M	99.6	99.7	71.9	71.1	81.8	81.4
	Recall_M	91.5	92.7	92.4	92.8	92	92.8
	F1_M	94.2	95.1	80.2	79.9	85.8	86
GNB	Accuracy	81.4	81.3	78.5	80.3	79.7	80.8
	Precision_N	77.9	78.2	80	81.6	78.5	79.3
	Recall_N	97.2	95.7	82.6	83	89.1	88.9
	F1_N	85.4	85.1	80.3	81.5	82.6	83.1
	Precision_M	84.7	87.4	74	78	77.9	81.8
	Recall_M	65.5	66.8	74.3	77.6	70.3	72.6
	F1_M	71.6	73.1	73.8	77.4	73.4	76.2
LR	Accuracy	89.9	91.6	89.4	88.9	89.7	90.1
	Precision_N	91.3	92.4	92.8	93.4	91.7	92.5
	Recall_N	93.2	95	86.4	84.5	89.6	89.4
	F1_N	91.5	93.1	89.2	88.5	90.2	90.5
	Precision_M	88.3	90.4	87.3	85.9	89.1	89
	Recall_M	86.7	88.1	92.5	93.3	89.9	90.9
	F1_M	86	87.9	89.6	89.2	88.9	89.4
SVC	Accuracy	88.2	90	90.6	91	89.6	90.5
	Precision_N	90.3	92.5	92.5	93.5	91.1	92.5
	Recall_N	90.3	91.2	89.1	88.8	89.7	89.9
	F1_N	89.6	91.2	90.5	90.9	90	90.9
	Precision_M	86.9	88.2	89.5	89.4	89.3	89.7
	Recall_M	86	88.8	92.1	93.3	89.5	91.2
	F1_M	85.1	87.2	90.5	91.1	88.9	90

Table B.22: Leave-One-Speaker+Distance-Out Results. The table presents the macro and micro averages across all LOSDO folds for each classifier.

Model	Metrics	LOSDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	77.5	74.7	80.8	78.5	79.3	76.7
	Precision_N	78.7	75.4	77.4	75.1	77.7	75
	Recall_N	82.7	82	93.6	93	88.7	88
	F1_N	79.7	77.6	84	82.3	82.1	80.2
	Precision_M	75.9	73.5	85.9	83.5	80.6	78.1
	Recall_M	72.2	67.4	68	63.9	69.8	65.5
	F1_M	72.8	69	74.2	70.5	73.7	70
QDA	Accuracy	92.6	92.3	81.8	81.1	86.6	86.2
	Precision_N	91.7	91.8	88.6	88.1	89.9	89.8
	Recall_N	99.5	99.3	80.2	79.4	88.9	88.5
	F1_N	94.6	94.4	82.7	82.1	88.3	88
	Precision_M	98.4	98	78.6	77	84.9	83.6
	Recall_M	85.8	85.3	83.3	82.8	84.4	83.9
	F1_M	87.2	86.1	78.9	77.9	82.6	81.7
DT	Accuracy	91.8	91.4	77.8	74.8	84.2	82.4
	Precision_N	91.3	90.5	87.4	85.9	89.3	88.2
	Recall_N	97.9	98.9	71.9	68.1	83.7	82.1
	F1_N	93.7	93.6	77.4	74.1	85.4	83.9
	Precision_M	98	98.9	70.1	66	77.3	74.5
	Recall_M	85.7	84	83.7	81.6	84.6	82.7
	F1_M	86.4	85.1	75.7	72.4	80.2	77.7
GNB	Accuracy	87.1	85.5	79.6	78.8	82.9	81.9
	Precision_N	83.8	82.4	81.5	81.1	82.4	81.6
	Recall_N	98	97.5	83.8	83.2	90.1	89.7
	F1_N	89.6	88.4	81.6	81	85.2	84.5
	Precision_M	86.6	84.1	75.7	73.5	80.4	78.2
	Recall_M	76.1	73.5	75.4	74.4	75.7	74
	F1_M	80.7	78	75	73.4	77.5	75.5
LR	Accuracy	95.3	95.3	84.2	82.9	89.2	88.6
	Precision_N	97.2	96.3	89.6	88.7	92.2	91.5
	Recall_N	93.5	94.4	84	82.5	88.3	87.9
	F1_N	95.2	95.3	85.5	84.2	89.7	89.1
	Precision_M	93.8	94.6	80.9	78.7	88	87.5
	Recall_M	97.1	96.3	84.5	83.2	90.1	89.2
	F1_M	95.4	95.4	81	79.1	88.3	87.5
SVC	Accuracy	94.5	93.6	85.7	83.9	89.6	88.3
	Precision_N	97.2	96.1	89.5	88.1	91.8	90.5
	Recall_N	91.7	91.2	86.7	85.3	89	88
	F1_N	94.3	93.5	87	85.5	90	88.8
	Precision_M	92.2	91.7	83.8	81.6	88.9	87.7
	Recall_M	97.2	96.1	84.7	82.5	90.2	88.7
	F1_M	94.6	93.8	82.5	80	89	87.6

Table B.23: Leave-One-Room+Speaker+Distance-Out Results. The table presents the macro and micro averages across all LORSDO folds for each classifier.

Model	Metrics	LORSDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	68.5	69.3	75.8	77.2	72.6	73.7
	Precision_N	67.2	68.4	71.2	72.7	69.4	70.8
	Recall_N	84.9	85.1	95.9	96.4	91.1	91.4
	F1_N	74.1	74.8	80.9	82.1	78	78.9
	Precision_M	69	68.4	88.8	90	79.4	80.2
	Recall_M	52	53.5	55.6	58.1	54	56
	F1_M	57.4	57.8	65	67	61.7	63
QDA	Accuracy	85.6	88.7	80	81.1	82.5	84.5
	Precision_N	83.9	87.5	84.1	87	84	87.2
	Recall_N	99.5	99.5	87.5	84.9	92.9	91.4
	F1_N	89.6	91.8	83.6	83.8	86.4	87.6
	Precision_M	99.5	99.5	79.2	78.6	85	85
	Recall_M	71.7	77.9	72.4	77.3	72.1	77.6
	F1_M	75.1	80.4	70.5	73.4	72.5	76.4
DT	Accuracy	88.7	89.2	74.4	71.2	80.8	79.3
	Precision_N	85.6	85.8	81.1	78.3	83.9	83.2
	Recall_N	99.5	99.5	69.8	62.7	83	79.1
	F1_N	91.1	91.4	72.3	67	82	79.8
	Precision_M	99.4	99.5	71.6	68.1	80.3	77.8
	Recall_M	78	79	78.9	79.7	78.5	79.4
	F1_M	83.3	84.5	73.1	71.5	77.2	76.6
GNB	Accuracy	74.9	78.7	71	74.8	72.7	76.5
	Precision_N	71.2	75	72	75.9	71.5	75.3
	Recall_N	97.7	96.6	87.1	86.5	91.8	91
	F1_N	81.2	83.4	76.9	79.2	78.9	81.1
	Precision_M	85	86.8	68	71.9	73.1	76.6
	Recall_M	52.1	60.8	54.9	63.2	53.6	62.1
	F1_M	58.8	67	56.4	64	57.5	65.3
LR	Accuracy	87.5	89.1	85	86.1	86.1	87.4
	Precision_N	87.9	89.2	87.3	89.9	87.2	89.2
	Recall_N	91.6	93.1	88.7	87.3	90	89.9
	F1_N	89.1	90.5	86.7	87.4	87.8	88.8
	Precision_M	86.4	88.6	85	84.5	85.3	85.9
	Recall_M	83.3	85.2	81.3	84.8	82.1	85
	F1_M	83.8	86	80.6	82.5	82.2	84.2
SVC	**Accuracy**	86.2	86.9	86.3	87.6	86.3	87.3
	Precision_N	86.3	87.6	87.4	89.8	86.7	88.5
	Recall_N	91.4	90.9	91.3	90.4	91.4	90.6
	F1_N	88.1	88.5	88.2	89.1	88.1	88.7
	Precision_M	85.6	86.1	88	87.7	86.8	86.8
	Recall_M	81	82.9	81.2	84.8	81.1	84
	F1_M	82.5	83.7	82.1	84.1	82.4	84.1

Bibliography

[1] Qualtrics. `https://www.qualtrics.com`, 2024. Location: Provo, Utah, USA.

[2] Obehioye Adubor, Rhomni St. John, and Aaron Steinfeld. Personal safety is more important than cost of damage during robot failure. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 403, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Neziha Akalin and Maria Riveiro. Let me explain why i didn't take the action you wanted! : Comparing different modalities for explanations in human-robot interaction. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 416–421, 2025.

[4] Abdullah I. Al-Shoshan. Speech and music classification and separation: A review. *Journal of King Saud University - Engineering Sciences*, 19(1):95–132, 2006.

[5] Pourya Aliasghari, Moojan Ghafurian, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. Effect of domestic trainee robots' errors on human teachers' trust. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, pages 81–88, 2021.

[6] Rosa Ma Alsina-Pagès, Joan Navarro, Francesc Alías, and Marcos Hervás. homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. *Sensors*, 17(4), 2017.

[7] Jakob Ambsdorf, Alina Munir, Yiyao Wei, Klaas Degkwitz, Harm Matthias Harms, Susanne Stannek, Kyra Ahrens, Dennis Becker, Erik Strahl, Tom Weber, and Stefan Wermter. Explain yourself! effects of explanations in human-robot interaction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 393–400, 2022.

[8] Dorsa Amir, Richard E Ahl, William Shelby Parsons, and Katherine McAuliffe. Children are more forgiving of accidental harms across development. *Journal of Experimental Child Psychology*, 205:105081, 2021.

[9] Daniel R Anderson and Tiffany A Pempek. Television and very young children. *American behavioral scientist*, 48(5):505–522, 2005.

[10] Naeimeh Anzabi, Anahita Etemad, and Hiroyuki Umemuro. Exploring the effects of self-disclosed backstory of social robots on development of trust in human-robot interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 431–435, 2023.

[11] Theo Araujo. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in human behavior*, 85:183–189, 2018.

[12] Elliot Aronson, Ben Willerman, and Joanne Floyd. The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science*, 4(6):227–228, 1966.

[13] Alexander M. Aroyo, Dario Pasquali, Austin Kothig, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Expectations vs. reality: Unreliability and transparency in a treasure hunt game with icub. *IEEE Robotics and Automation Letters*, 6(3):5681–5688, 2021.

[14] Chatchalita Asavanant and Hiroyuki Umemuro. Personal space violation by a robot: An application of expectation violation theory in human-robot interaction. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, pages 1181–1188, 2021.

[15] Francisco J Ayala. The difference of being human: Morality. *Proceedings of the National Academy of Sciences*, 107(supplement_2):9015–9022, 2010.

[16] Sumair Aziz, Muhammad Awais, Tallha Akram, Umar Shahbaz Khan, Musaed A. Alhussein, and Khursheed Aurangzeb. Automatic scene recognition through acoustic classification for behavioral robotics. *Electronics*, 8, 2019.

[17] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Trans. Interact. Intell. Syst.*, 8(4), November 2018.

[18] Karen Banai and Rachel Yifat. Perceptual anchoring in preschool children: Not adultlike, but there. *PLoS One*, 6(5):e19769, 2011.

[19] Jaime Banks. Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, 13(8):2021–2038, 2021.

[20] Jon Philip Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. *Interspeech*, 2018.

[21] Rachel Barr. Transfer of learning between 2d and 3d sources during infancy: Informing theory and practice. *Developmental review*, 30(2):128–154, 2010.

[22] Roberto Basili, Alfredo Serafini, and Armando Stellato. Classification of musical genre: a machine learning approach. *ISMIR*, 2004.

[23] Roy F Baumeister. Free will, consciousness, and cultural animals. *Are we free*, pages 65–85, 2008.

[24] Tanya Behne, Malinda Carpenter, Josep Call, and Michael Tomasello. Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology*, 41(2):328–337, 2005.

[25] Debra Bernstein and Kevin Crowley. Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences*, 17(2):225–247, 2008.

[26] Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391, 2017.

[27] Yochanan E Bigman and Kurt Gray. People are averse to machines making moral decisions. *Cognition*, 181:21–34, 2018.

[28] Yochanan E Bigman, Adam Waytz, Ron Alterovitz, and Kurt Gray. Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences*, 23(5):365–368, 2019.

[29] Logan Blue, Luis Vargas, and Patrick Traynor. Hello, is it me you're looking for? differentiating between human and electronic speakers for voice interface security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 123–133, 2018.

[30] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.

[31] Jean-François Bonnefon, Iyad Rahwan, and Azim Shariff. The moral psychology of artificial intelligence. *Annual review of psychology*, 75(1):653–675, 2024.

[32] Celina K Bowman-Smith, Charlotte Aitken, Thuvaraka Mahenthiran, Edith Law, and Elizabeth S Nilsen. Teaching social robots: the effect of robot mistakes on children's learning-through-teaching. *Frontiers in Developmental Psychology*, 3:1526486, 2025.

[33] C. Breazeal and B. Scassellati. How to build robots that make friends and influence people. In *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, volume 2, pages 858–863 vol.2, 1999.

[34] Cynthia Breazeal, Paul L Harris, David DeSteno, Jacqueline M Kory Westlund, Leah Dickens, and Sooyeon Jeong. Young children treat robots as informants. *Topics in cognitive science*, 8(2):481–491, 2016.

[35] L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. 1984.

[36] Kimberly A Brink, Kurt Gray, and Henry M Wellman. Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child development*, 90(4):1202–1214, 2019.

[37] Kimberly A Brink and Henry M Wellman. Robot teachers for children? young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology*, 56(7):1268, 2020.

[38] Daniel J Brooks. *A human-centric approach to autonomous robot failures*. PhD thesis, University of Massachusetts Lowell, 2017.

[39] Daniel J Brooks, Momotaz Begum, and Holly A Yanco. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 487–492. IEEE, 2016.

[40] Johannes Bührke, Alfred Benedikt Brendel, Sascha Lichtenberg, Maike Greve, and Milad Mirbabaie. Is making mistakes human? on the perception of typing errors in chatbot communication. 2021.

[41] Harriet R Cameron, Simon Castle-Green, Muhammad Chughtai, Liz Dowthwaite, Ayse Kucukyilmaz, Horia A Maior, Victor Ngo, Eike Schneiders, and Bernd C Stahl. A taxonomy of domestic robot failure outcomes: Understanding the impact of failure on trustworthiness of domestic robots. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–14, 2024.

[42] Kate Candon, Nicholas C. Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. React: Two datasets for analyzing both human reactions and evaluative feedback to robots over time. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 885–889, New York, NY, USA, 2024. Association for Computing Machinery.

[43] Jennifer Carlson and Robin R Murphy. How ugvs physically fail in the field. *IEEE Transactions on robotics*, 21(3):423–437, 2005.

[44] Malinda Carpenter, Josep Call, and Michael Tomasello. Twelve- and 18-month-olds copy actions in terms of goals. *Developmental Science*, 8(1):F13–F20, January 2005.

[45] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), may 2011.

[46] Xiaoyu Chang, Yanheng Li, Sijia Liu, Ling Ma, and Ray Lc. "sorry to keep you waiting": Recovering from negative consequences resulting from service robot unintended rejection. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 96–105, New York, NY, USA, 2024. Association for Computing Machinery.

[47] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. Bathroom activity monitoring based on sound. In Hans W. Gellersen, Roy Want, and Albrecht Schmidt, editors, *Pervasive Computing*, pages 47–61, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[48] Sonia Chernova and Andrea Thomaz. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8:1–121, 04 2014.

[49] Nadia Chernyak and Heather E Gary. Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. In *Young children's developing understanding of the biological world*, pages 73–90. Routledge, 2019.

[50] Vivienne Bihe Chi and Bertram F. Malle. Interactive human-robot teaching recovers and builds trust, even with imperfect learners. In *2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 127–136, 2024.

[51] Sungwoo Choi, Anna S Mattila, and Lisa E Bolton. To err is human (-oid): how do consumers react to robot service failure and recovery? *Journal of Service Research*, 24(3):354–371, 2021.

[52] Dimitri A. Christakis, Frederick J. Zimmerman, David L. DiGiuseppe, and Carolyn A. McCarty. Early television exposure and subsequent attentional problems in children. *Pediatrics*, 113(4):708–713, 04 2004.

[53] Selina M. Chu, Shrikanth S. Narayanan, C.-C. Jay Kuo, and Maja J. Matarić. 'where am i? scene recognition for mobile robots using audio features. *IEEE International harry zhan on Multimedia and Expo*, pages 885–888, 2006.

[54] Houston Claure, Inyoung Shin, J. Gregory Trafton, and Marynel Vázquez. Did the robot really intend to harm me? the effect of perceived agency and intention on fairness judgments. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 889–898, 2025.

[55] David L Coker Jr and Kristen D Ritchey. Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children*, 76(2):175–193, 2010.

[56] Judith H Danovitch and Frank C Keil. Young humeans: The role of emotions in children's evaluation of moral reasoning abilities. *Developmental Science*, 11(1):33–39, 2008.

[57] Kate Darling, Palash Nandy, and Cynthia Breazeal. Empathic concern and the effect of stories in human-robot interaction. In *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 770–775. IEEE, 2015.

[58] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 351–360, 2021.

[59] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258. IEEE, 2013.

[60] Vladimir Despotovic, Peter Pocta, and Andrej Zgank. Audio-based active and assisted living: A review of selected applications and future trends. *Computers in Biology and Medicine*, 149:106027, 2022.

[61] Alexander P D'Esterre, Michael T Rizzo, and Melanie Killen. Unintentional and intentional falsehoods: The role of morally relevant theory of mind. *Journal of Experimental Child Psychology*, 177:53–69, 2019.

[62] Ha Do, Weihua Sheng, and Meiqin Liu. Human-assisted sound event recognition for home service robots. *Robotics and Biomimetics*, 3, 12 2016.

[63] Ha Manh Do, Minh Pham, Weihua Sheng, Dan Yang, and Meiqin Liu. Rish: A robot-integrated smart home for elderly care. *Robotics and Autonomous Systems*, 101:74–92, 2018.

[64] Sabine Doebel and Melissa A Koenig. Children's use of moral behavior in selective trust: Discrimination versus learning. *Developmental psychology*, 49(3):462, 2013.

[65] Bo Dong, Cristian Lumezanu, Yuncong Chen, Dongjin Song, Takehiko Mizoguchi, Haifeng Chen, and Latifur Khan. At the speed of sound: Efficient audio scene classification. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 301–305, New York, NY, USA, 2020. Association for Computing Machinery.

[66] Hazım Kemal Ekenel and Tomas Semela. Multimodal genre classification of tv programs and youtube videos. *Multimedia Tools and Applications*, 63:547–567, 2013.

[67] Connor Esterwood and Lionel P Robert. Having the right attitude: How attitude impacts trust repair in human—robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 332–341. IEEE, 2022.

[68] Connor Esterwood and Lionel P Robert. A literature review of trust repair in hri. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, pages 1641–1646. IEEE, 2022.

[69] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, jun 2008.

[70] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[71] Brian Fiala, Adam Arico, and Shaun Nichols. You, robot. In *Current controversies in experimental philosophy*, pages 31–47. Routledge, 2014.

[72] Tim Fischer, Johannes Schneider, and Wilhelm Stork. Classification of breath and snore sounds using audio data recorded with smartphones in the home environment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 226–230, 2016.

[73] Cliff Fitzgerald. Developing baxter. In *2013 IEEE conference on technologies for practical robot applications (TePRA)*, pages 1–6. IEEE, 2013.

[74] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, 57:238, 1989.

[75] Teresa Flanagan, Nicholas C Georgiou, Brian Scassellati, and Tamar Kushnir. School-age children are more skeptical of inaccurate robots than adults. *Cognition*, 249:105814, 2024.

[76] Teresa Flanagan, Joshua Rottman, and Lauren H Howard. Constrained choice: Children's and adults' attribution of choice to a humanoid robot. *Cognitive Science*, 45(10):e13043, 2021.

[77] Teresa Flanagan, Gavin Wong, and Tamar Kushnir. The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Developmental psychology*, 59(6):1017, 2023.

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

[79] Matija Franklin, Hal Ashton, Edmond Awad, and David Lagnado. Causal framework of artificial autonomous agent responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 276–284, New York, NY, USA, 2022. Association for Computing Machinery.

[80] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[81] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. Attributing blame to robots: I. the influence of robot autonomy. *Human factors*, 63(4):592–602, 2021.

[82] Susan R Fussell, Sara Kiesler, Leslie D Setlock, and Victoria Yew. How people anthropomorphize robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 145–152, 2008.

[83] Cecilia Gabriela Morales Garza. Failure is an option: How the severity of robot errors affects human-robot interaction. *Pittsburgh: Carnegie Mellon University*, 2018.

[84] Denise Y Geiskkovitch, Raquel Thiessen, James E Young, and Melanie R Glenwright. What? that's not a chair!: How robot informational errors affect children's trust towards robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 48–56. IEEE, 2019.

[85] Nicholas C Georgiou, Rebecca Ramnauth, Emmanuel Adeniran, Michael Lee, Lila Selin, and Brian Scassellati. Is someone there or is that the tv? detecting social presence using sound. *ACM Transactions on Human-Robot Interaction*, 12(4):1–33, 2023.

[86] Nicholas C Georgiou, Shuangge Wang, Joel Banks, Kate Candon, Drazen Brscic, and Brian Scassellati. When teaching a robot, people employ different feedback strategies: Some are more effective than others. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[87] György Gergely, Harold Bekkering, and Ildikó Király. Rational imitation in preverbal infants. *Nature*, 415(6873):755–755, February 2002.

[88] Romi Gideoni, Shanee Honig, and Tal Oron-Gilad. Is it personal? the impact of personally relevant robotic failures (perfs) on humans' trust, likeability, and willingness to use the robot. *International journal of social robotics*, 16(6):1049–1067, 2024.

[89] Lauren N Girouard-Hallam and Judith H Danovitch. Children's trust in and learning from voice assistants. *Developmental Psychology*, 58(4):646, 2022.

[90] Lauren N Girouard-Hallam, Yu Tong, Fuxing Wang, and Judith H Danovitch. What can the internet do?: Chinese and american children's attitudes and beliefs about the internet. *Cognitive Development*, 66:101338, 2023.

[91] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology*, 6:931, 2015.

[92] Sabine Gless, Emily Silverman, and Thomas Weigend. If robots cause harm, who is to blame? self-driving cars and criminal liability. *New Criminal Law Review*, 19(3):412–436, 2016.

[93] Lacrimioara Grama and Corneliu Rusu. Adding audio capabilities to tiago service robot. In *2018 International Symposium on Electronics and Telecommunications (ISETC)*, pages 1–4, 2018.

[94] 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.

[95] Kurt Gray and Daniel M Wegner. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130, 2012.

[96] Haley N Green, Md Mofijul Islam, Shahira Ali, and Tariq Iqbal. Who's laughing nao? examining perceptions of failure in a humorous robot partner. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 313–322. IEEE, 2022.

[97] Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.

[98] Hyowon Gweon. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in cognitive sciences*, 25(10):896–910, 2021.

[99] J.A. Haigh and J.S. Mason. Robust voice activity detection using cepstral features. In *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, volume 3, pages 321–324 vol.3, 1993.

[100] Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. "an error occurred!" - trust repair with virtual robot using levels of mistake explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 218–226, New York, NY, USA, 2021. Association for Computing Machinery.

[101] Paul L Harris and Kathleen H Corriveau. Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1179–1187, 2011.

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

[103] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell.*, 27:83–85, 11 2004.

[104] Shanee Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861, 2018.

[105] Jessica S Horst and Michael C Hout. The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, 48(4):1393–1409, 2016.

[106] Jindan Huang, Reuben M Aronson, and Elaine Schaertl Short. Modeling variation in human feedback with user inputs: An exploratory methodology. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 303–312, 2024.

[107] Vikram K Jaswal, A Carrington Croft, Alison R Setia, and Caitlin A Cole. Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, 21(10):1541–1547, 2010.

[108] Seok-Hoon Kim Jeong-Sik Park. Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices. *Intelligent Automation & Soft Computing*, 26(1):149–159, 2020.

[109] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. *In Multimedia and Expo, 2002. ICME'02*, 1:113–116, 2002.

[110] Jennifer L Jipson and Susan A Gelman. Robots and rodents: Children's inferences about living and nonliving kinds. *Child development*, 78(6):1675–1688, 2007.

[111] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Nathan G Freier, Rachel L Severson, Brian T Gill, Jolina H Ruckert, and Solace Shen. "robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2):303, 2012.

[112] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather E Gary, Aimee L Reichert, Nathan G Freier, and Rachel L Severson. Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 33–40, 2012.

[113] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather E Gary, Aimee L Reichert, Nathan G Freier, and Rachel L Severson. Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 33–40, 2012.

[114] Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. "what if it is wrong": Effects of power dynamics and trust repair strategy on trust and compliance in hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 271–280, New York, NY, USA, 2023. Association for Computing Machinery.

[115] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[116] Khaled Kassem, Patrick Gietl, Florian Michahelles, and Andrii Matviienko. Roboteach: How student robots' preexisting proficiency and learning rate affect human teachers demonstrating object placement. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[117] Merel Keijsers, Hussain Kazmi, Friederike Eyssel, and Christoph Bartneck. Teaching robots a lesson: determinants of robot punishment. *International Journal of Social Robotics*, 13:41–54, 2021.

[118] Charles C Kemp, Aaron Edsinger, Henry M Clever, and Blaine Matulevich. The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3150–3157. IEEE, 2022.

[119] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. Social robot tutoring for child second language learning. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 231–238. IEEE, 2016.

[120] Taylor A. Kessler Faulkner, Elaine Schaertl Short, and Andrea L. Thomaz. Interactive reinforcement learning with inaccurate feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7498–7504, 2020.

[121] Parag Khanna. Adapting robotic explanations for robotic failures in human robot collaboration. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1863–1865. IEEE, 2025.

[122] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. Effects of explanation strategies to resolve failures in human-robot collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1829–1836. IEEE, 2023.

[123] Zahra Rezaei Khavas, Monish Reddy Kotturu, Reza Azadeh, and Paul Robinette. Do humans have different expectations regarding humans and robots'

morality?*. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1126–1133, 2024.

[124] Zahra Rezaei Khavas, Amin Majdi, S. Reza Ahmadzadeh, and Paul Robinette. Human trust after drone failure: Study of the effects of drone type and failure type on human-drone trust. In *2023 20th International Conference on Ubiquitous Robots (UR)*, pages 685–692, 2023.

[125] Hyun-Don Kim, Jinsung Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1705–1711, 2008.

[126] Taemie Kim and Pamela Hinds. Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 80–85, 2006.

[127] Taenyun Kim and Hayeon Song. How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:101595, 2021.

[128] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[129] W. Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE International Conference on Development and Learning*, pages 292–297, 2008.

[130] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[131] Melissa A Koenig and Paul L Harris. Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, 76(6):1261–1277, 2005.

[132] Melissa A Koenig and Paul L Harris. The role of social cognition in early trust. *Trends in Cognitive Sciences*, 9(10):457–459, 2005.

[133] Melissa A Koenig and Amanda L Woodward. Sensitivity of 24-month-olds to the prior inaccuracy of the source: possible mechanisms. *Developmental psychology*, 46(4):815, 2010.

[134] Takanori Komatsu. How do people judge moral wrongness in a robot and in its designers and owners regarding the consequences of the robot's behaviors? In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1168–1171. IEEE, 2016.

[135] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. Behavioural responses to robot conversational failures. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 53–62, 2020.

[136] Dimosthenis Kontogiorgos and Julie Shah. Questioning the robot: Using human non-verbal cues to estimate the need for explanations. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '25, page 717–728. IEEE Press, 2025.

[137] Dimosthenis Kontogiorgos, Sanne Van Waveren, Olle Wallberg, Andre Pereira, Iolanda Leite, and Joakim Gustafson. Embodiment effects in interactions with failing robots. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[138] Tobias Kopp, Marco Baumgartner, and Steffen Kinkel. How linguistic framing affects factory workers' initial trust in collaborative robots: The interplay between anthropomorphism and technological replacement. *International Journal of Human-Computer Studies*, 158:102730, 2022.

[139] Jacqueline M Kory-Westlund and Cynthia Breazeal. Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. *Frontiers in Robotics and AI*, 6:54, 2019.

[140] Konstantinos Koutroumbas and Sergios Theodoridis. *Pattern recognition*. Academic Press, 2008.

[141] Esther S. Kox, Juul van den Boogaard, Vesa Turjaka, and José H. Kerstholt. The journey or the destination: The impact of transparency and goal attainment on trust in human-robot teams. *J. Hum.-Robot Interact.*, 14(2), December 2024.

[142] Ken'ichi Kumatani, Sankaran Panchapagesan, Minhua Wu, Minjae Kim, Nikko Strom, Gautam Tiwari, and Arindam Mandal. Direct modeling of raw audio with dnns for wake word detection. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 252–257, 2017.

[143] Tamar Kushnir and Melissa A Koenig. What i don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology*, 53(5):826–835, 2017.

[144] Tamar Kushnir, Christopher Vredenburgh, and Lauren A. Schneider. "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, 49(3):446–453, 2013.

[145] Veton Këpuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). *2018*

IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018.

[146] Michael Laakasuo. Moral uncanny valley revisited – how human expectations of robot morality based on robot appearance moderate the perceived morality of robot decisions in high conflict moral dilemmas. *Frontiers in Psychology*, 14, 2023.

[147] Michael Laakasuo, Jussi Palomäki, and Nils Köbis. Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, 13(7):1679–1688, 2021.

[148] Ali Ladak, Steve Loughnan, and Matti Wilks. The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, 33(1):27–34, 2024.

[149] Asheley R Landrum, Baxter S Eaves, and Patrick Shafto. Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, 19(3):109–111, 2015.

[150] Elizabeth Lapidow, Tushita Tandon, Mariel Goddu, and Caren M Walker. A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, zoom, and lookit methodologies. *Frontiers in psychology*, 12:731404, 2021.

[151] K Lee, L Smith, A Dragan, and P Abbeel. B-pref: Benchmarking preference-based reinforcement learning. *Neural Information Processing Systems (NeurIPS)*, 2021.

[152] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210. IEEE, 2010.

[153] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. People may punish, but not blame robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[154] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children. *International Journal of Social Robotics*, 6:329–341, 2014.

[155] Iolanda Leite, André Pereira, Carlos Martinho, and Ana Paiva. Are emotional robots more fun to play with? In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 77–82. IEEE, 2008.

[156] Séverin Lemaignan, Julia Fink, Francesco Mondada, and Pierre Dillenbourg. You're doing it wrong! studying unexpected behaviors in child-robot interaction. In *International conference on social robotics*, pages 390–400. Springer, 2015.

[157] Gregory LeMasurier, Alvika Gautam, Zhao Han, Jacob W Crandall, and Holly A Yanco. Reactive or proactive? how robots should explain failures. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 413–422, 2024.

[158] Gregory LeMasurier, Christian Tagliamonte, Jacob Breen, Daniel Maccaline, and Holly A. Yanco. Templated vs. generative: Explaining robot failures. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1346–1353, 2024.

[159] Xuying Leo and Young Eun Huh. Who gets the blame for service failures? attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior*, 113:106520, 2020.

[160] Xiaoqian Li and W Quin Yow. Younger, not older, children trust an inaccurate human informant more than an inaccurate robot informant. *Child Development*, 95(3):988–1000, 2024.

[161] Gabriel Lima, Meeyoung Cha, Chihyung Jeon, and Kyung Sin Park. The conflict between people's urge to punish ai and legal systems. *Frontiers in Robotics and AI*, 8:756242, 2021.

[162] Jirachaya Fern Limprayoon, Nicholas C. Georgiou, Natnaree Proud Ua-Arak, and Brian Scassellati. The effects of a gossiping robot on team cohesion. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 2066–2071, 2024.

[163] Shannon Liu, Maria Teresa Parreira, and Wendy Ju. "i'm done": Describing human reactions to successive robot failure. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1458–1462, 2025.

[164] Robert Loftin, James MacGlashan, Bei Peng, Matthew Taylor, Michael Littman, Jeff Huang, and David Roberts. A strategy-aware technique for learning behaviors from discrete human feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[165] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30:30–59, 2016.

[166] Beth Logan. Mel frequency cepstral coefficients for music modeling. 2000.

[167] Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 344–351, New York, NY, USA, 2018. Association for Computing Machinery.

[168] Matthew B Luebbers, Aaquib Tabrez, Kanaka Samagna Talanki, and Bradley Hayes. Recency bias in task performance history affects perceptions of robot competence and trustworthiness. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11274–11280. IEEE, 2024.

[169] Joseph B. Lyons, Izz aldin Hamdan, and Thy Q. Vo. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, 138:107473, 2023.

[170] Joseph B. Lyons, Izz aldin Hamdan, and Thy Q. Vo. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior*, 138:107473, 2023.

[171] Lili Ma and Patricia A Ganea. Dealing with conflicting information: Young children's reliance on what they see versus what they are told. *Developmental Science*, 13(1):151–160, 2010.

[172] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2285–2294. JMLR.org, 2017.

[173] Akihiro Maehigashi, Kenta Kubo, Yun Nungduk, and Seiji Yamada. Effects of robot bowing during apology on trust repair. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1478–1482. IEEE, 2025.

[174] Christof Mahieu, Femke Ongenae, Femke De Backere, Pieter Bonte, Filip De Turck, and Pieter Simoens. Semantics-based platform for context-aware and personalized robot interaction in the internet of robotic things. *Journal of Systems and Software*, 149:138–157, 2019.

[175] 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 the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 117–124, New York, NY, USA, 2015. Association for Computing Machinery.

[176] Bertram F. Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. Which robot am i thinking about? the impact of action and appearance on people's evaluations of a moral robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, page 125–132. IEEE Press, 2016.

[177] Thanassis Mavropoulos, Georgios Meditskos, Spyridon Symeonidis, Eleni Kamateri, Maria Rousi, Dimitris Tzimikas, Lefteris Papageorgiou, Christos Eleftheriadis, George Adamopoulos, Stefanos Vrochidis, et al. A context-aware conversational agent in the rehabilitation domain. *Future Internet*, 11(11):231, 2019.

[178] J. Maxime, X. Alameda-Pineda, L. Girin, and R. Horaud. 'sound representation and classification benchmark for domestic robots. *IEEE International harry zhang on Robotics and Automation (ICRA)*, pages 6285–6292, 2014.

[179] Carolynn E McElroy, Caroline M Kelsey, Janine Oostenbroek, and Amrisha Vaish. Beyond accidents: Young children's forgiveness of third-party intentional transgressors. *Journal of Experimental Child Psychology*, 228:105607, 2023.

[180] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. 2015.

[181] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Maria Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. pages 274–278, 10 2020.

[182] Douglas L Medin, Salino G García, et al. Conceptualizing agency: Folkpsychological and folkcommunicative perspectives on plants. *Cognition*, 162:103–123, 2017.

[183] Katherine Metcalf, Miguel Sarabia, Masha Fedzechkina, and Barry-John Theobald. Can you rely on synthetic labellers in preference-based reinforcement learning? it's complicated. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10128–10136, Mar. 2024.

[184] Candice M Mills and Fadwa B Elashi. Children's skepticism: Developmental and individual differences in children's ability to detect and explain distorted claims. *Journal of experimental child psychology*, 124:1–17, 2014.

[185] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4:21, 2017.

[186] Nobuyuki Miyake, Tetsuya Takiguchi, and Yasuo Ariki. Noise detection and classification in speech signals with boosting. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 778–782, 2007.

[187] M. H. Moattar and M. M. Homayounpour. A simple but efficient real-time voice activity detection algorithm. In *2009 17th European Signal Processing Conference*, pages 2549–2553, 2009.

[188] Cecilia G Morales, Elizabeth J Carter, Xiang Zhi Tan, and Aaron Steinfeld. Interaction needs and opportunities for failing robots. In *Proceedings of the 2019 on designing interactive systems conference*, pages 659–670, 2019.

[189] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4):3005–3054, aug 2022.

[190] Javier Movellan, Micah Eckhardt, Marjo Virnes, and Angelica Rodriguez. Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 307–308, 2009.

[191] Lauren J Myers, Emily Crawford, Claire Murphy, Edoukou Aka-Ezoua, and Christopher Felix. Eyes in the room trump eyes on the screen: Effects of a responsive co-viewer on toddlers' responses to and learning from video chat. *Journal of Children and Media*, 12(3):275–294, 2018.

[192] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. *In Proceedings of the International harry zhang on Music Information Retrieval (ISMIR)*, 2011.

[193] Eddy Nahmias, Corey Hill Allen, and Bradley Loveall. When do robots have free will? exploring the relationships between (attributions of) consciousness and free will. *Free will, causality, and neuroscience*, 338:57–80, 2020.

[194] Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, 18(5):561–584, 2005.

[195] Marco Nalin, Ilaria Baroni, Alberto Sanna, and Clara Pozzi. Robotic companion for diabetic children: emotional and educational support to diabetic children, through an interactive robot. In *Proceedings of the 11th international conference on interaction design and children*, pages 260–263, 2012.

[196] Andreas Naoum, Parag Khanna, Elmira Yadollahi, Mårten Björkman, and Christian Smith. Adapting robot's explanation for failures based on observed human behavior in human-robot collaboration. *arXiv preprint arXiv:2504.09717*, 2025.

[197] Birthe Nesset, Marta Romeo, Gnanathusharan Rajendran, and Helen Hastie. Robot broken promise? repair strategies for mitigating loss of trust for repeated failures. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1389–1395, 2023.

[198] Mark Nielsen, Gabrielle Simcock, and Linda Jenkins. The effect of social engagement on 24-month-olds' imitation from live and televised models. *Developmental science*, 11(5):722–731, 2008.

[199] Sari RR Nijssen, Barbara CN Müller, Tibor Bosse, and Markus Paulus. Can you count on a calculator? the role of agency and affect in judgments of robots as moral agents. *Human–Computer Interaction*, 38(5-6):400–416, 2023.

[200] Cara O'Brien, Molly O'Mara, Johann Issartel, and Conor McGinn. Exploring the design space of therapeutic robot companions for children. In *proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 243–251, 2021.

[201] Linda Onnasch and Clara Laudine Hildebrandt. Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(1):1–24, 2021.

[202] Linda Onnasch and Eileen Roesler. Anthropomorphizing robots: The effect of framing in human-robot collaboration. In *Proceedings of the human factors and ergonomics Society annual meeting*, volume 63, pages 1311–1315. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[203] Janine Oostenbroek and Amrisha Vaish. The emergence of forgiveness in young children. *Child development*, 90(6):1969–1986, 2019.

[204] Bauyrzhan Ospan, Nawaz Khan, Juan Augusto Wrede, Mario Quinde, and Kenzhegali Nurgaliyev. Context aware virtual assistant with case-based conflict resolution in multi-user smart home environment. *International Conference on Computing and Network Communications (CoCoNet)*, pages 36–44, 2018.

[205] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

[206] Sharnil Pandya and Hemant Ghayvat. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. *Advanced Engineering Informatics*, 47:101238, 2021.

[207] Maria Teresa Parreira, Isabel Neto, Filipa Rocha, and Wendy Ju. Calling for backup: How children navigate successive robot communication failures. *arXiv preprint arXiv:2601.00754*, 2026.

[208] Elisabeth S. Pasquini, Kathleen H. Corriveau, Melissa Koenig, and Paul L. Harris. Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43(5):1216–1226, 2007.

[209] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles

Louppe. E.scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[210] Héctor Peiteado, Inmaculada Hernáez, Artzai Picon, Javier Camarena, and Eva Navas. Audio classification techniques in home environments for elderly/dependant people. pages 320–323, 07 2010.

[211] Babiche L. Pompe, Ella Velner, and Khiet P. Truong. The robot that showed remorse: Repairing trust with a genuine apology. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 260–265, 2022.

[212] T. Lakshmi Priya, N.R. Raajan, N. Raju, P. Preethi, and S. Mathini. Speech and non-speech identification and classification using knn algorithm. *Procedia Engineering*, 38:952–958, 2012. INTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING.

[213] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O. Arras. Errare humanum est: Erroneous robots in human-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 501–506, 2016.

[214] Reza Rawassizadeh, Taylan Sen, Sunny Jung Kim, Christian Meurisch, Hamidreza Keshavarz, Max Mühlhäuser, and Michael Pazzani. Manifestation of virtual assistants and robots into daily life: Vision and challenges. *CCF Transactions on Pervasive Computing and Interaction*, 1:163–174, 2019.

[215] Madeline G Reinecke, Matti Wilks, and Paul Bloom. Developmental changes in perceived moral standing of robots. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 43, 2021.

[216] Zahra Rezaei Khavas, Monish Reddy Kotturu, S. Reza Ahmadzadeh, and Paul Robinette. Do humans trust robots that violate moral trust? *J. Hum.-Robot Interact.*, 13(2), June 2024.

[217] Michael T Rizzo, Leon Li, Amanda R Burkholder, and Melanie Killen. Lying, negligence, or lack of knowledge? children's intention-based moral reasoning about resource claims. *Developmental psychology*, 55(2):274, 2019.

[218] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 101–108. IEEE, 2016.

[219] Eileen Roesler. Anthropomorphic framing and failure comprehensibility influence different facets of trust towards industrial robots. *Frontiers in Robotics and AI*, 10:1235017, 2023.

[220] Eileen Roesler, Johannes Pickl, and Felix W Siebert. Investigating the impact of anthropomorphic framing and product value on user acceptance of delivery robots. In *International Conference on Human-Computer Interaction*, pages 347–357. Springer, 2023.

[221] Samuel Ronfard and Jonathan D Lane. Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child development*, 89(2):414–429, 2018.

[222] Samuel Ronfard and Jonathan D Lane. Children's and adults' epistemic trust in and impressions of inaccurate informants. *Journal of experimental child psychology*, 188:104662, 2019.

[223] Holly Rosenkrantz. What should a first grader know? *U.S. News*, 2021.

[224] Andres Rosero, Elizabeth Dula, Harris Kelly, Bertram F. Malle, and Elizabeth K. Phillips. Human perceptions of social robot deception behaviors: an exploratory analysis. *Frontiers in Robotics and AI*, 11, 2024.

[225] Alessandra Rossi, Antonio Andriella, Silvia Rossi, Carme Torras, and Guillem Alenyà. Evaluating the effect of theory of mind on people's trust in a faulty robot. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 477–482, 2022.

[226] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn*, 9(1):137–154, 2018.

[227] Diane N Ruble, Ann K Boggiano, Nina S Feldman, and Judith H Loebl. Developmental analysis of the role of social comparison in self-evaluation. *Developmental Psychology*, 16(2):105, 1980.

[228] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013.

[229] Lydia Paulin Schidelko, Britta Schünemann, Hannes Rakoczy, and Marina Proft. Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, 12:703238, 2021.

[230] Eric Schniter, Roman M Sheremeta, and Daniel Sznycer. Building and rebuilding trust with promises and apologies. *Journal of Economic Behavior & Organization*, 94:242–256, 2013.

[231] Laura E Schulz and Elizabeth Baraff Bonawitz. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, 43(4):1045, 2007.

[232] Sofia Serholt and Wolmet Barendregt. Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. In *Proceedings of the 9th nordic conference on human-computer interaction*, pages 1–10, 2016.

[233] Madison R. Shippy, Brian J. Zhang, and Naomi T. Fitter. Oh &$#%! how do people feel about robots that leverage profanity? In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1873–1880, 2025.

[234] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226. IEEE, 2010.

[235] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. No fair!! an interaction with a cheating robot. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '10, page 219–226. IEEE Press, 2010.

[236] Debra Slocum, Alfred Allan, and Maria M. Allan. An emerging theory of apology. *Australian Journal of Psychology*, 63(2):83–92, 2011.

[237] Craig E Smith, Deborah Anderson, and Anna Straussberger. Say you're sorry: Children distinguish between willingly given and coerced expressions of remorse. *Merrill-Palmer Quarterly*, 64(2):275–308, 2018.

[238] Craig E Smith, Diyu Chen, and Paul L Harris. When the happy victimizer says sorry: Children's understanding of apology and emotion. *British Journal of Developmental Psychology*, 28(4):727–746, 2010.

[239] Craig E Smith and Paul L Harris. He didn't want me to feel sad: Children's reactions to disappointment and apology. *Social Development*, 21(2):215–228, 2012.

[240] David M Sobel and Tamar Kushnir. Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, 120(4):779, 2013.

[241] Kristyn Sommer, Rebecca Davidson, Kristy L Armitage, Virginia Slaughter, Janet Wiles, and Mark Nielsen. Preschool children overimitate robots, but do so less than they overimitate humans. *Journal of Experimental Child Psychology*, 191:104702, 2020.

[242] Kai-Tai Song, Meng-Ju Han, and Shih-Chieh Wang. Speech signal-based emotion recognition and its application to entertainment robots. *Journal of the Chinese Institute of Engineers*, 37(1):14–25, 2014.

[243] Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

[244] A G Starr, R J Wynne, and I Kennedy. Failure analysis of mature robots in automated production. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 213(8):813–824, 1999.

[245] Gerald Steinbauer. A survey about faults of robots used in robocup. In *RoboCup 2012: Robot Soccer World Cup XVI*, volume 7500 of *Lecture Notes in Computer Science*, pages 344–355. ., 2012. RoboCup International Symposium : RoboCup 2012 ; Conference date: 24-06-2012.

[246] Maia Stiber and Chien-Ming Huang. Not all errors are created equal: Exploring human responses to robot errors with varying severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 97–101, 2020.

[247] Deborah J Stipek and Denise H Daniels. Declining perceptions of competence: A consequence of changes in the child or in the educational environment? *Journal of Educational Psychology*, 80(3):352, 1988.

[248] Rebecca Stower, Anna Gautier, Maciej Wozniak, Patric Jensfelt, Jana Tumova, and Iolanda Leite. Take a chance on me: How robot performance and risk behaviour affects trust and risk-taking. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 391–399, 2025.

[249] Rebecca Stower, Arvid Kappas, and Kristyn Sommer. When is it right for a robot to be wrong? children trust a robot over a human in a selective trust task. *Computers in Human Behavior*, 157:108229, 2024.

[250] Gabrielle A Strouse and Patricia A Ganea. Toddlers' word learning and transfer from electronic and print books. *Journal of experimental child psychology*, 156:129–142, 2017.

[251] Gabrielle A Strouse and Jennifer E Samson. Learning from video: A meta-analysis of the video deficit in children ages 0 to 6 years. *Child development*, 92(1):e20–e38, 2021.

[252] Gabrielle A Strouse, Georgene L Troseth, Katherine D O'Doherty, and Megan M Saylor. Co-viewing supports toddlers' word learning from contingent and non-contingent video. *Journal of experimental child psychology*, 166:310–326, 2018.

[253] Aleksandra Swiderska and Dennis Küster. Robots as malevolent moral agents: Harmful behavior results in dehumanization, not anthropomorphism. *Cognitive science*, 44(7):e12872, 2020.

[254] Katelyn Swift-Spong, Cheng K. Fred Wen, Donna Spruijt-Metz, and Maja J Matarić. Comparing backstories of a socially assistive robot exercise buddy for adolescent youth. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, page 1013–1018. IEEE Press, 2016.

[255] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. Clear evaluation of acoustic event detection and classification systems. In *Proceedings of the 1st International Evaluation harry zhang on Classification of Events, Activities and Relationships*, CLEAR'06, page 311–322, Berlin, Heidelberg, 2006. Springer-Verlag.

[256] Andrea L. Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6):716–737, 2008.

[257] Andrea Lockerd Thomaz, Cynthia Breazeal, et al. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, pages 1000–1005. Boston, MA, 2006.

[258] Sydney Thompson, Austin Narcomey, Alexander Lew, and Marynel Vázquez. Shutter: A low-cost and flexible social robot platform for in-the-wild deployments. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 94–96, New York, NY, USA, 2024. Association for Computing Machinery.

[259] Leimin Tian and Sharon Oviatt. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–32, 2021.

[260] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pages 3–12, 2020.

[261] Michael Tomasello. *Becoming human: A theory of ontogeny*. Belknap Press, 2019.

[262] Georgette L Troseth and Judy S DeLoache. The medium can obscure the message: Young children's understanding of video. *Child development*, 69(4):950–965, 1998.

[263] George Tzanetakis and Perry R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10:293–302, 2002.

[264] United States Bureau of Labor Statistics. American time use survey summary. `https://www.bls.gov/news.release/pdf/atus.pdf`, 2019.

[265] Anastasios Vafeiadis, Konstantinos Votis, Dimitrios Giakoumis, Dimitrios Tzovaras, Liming Chen, and Raouf Hamzaoui. Audio content analysis for unobtrusive event detection in smart homes. *Engineering Applications of Artificial Intelligence*, 89:103226, 2020.

[266] Amrisha Vaish, Manuela Missana, and Michael Tomasello. Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology*, 29(1):124–130, 2011.

[267] Diede P.M. Van der Hoorn, Anouk Neerincx, and Maartje M.A. de Graaf. "i think you are doing a bad job!": The effect of blame attribution by a robot in human-robot collaboration. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, page 140–148, New York, NY, USA, 2021. Association for Computing Machinery.

[268] Sanne Van Waveren, Elizabeth J Carter, and Iolanda Leite. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM international conference on intelligent virtual agents*, pages 151–158, 2019.

[269] John Voiklis, Boyoung Kim, Corey Cusimano, and Bertram F. Malle. Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 775–780, 2016.

[270] Tobias Kopp, Marco Baumgartner, and Steffen Kinkel. "it's not paul, it's a robot": The impact of linguistic framing and the evolution of trust and distrust in a collaborative robot during a human-robot interaction. *International Journal of Human-Computer Studies*, 178:103095, 2023.

[271] Lennart Wachowiak, Andrew Fenn, Haris Kamran, Andrew Coles, Oya Celiktutan, and Gerard Canal. When do people want an explanation from a robot? In *2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 752–761, 2024.

[272] Owen Waddington, Keith Jensen, and Bahar Köymen. Boundaries of apologies: Children avoid transgressors who give the same apology for a repeat offence. *Cognitive Development*, 64:101264, October 2022.

[273] Fuxing Wang, Yu Tong, and Judith Danovitch. Who do i believe? children's epistemic trust in internet, teacher, and peer informants. *Cognitive Development*, 50:248–260, 2019.

[274] Shuangge Wang, Anjiabei Wang, Sofiya Goncharova, Brian Scassellati, and Tesca Fitzgerald. Effects of robot competency and motion legibility on human correction feedback. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 789–799, 2025.

[275] Xijing Wang and Eva G Krumhuber. Mind perception of robots varies with their economic versus social function. *Frontiers in psychology*, 9:1230, 2018.

[276] Adrian F Ward, Andrew S Olsen, and Daniel M Wegner. The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, 24(8):1437–1445, 2013.

[277] Auriel Washburn, Akanimoh Adeleye, Thomas An, and Laurel D. Riek. Robot errors in proximate hri: How functionality framing affects perceived reliability and trust. *J. Hum.-Robot Interact.*, 9(3), May 2020.

[278] A. Watson. Genre breakdown of the top 250 tv programs in the united states in 2017. `https://www.statista.com/statistics/201565/most-popular-genres-in-us-primetime-tv/`, 2019.

[279] Kara Weisman, Carol S Dweck, and Ellen M Markman. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 114(43):11374–11379, 2017.

[280] Kara Weisman, Cristine H Legare, Rachel E Smith, Vivian A Dzokoto, Felicity Aulino, Emily Ng, John C Dulin, Nicole Ross-Zehnder, Joshua D Brahinsky, and Tanya Marie Luhrmann. Similarities and differences in concepts of mental life among adults and children in five cultures. *Nature Human Behaviour*, 5(10):1358–1368, 2021.

[281] David Westerman, Aaron C Cross, and Peter G Lindmark. I believe in a thing called bot: Perceptions of the humanness of "chatbots". *Communication Studies*, 70(3):295–312, 2019.

[282] Alicja Wróbel, Karolina Źróbek, Marie-Monique Schaper, Paulina Zguda, and Bipin Indurkhya. Age-appropriate robot design: in-the-wild child-robot interaction studies of perseverance styles and robot's unexpected behavior. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1451–1458. IEEE, 2023.

[283] Jie Xie and Mingying Zhu. Investigation of acoustic and visual features for acoustic scene classification. *Expert Systems with Applications*, 126:20–29, 2019.

[284] Jin Xu and Ayanna Howard. Evaluating the impact of emotional apology on human-robot trust. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, pages 1655–1661. IEEE, 2022.

[285] Kai Chi Yam, Yochanan E Bigman, Pok Man Tang, Remus Ilies, David De Cremer, Harold Soh, and Kurt Gray. Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*, 106(10):1557, 2021.

[286] Hongyan Yang, Hong Xu, Yan Zhang, Yan Liang, and Ting Lyu. Exploring the effect of humor in robot failure. *Annals of Tourism Research*, 95:103425, 2022.

[287] Shannon Yasuda, Devon Doheny, Nicole Salomons, Sarah Strohkorb Sebo, and Brian Scassellati. Perceived agency of a social norm violating robot. In *Proceedings of the annual meeting of the Cognitive Science Society*, 2020.

[288] Hang Yu, Reuben M. Aronson, Katherine H. Allen, and Elaine Schaertl Short. From "thumbs up" to "10 out of 10": Reconsidering scalar feedback in interactive reinforcement learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4121–4128, 2023.

[289] Ceciley Zhang and Sun Kyong Lee. "but you shouldn't blame me": A cross-national comparison of the effects of performance failures and trust repairs in human–robot interactions. *ACM Transactions on Human-Robot Interaction*, 15(1):1–28, 2025.

[290] Harry Zhang. The optimality of naive bayes. volume 2, 2004.

[291] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, dec 1997.