Perceived Morality of Robot and Human Transgressors Varies By Perceived Ability to Feel

Nicholas C. Georgiou*
Yale University
New Haven, CT, USA
nicholas.georgiou@yale.edu

Teresa Flanagan*
University of Chicago
Chicago, IL, USA
tflanagan2@uchicago.edu

Brian Scassellati
Yale University
New Haven, CT, USA
brian.scassellati@yale.edu

Tamar Kushnir
Duke University
Durham, NC, USA
tamar.kushnir@duke.edu

***Abstract*—Mistakes, failures, and transgressions committed by a robot are inevitable as robots become more involved in our society. When a wrong behavior occurs, it is important to understand what factors might affect how the robot is perceived by people. In this paper, we investigated how the type of transgressor (*human* or *robot*) and type of backstory depicting the transgressor's mental capabilities (*default*, *physio-emotional*, *socio-emotional*, or *cognitive*) shaped participants' perceptions of the transgressor's morality. We performed an online, between-subjects study in which 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 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.**

***Index Terms*—robot ethics; backstories; morality; human-robot interaction.**

I. 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 and 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 [1]–[6]. 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 [7], [8]. 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.

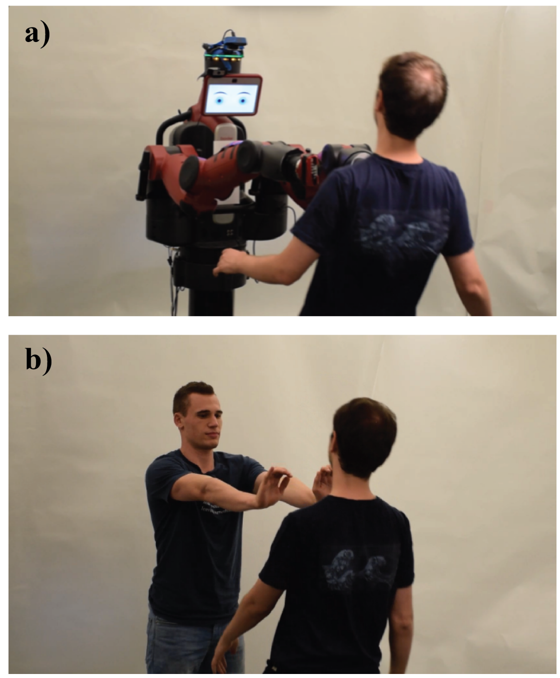

Fig. 1. The between-subjects study was centered on a video depicting a physical transgression (i.e., pushing down a human) committed by either a a) robot or b) human. Along with the human or robot condition, there were also conditions based on what mental capability backstory was highlighted about the transgressor (default vs. socio-emotional vs physio-emotional vs. cognitive) when they were first introduced to the participant.

In this paper, 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, a snippet of which can be seen in Fig. 1. 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.

* Authors contributed equally to this work.

II. 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 [7]–[17]. Other components involved with morality include situational awareness (e.g., moral or emotional knowledge), intentionality, desire, and free will [10], [15], [18]. 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 [19]. Weisman et al. propose that our mental perceptions of others include three dimensions [20]: 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 [21]–[23]. This leads to questions regarding under what circumstances we perceive and attribute morality to robots [24]–[27].

People do consider robots as moral agents in certain scenarios [11], [28], [29], although robots are not necessarily expected to treat moral decision-making in the same ways as humans [30]. People's perceptions of robot morality can be influenced by a variety of factors, including the robot's perceived autonomy [31], transparency [32], appearance [33]–[35], and affective capability [12], [36], [37].

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 [38], without a main focus on the perceived morality 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) [18].

Lastly, prior work has highlighted the power of framing and backstories in different contexts in HRI [39]–[48]. 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?*

III. 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.

A. 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 [49]. 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.

B. Materials

The study was conducted on the online survey platform, Qualtrics [50]. 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.

C. Conditions

Participants were randomly assigned to one of 16 conditions in a 2 (**Transgressor**: Human, Robot) $\times$ 4 (**Story Info**: Default, Physio-Emotional, Socio-Emotional, Cognitive) $\times$ 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 [51]. 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 [20]. A detailed description of each story that was provided for the four conditions can be found in Table I.

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.

D. Procedure

1) Introduction Phase: Depending on the Transgressor and Story Info condition, the participant was told a story, which can be viewed in Table I.
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[1] 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 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 E, Moral Judgments Questionnaire.

A Mental Perception Survey (see Section E) 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 introduced in the Introduction Phase, while the other half were shown the survey after the Moral Judgments Phase.

[1]Videos can be found in the Supplementary Materials.

TABLE I
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.

E. 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 [7]–[13], [16]–[19], [23], [31], [32], [52], [53]. 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 think 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, fear, tiredness), 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 [20]. The full list of questions, which can be found in the Supplemental Materials, 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 that were asked about the transgressor's social-emotions were directly related to the socio-emotional story, and the questions that were asked about the transgressor's cognitive abilities were directly related to the cognitive story. These questions served both as a manipulation check for each story condition, and as indicators of how a story in one (or none) of these areas can influence how the agent is perceived in other areas – although this aspect of the experiment is not discussed in detail in this paper.

IV. 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[2]. For each of our Moral Judgments 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.

[2]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.

Figures 2 and 3 show the means or 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.

A. 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.

1) 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: $M = 2.26$, $SD = 0.83$).

2) 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 ($M = 1.66$, $SD = 1.15$).

B. Moral Judgments Questionnaire

1) ***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 2B), and story, $F(3, 712) = 2.60$, $p = .051$ (see Figure 2C).

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 2A).

*2) **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 2B), and story, $F(3, 712) = 3.93$, $p = .0085$ (see Figure 2C).

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 2A). 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 2A).

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, $ps < .0007$.

*3) **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 2B). 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 2C). 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 2A).

*4) **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 2B), and story, $F(3, 712) = 10.78$, $p < .0001$ (see Figure 2C).

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 2A).

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 2A).

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, $ps < .0001$.

*5) **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 2B), and story, $F(3, 712) = 19.16$, $p < .0001$ (see Figure 2C).

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 2A).

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 2A).

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-

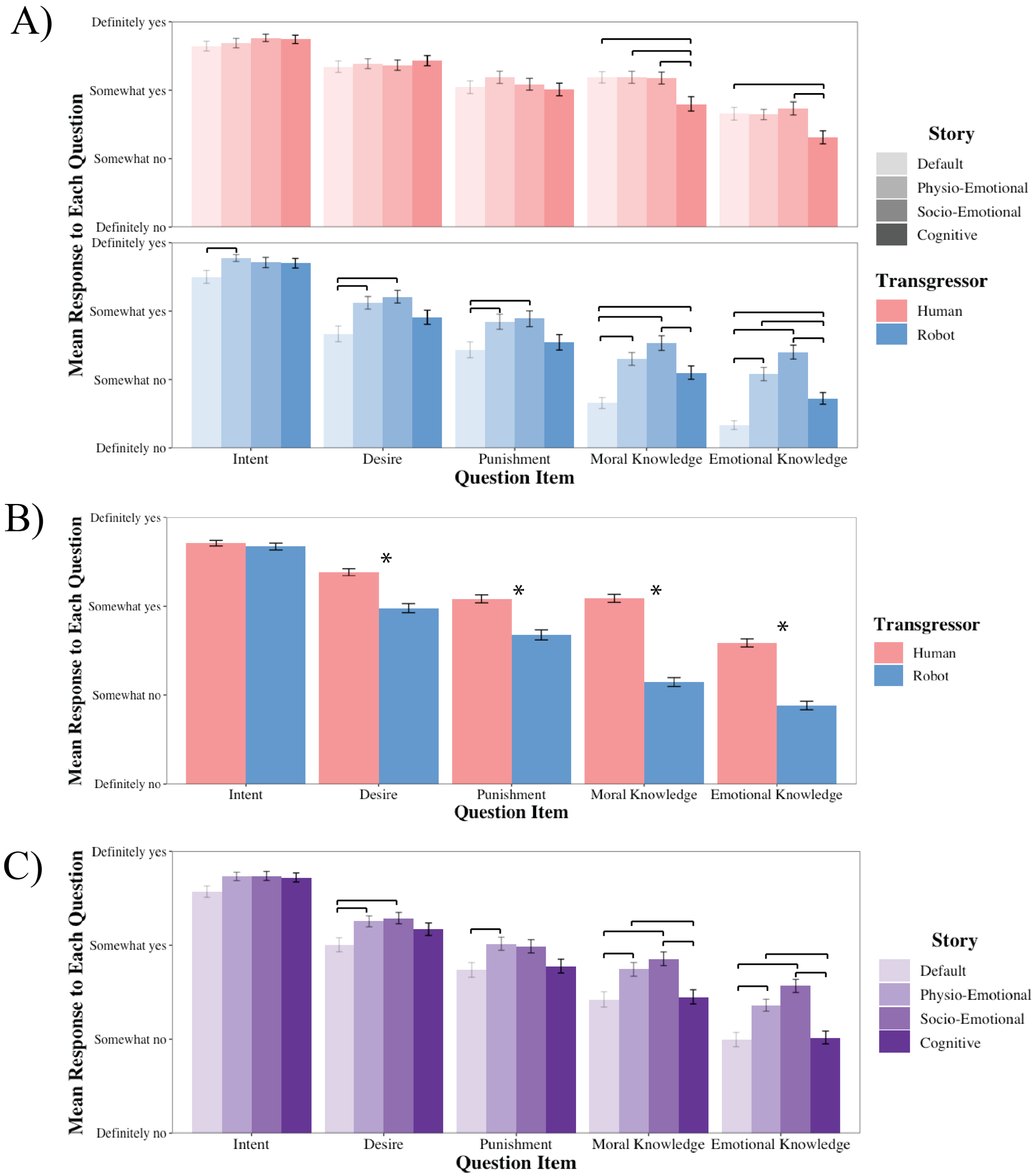

Fig. 2. **Participant Responses to the Moral Judgments Questionnaire.** A) The graphs display the mean responses, with standard error, for each measure, split by transgressor type and story type. B) The graph is collapsed across story types and displays the mean response, with standard error, by transgressor 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.

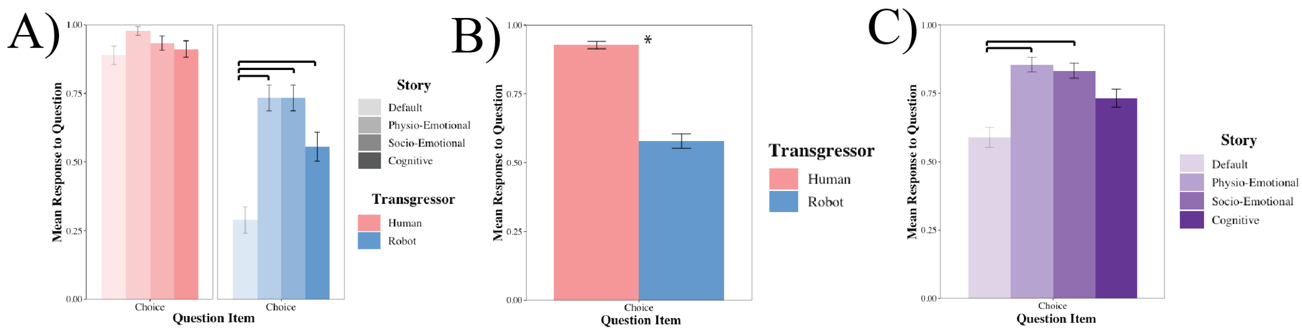

Fig. 3. **Participant Responses to the Choice Measure.** A) The graphs show the mean responses with standard error, split by transgressor type and story type. B) The graph is collapsed across story types and displays the mean response, with standard error, by transgressor type. C) The graph is collapsed 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.

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, $ps < .0001$.

*6) **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 3B). 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 3C). 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 3A). 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 $p < .0001$.

V. 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 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 anthropomorphizing robots can lead to stronger connections and bonds with them [54]–[56], which can be particularly important when they are used for emotional support. However, anthropomorphizing robots becomes a more complicated issue when a feeling robot 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 [10]. 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.

VI. 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 [57], [58], 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 [13]. 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, works have 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 [59]–[62]. 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.

VII. Conclusion

To investigate how different perceptions of mind influence a transgressor's perceived morality, 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 find 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.

VIII. Acknowledgments

The authors thank undergraduate research assistants, Lila Selin and Collin Pitts, for their work on this project. We also thank Morgan Vanderwall and Kate Candon for their help in improving the manuscript. This work was supported by the National Science Foundation (NSF) awards SL-1955280, IIS-2106690, and IIS-1955653, and Office of Naval Research (ONR) award N00014-24-1-2124. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or ONR.

REFERENCES

[1] S. Honig and T. Oron-Gilad, “Understanding and resolving failures in human-robot interaction: Literature review and model development,” *Frontiers in psychology*, vol. 9, p. 861, 2018.

[2] H. Yang, H. Xu, Y. Zhang, Y. Liang, and T. Lyu, “Exploring the effect of humor in robot failure,” *Annals of Tourism Research*, vol. 95, p. 103425, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160738322000767

[3] T. Flanagan, N. C. Georgiou, B. Scassellati, and T. Kushnir, “School-age children are more skeptical of inaccurate robots than adults,” *Cognition*, vol. 249, p. 105814, 2024.

[4] J. B. Lyons, I. aldin Hamdan, and T. Q. Vo, “Explanations and trust: What happens to trust when a robot partner does something unexpected?” *Computers in Human Behavior*, vol. 138, p. 107473, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S074756322200293X

[5] S. Choi, A. S. Mattila, and L. E. Bolton, “To err is human (-oid): how do consumers react to robot service failure and recovery?” *Journal of Service Research*, vol. 24, no. 3, pp. 354–371, 2021.

[6] K. C. Yam, Y. E. Bigman, P. M. Tang, R. Ilies, D. De Cremer, H. Soh, and K. Gray, “Robots at work: People prefer—and forgive—service robots with perceived feelings.” *Journal of Applied Psychology*, vol. 106, no. 10, p. 1557, 2021.

[7] G. Lima, M. Cha, C. Jeon, and K. S. Park, “The conflict between people’s urge to punish ai and legal systems,” *Frontiers in Robotics and AI*, vol. 8, p. 756242, 2021.

[8] S. Gless, E. Silverman, and T. Weigend, “If robots cause harm, who is to blame? self-driving cars and criminal liability,” *New Criminal Law Review*, vol. 19, no. 3, pp. 412–436, 2016.

[9] T. 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)*. IEEE, 2016, pp. 1168–1171.

[10] Y. E. Bigman, A. Waytz, R. Alterovitz, and K. Gray, “Holding robots responsible: The elements of machine morality,” *Trends in cognitive sciences*, vol. 23, no. 5, pp. 365–368, 2019.

[11] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier, and R. 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*, 2012, pp. 33–40.

[12] M. Lee, P. Ruijten, L. Frank, Y. de Kort, and W. IJsselsteijn, “People may punish, but not blame robots,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445284

[13] X. Leo and Y. E. Huh, “Who gets the blame for service failures? attribution of responsibility toward robot versus human service providers and service firms,” *Computers in Human Behavior*, vol. 113, p. 106520, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563220302727

[14] M. Keijsers, H. Kazmi, F. Eyssel, and C. Bartneck, “Teaching robots a lesson: determinants of robot punishment,” *International Journal of Social Robotics*, vol. 13, pp. 41–54, 2021.

[15] M. Franklin, H. Ashton, E. Awad, and D. Lagnado, “Causal framework of artificial autonomous agent responsibility,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 276–284. [Online]. Available: https://doi.org/10.1145/3514094.3534140

[16] E. Nahmias, S. Morris, T. Nadelhoffer, and J. Turner, “Surveying freedom: Folk intuitions about free will and moral responsibility,” *Philosophical Psychology*, vol. 18, no. 5, pp. 561–584, 2005.

[17] R. F. Baumeister, “Free will, consciousness, and cultural animals,” *Are we free*, pp. 65–85, 2008.

[18] E. Nahmias, C. H. Allen, and B. Loveall, “When do robots have free will? exploring the relationships between (attributions of) consciousness and free will,” *Free will, causality, and neuroscience*, vol. 338, pp. 57–80, 2020.

[19] K. Gray and D. M. Wegner, “Moral typecasting: divergent perceptions of moral agents and moral patients.” *Journal of personality and social psychology*, vol. 96, no. 3, p. 505, 2009.

[20] K. Weisman, C. S. Dweck, and E. M. Markman, “Rethinking people’s conceptions of mental life,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 43, pp. 11 374–11 379, 2017.

[21] K. Gray and D. M. Wegner, “Feeling robots and human zombies: Mind perception and the uncanny valley,” *Cognition*, vol. 125, no. 1, pp. 125–130, 2012.

[22] X. Wang and E. G. Krumhuber, “Mind perception of robots varies with their economic versus social function,” *Frontiers in psychology*, vol. 9, p. 1230, 2018.

[23] M. G. Reinecke, M. Wilks, and P. Bloom, “Developmental changes in perceived moral standing of robots,” in *Proceedings of the annual meeting of the Cognitive Science Society*, vol. 43, no. 43, 2021.

[24] J.-F. Bonnefon, I. Rahwan, and A. Shariff, “The moral psychology of artificial intelligence,” *Annual review of psychology*, vol. 75, no. 1, pp. 653–675, 2024.

[25] A. Ladak, S. Loughnan, and M. Wilks, “The moral psychology of artificial intelligence,” *Current Directions in Psychological Science*, vol. 33, no. 1, pp. 27–34, 2024.

[26] L. Floridi and J. W. Sanders, “On the morality of artificial agents,” *Minds and machines*, vol. 14, pp. 349–379, 2004.

[27] A. F. Ward, A. S. Olsen, and D. M. Wegner, “The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead,” *Psychological Science*, vol. 24, no. 8, pp. 1437–1445, 2013.

[28] J. Voiklis, B. Kim, C. Cusimano, and B. F. Malle, “Moral judgments of human vs. robot agents,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 775–780.

[29] J. Banks, “Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust,” *International Journal of Social Robotics*, vol. 13, no. 8, pp. 2021–2038, 2021.

[30] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. 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*, ser. HRI ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 117–124. [Online]. Available: https://doi.org/10.1145/2696454.2696458

[31] C. Furlough, T. Stokes, and D. J. Gillan, “Attributing blame to robots: I. the influence of robot autonomy,” *Human factors*, vol. 63, no. 4, pp. 592–602, 2021.

[32] T. Kim and P. 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*, 2006, pp. 80–85.

[33] M. 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*, vol. 14, 2023.

[34] B. F. Malle, M. Scheutz, J. Forlizzi, and J. 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*, ser. HRI ’16. IEEE Press, 2016, p. 125–132.

[35] M. Laakasuo, J. Palomäki, and N. Köbis, “Moral uncanny valley: A robot’s appearance moderates how its decisions are judged,” *International Journal of Social Robotics*, vol. 13, no. 7, pp. 1679–1688, 2021.

[36] Y. E. Bigman and K. Gray, “People are averse to machines making moral decisions,” *Cognition*, vol. 181, pp. 21–34, 2018.

[37] S. R. Nijssen, B. C. Müller, T. Bosse, and M. Paulus, “Can you count on a calculator? the role of agency and affect in judgments of robots as moral agents,” *Human–Computer Interaction*, vol. 38, no. 5-6, pp. 400–416, 2023.

[38] A. Swiderska and D. Küster, “Robots as malevolent moral agents: Harmful behavior results in dehumanization, not anthropomorphism,” *Cognitive science*, vol. 44, no. 7, p. e12872, 2020.

[39] E. Roesler, “Anthropomorphic framing and failure comprehensibility influence different facets of trust towards industrial robots,” *Frontiers in Robotics and AI*, vol. 10, p. 1235017, 2023.

[40] E. Roesler, J. Pickl, and F. W. Siebert, “Investigating the impact of anthropomorphic framing and product value on user acceptance of delivery robots,” in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 347–357.

[41] T. Kopp, M. Baumgartner, and S. 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*, vol. 158, p. 102730, 2022.

[42] T. Kopp, M. Baumgartner, and S. 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*, vol. 178, p. 103095, 2023.

[43] L. Onnasch and C. L. Hildebrandt, "Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 1, pp. 1–24, 2021.

[44] K. Swift-Spong, C. K. F. Wen, D. Spruijt-Metz, and M. 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)*. IEEE Press, 2016, p. 1013–1018. [Online]. Available: https://doi.org/10.1109/ROMAN.2016.7745233

[45] K. Darling, P. Nandy, and C. 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)*. IEEE, 2015, pp. 770–775.

[46] L. Onnasch and E. Roesler, "Anthropomorphizing robots: The effect of framing in human-robot collaboration," in *Proceedings of the human factors and ergonomics Society annual meeting*, vol. 63, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 1311–1315.

[47] N. Anzabi, A. Etemad, and H. 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*, 2023, pp. 431–435.

[48] J. M. Kory-Westlund and C. 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*, vol. 6, p. 54, 2019.

[49] S. Palan and C. Schitter, "Prolific. ac—a subject pool for online experiments," *Journal of behavioral and experimental finance*, vol. 17, pp. 22–27, 2018.

[50] "Qualtrics," https://www.qualtrics.com, 2024, location: Provo, Utah, USA.

[51] C. Fitzgerald, "Developing baxter," in *2013 IEEE conference on technologies for practical robot applications (TePRA)*. IEEE, 2013, pp. 1–6.

[52] F. J. Ayala, "The difference of being human: Morality," *Proceedings of the National Academy of Sciences*, vol. 107, no. supplement_2, pp. 9015–9022, 2010.

[53] T. Flanagan, G. Wong, and T. Kushnir, "The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies." *Developmental psychology*, vol. 59, no. 6, p. 1017, 2023.

[54] I. Leite, A. Pereira, C. Martinho, and A. Paiva, "Are emotional robots more fun to play with?" in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 77–82.

[55] 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)*, vol. 2, 1999, pp. 858–863 vol.2.

[56] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children," *International Journal of Social Robotics*, vol. 6, pp. 329–341, 2014.

[57] D. L. Medin, S. G. García *et al.*, "Conceptualizing agency: Folkpsychological and folkcommunicative perspectives on plants," *Cognition*, vol. 162, pp. 103–123, 2017.

[58] K. Weisman, C. H. Legare, R. E. Smith, V. A. Dzokoto, F. Aulino, E. Ng, J. C. Dulin, N. Ross-Zehnder, J. D. Brahinsky, and T. M. Luhrmann, "Similarities and differences in concepts of mental life among adults and children in five cultures," *Nature Human Behaviour*, vol. 5, no. 10, pp. 1358–1368, 2021.

[59] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! an interaction with a cheating robot," in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '10. IEEE Press, 2010, p. 219–226.

[60] A. Rosero, E. Dula, H. Kelly, B. F. Malle, and E. K. Phillips, "Human perceptions of social robot deception behaviors: an exploratory analysis," *Frontiers in Robotics and AI*, vol. 11, 2024. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2024.1409712

[61] S. Yasuda, D. Doheny, N. Salomons, S. S. Sebo, and B. Scassellati, "Perceived agency of a social norm violating robot," in *Proceedings of the annual meeting of the Cognitive Science Society*, 2020.

[62] J. F. Limprayoon, N. C. Georgiou, N. P. Ua-Arak, and B. Scassellati, "The effects of a gossiping robot on team cohesion," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 2066–2071.