Did the Robot Really Intend to Harm Me? The Effect of Perceived Agency and Intention on Fairness Judgments

Houston Claure*
Yale University
New Haven, Connecticut
houston.claure@yale.edu

Inyoung Shin*
Yale University
New Haven, Connecticut
inyoung.shin@yale.edu

J. Gregory Trafton
U.S. Naval Research Laboratory
Washington, D.C.
greg.trafton@nrl.navy.mil

Marynel Vázquez
Yale University
New Haven, Connecticut
marynel.vazquez@yale.edu

***Abstract*—Determining whether a robot's actions will be perceived as fair or unfair is complicated in Human-Robot Interaction (HRI), where factors like the robot's perceived agency and intent may influence these judgments. We report findings from two experiments that examine how people evaluate fairness after reviewing a scenario where a robot harms a human. In these experiments, we manipulate different aspects of the context: the fairness of a situation (Fair vs. Unfair); the perceived agency of a robot that commits the harm (High Agency vs. Low Agency); and the perceived intention behind the harmful action (Intentional vs. Unintentional). We examine fairness as a multifaceted construct, using Fairness Theory to capture three key components: reduced welfare, conduct, and moral transgression. We find that this multifaceted perspective can capture nuances in fairness judgments. When robots are perceived to have greater decision-making autonomy, humans tend to assign higher moral responsibility, especially when harmful actions appear intentional. Conversely, when robots are seen as merely following predetermined programming, people focus more on the possibility that the programming could have been designed differently. These findings highlight how agency and intention need to be considered when investigating fairness in HRI.**

***Index Terms*—Fairness; Agency; Human-Robot Interaction; Fairness Theory**

I. Introduction

Fairness is essential for building trust and sustaining collaboration [1]–[3]. When people believe they are treated unfairly by others—whether individuals, groups, organizations, or society—they tend to assign blame, seek punishment, and attempt to correct the situation [4]–[6]. With advancements in artificial intelligence (AI) and robotics, the possibility of robots acting as social collaborators or even decision-makers has increased. Accordingly, investigating human responses to a robot's unfair behavior has gained momentum within the field of Human-Robot Interaction (HRI) [7]–[14]. However, the lack of an united theoretical framework for fairness in HRI has led to a variety of interpretations and methods for identifying and measuring fairness. The most common approach investigates human judgments of fairness through a distributive lens, where fairness judgments are thought to arise from comparisons between the outcomes or treatment a robot provides to one individual versus others [7]–[10], [15]. There are limitations to this approach as robots have been shown to trigger unfair judgments from humans through a variety of different actions that do not revolve around the distribution of resources (e.g. a robot cheating during a rock, paper, scissors game [11]). Moreover, fairness is a multifaceted concept [6]. People have shown to consider both the outcomes they receive and whether the procedures leading to those outcomes were appropriate [16], [17]. Additionally, they evaluate whether the robot acted in line with social norms [11] and whether it had justifiable reasons for its behavior [6].

A more recent approach applies Fairness Theory [4] from organizational psychology and aims to establish a framework for studying fairness in HRI [12]. Fairness Theory takes a multifaceted approach, incorporating three key components: whether the entity's action has reduced the individual's well-being (reduced welfare), whether a different action could have been taken (conduct), and whether the action violated social norms (moral transgression). However, humans may not always perceive a robot to be in control over its actions and therefore may interpret the robot's intentions differently from humans [18], [19]. These unique dynamics in HRI lead us to examine how the perceived agency of robots and the perceived intentions behind their actions influence how people assess fairness.

In two online studies, we demonstrate how the perceived agency of a robot, the perceived intention of a robot's actions, and their interaction influence fairness perceptions. In this paper, we operationalize fairness using different theories and frameworks. The first study examines how the perceived agency level of a robot influences fairness perceptions drawing on different frameworks including single-item fairness, distributive justice, procedural justice, and Fairness Theory. Our second experiment extends the first study by focusing on agency and intentions in the context of unfairness, examining how fairness perceptions vary based on the levels of a robot's perceived agency and the intentions behind harmful actions.

Our paper makes three key contributions to the literature on fairness in HRI. First, we examine the effects of perceived agency and intention on a robot's actions based on various

*Both authors contributed equally

979-8-3503-7893-1/25/$31.00 ©2025 IEEE

fairness judgments, demonstrating that perceptions of fairness related to accountability are specifically influenced by these factors. Second, we provide evidence that fairness toward robots is multifaceted, incorporating outcome-based views, counterfactual thinking, and morality. Finally, we offer guidelines for future research, recommending appropriate frameworks and measures based on research goals and contexts, with particular attention to varying degrees of perceived agency and intention of robots.

II. Related Works

A. Fairness in HRI

Researchers in HRI often use the concept of fairness as an indicator of how people judge robots and their actions. As the concept of fairness is multifaceted, its applications are diverse and difficult to define within a single dimension [4], [6]. The most common approach to applying fairness in HRI is to examine it in terms of human reactions to unequal resource distribution by robots embedded in human groups [7]–[12], [15]. In these settings, robots allocate resources ranging from tangible items such as tools [20] to more communicative resources like gaze [21], [22]. Humans are particularly sensitive to how robots distribute these resources. Receiving fewer resources than others can lead to negative impacts on cooperative behaviors among group members, such as trust in the robot [7], group cohesion [12], and performance on team tasks [23]. Another application of fairness revolves around social norms and ethical standards. Here, fairness judgments gauge whether a robot's actions adhere to or violate expected rules and norms within a given context [11], [24]–[27]. For instance, in a rock-paper-scissors game, a robot that changed its action to win was perceived as behaving unfairly [11]. Lastly, fairness has been examined in the context of robots perpetuating or challenging socially structured prejudices and biases [28], [29], particularly in how robots exhibit bias or impartiality during interactions based on a person's attributes, such as skin color [30] or gender [28].

Given the different conceptualizations and applications of fairness in HRI, developing a unified measure of fairness is challenging. Various methods to assess fairness toward robots have been proposed, but each comes with its limitations. The most common approach relies on heuristics, often asking a single question about whether people believe a robot's action is fair or unfair [12], [29]. While this can apply to a range of fairness judgments, a single question does not clarify which specific dimension of fairness is being addressed [31]. Other researchers propose adapting the organizational justice framework for HRI [32], [33], which was originally developed to explain how workers react to unfair treatment within organizations [31]. However, the forms of justice identified in this framework primarily focus on how resources are distributed, such as the fairness of allocation (distributive justice) [34] or the procedures used to determine allocations (procedural justice) [35]. These justice measures also tend to emphasize formal rules and ethical standards, whereas fairness in HRI often involves subjective interpretations influenced by morality, self-interest, and other motivations [4]. This makes it challenging to apply the organizational justice framework to the context where individuals evaluate the fairness of a single entity like a robot.

This study adopts a more recent approach to measuring fairness in HRI, Fairness Theory [4], which encompasses various fairness motives and identifies the conditions under which individuals hold entities accountable for their actions

B. Fairness Theory

At its core, Folger and Cropanzano's Fairness Theory [4] explains how an entity (i.e., an individual, a group, an organization) is judged as accountable for its actions that cause negative outcomes. It posits that accountability for fairness judgments has three interrelated components: reduced welfare (i.e., the extent to which an entity's action diminishes well-being), conduct (i.e., the extent to which an entity could have acted differently to produce better outcomes), and moral transgression (i.e., the extent to which the action violates moral norms). According to Fairness Theory, all three components must be present for an entity to be held accountable for unfair actions. Specifically, reduced welfare provides a prerequisite condition under which people can blame an entity for negative outcomes. Conduct and moral transgression address whether the entity violates expectations by not acting differently to achieve better outcomes and not adhering to moral norms. For example, consider a pedestrian who was hit by a car. If the accident was due to external factors such as a car malfunction, the driver's conduct and moral transgression could be seen as moderate by the pedestrian. However, if the accident happens due to the driver's ignoring a red light, both conduct and moral transgression would likely be seen as severe, leading to greater accountability.

Fairness Theory provides a unified framework that can be used to integrate prior definitions of fairness in HRI. For example, Fairness Theory encompasses how people form fairness judgments based on the distribution of rewards or resources compared to others (distributive justice) [36], the process used to reach that conclusion (procedural justice) [37], whether the entity adheres to expected norms and rules [11], and broader issues of morality, including prejudice and bias [28]. However, nuances in how fairness toward a robot is judged complicate the application of Fairness Theory in HRI. Fairness Theory was developed to assess unfairness in human interactions, assuming that the entity being judged has some level of capability to act and form intentions independently. In particular, conduct and moral transgressions are based on violations of expectations, which can occur when humans believe the entity has the capability to act better or adhere to morality [4]. But robots are not always perceived as capable of human-like behaviors. People often interpret the intentions behind their actions differently from human intentions. To hold a robot accountable for its unfair actions according to Fairness theory, some assurance of its agency and intentions may be required [12].

C. Robots as Targets of Blame

Prior HRI research on moral blame suggests that humans attribute different mental models to robots when attributing blame to robots compared to humans [38]–[40]. To blame a robot, people first consider whether it has the ability to cause negative outcomes [40], [41]. This is closely related to its perceived agency, or the robot's perceived cognitive ability to make independent decisions [42]. People are more likely to blame robots with high perceived agency than those seen as just following pre-programmed, simple instructions [24], [38], [43], [44]. Although fairness judgments differ from moral blame, both involve holding entities accountable. We propose that perceived agency influences fairness judgments in a similar way. The relationship between perceived agency and fairness, however, has yet to be fully explored. We thus ask the question:

RQ1: How does the perceived agency of a robot influence fairness judgments towards the robot?

Another criterion used in the blaming process for robots is the perceived intention behind actions, or whether harmful or immoral actions are perceived to be caused deliberately [40], [45]. If wrong actions are perceived as accidental or unintentional, less moral accountability is attributed because no malicious intent is involved [40]. The perceived intention of harmful actions is generally examined when an entity is perceived to have agency [41], [46]. However, even when a robot is perceived to lack the ability to make its own independent decisions, its actions can still be perceived as intentional. In such cases, the perceived intentions reflect those of the users, programmers, or institutions behind the robot [38], [45], [47], [48]. Reflecting the similarity between moral judgment and fairness judgment, we propose that fairness attribution becomes stronger when a robot's harmful actions are intentionally committed. However, the ways in which the perceived intentions affect fairness judgments across levels of perceived agency remain underexplored. We thus pose two research questions regarding the perceived intentions of a robot's actions:

RQ2: How does the perceived intention of a robot's action influence fairness judgments towards the robot?

RQ3: How do the perceived intentions behind a robot's actions, in combination with its perceived agency, influence fairness judgments toward the robot?

D. Hypotheses

There are three main frameworks used in HRI to measure fairness perceptions of robots. Rather than expecting all types of fairness perceptions to be influenced by perceived agency and intentions, we anticipate that those measured through accountability-based frameworks are more likely to be impacted by these factors.

Our first research question focuses on the effects of perceived agency on fairness perceptions. When humans look to attribute blame to a robot for an unfair action, they first assess whether the robot had the capability to cause the negative event [40]. Prior research suggests that robots perceived as having higher capabilities and agency are assigned more blame [49] and moral responsibility compared to robots that seem to be following simple programmed instructions [38], [45], [50]. People may believe that robots with higher agency could or should have acted differently, yet chose not to do so through their own decision-making. Given that the perceived agency of robots is tied to accountability judgments, we anticipate that only fairness perceptions related to accountability will be affected by the robot's perceived agency.

Measuring the fairness of a robot's action using a single-item question has been widely used in various contexts related to equal distributions of tangible resources [51], following the normative rules and procedures [11], moral responsibilities [52], and social bias (e.g., gender and race) [28]. Given this global application, we anticipate that fairness judgments measured through the single item capture some aspects of accountability. Broadly speaking, the single-item question asks how fair people think a robot is [11], [13]. Studies that define fairness as the balance between self-interest and concern for others specify questions around how fair the robot or its actions are toward the individual [12]. We do not suspect that such subtle differences make a significant variation in fairness perceptions. We thus focus on the single item that captures the robot's actions toward the individual and hypothesize:

H1. Fairness and Agency on Unfairness Perceptions: In unfair contexts, single-item unfairness ratings will be significantly higher when the perceived agency of a robot is high compared to when it is low.

Fairness Theory is the framework built upon accountability judgments. Although it posits that all three components—reduced welfare, conduct, and moral transgressions—must be present for an entity to hold accountability, conduct, and moral transgression are more strongly related to the accountability process than reduced welfare [4]. Conduct is judged more harshly when a robot with high perceived agency fails to act for better outcomes. Moral transgressions are also seen as more severe when a human-like robot, which seems aware of moral norms, violates them [49]. Therefore, we hypothesize:

H2. Fairness and Agency on Fairness Theory: In unfair contexts, humans will show higher ratings of a) conduct and b) moral transgressions when a robot is seen to have high agency compared to when it has low agency.

By contrast, reduced welfare is linked to the experience of negative outcomes, which remains consistent regardless of how humans perceive a robot's agency. For example, the negativity of harm caused by either a robot perceived to have human-like capabilities or one following simple programming rules would be perceived similarly. Therefore, we do not propose a hypothesis regarding reduced welfare and perceived agency.

Work that applies distributive [7] and procedural justice [53] evaluates fairness based on outcomes a person receives or procedures involved in that decision [31]. These outcome-based measures may not vary based on the nature of the entity being judged. Therefore, we do not hypothesize that

Fig. 1. Visuals used for the high and low agency conditions. Scenarios 1 and 2 were used for both studies.

the perceived agency of a robot influences fairness perceptions measured by distributive and procedural justice.

Our second and third research questions focus on the perceived intentions behind a robot's harmful actions and its interaction with the perceived agency. These questions are built upon work that suggests that people judge intentions after assessing agency [40]. In fact, when a robot perceived to be capable of making independent decisions causes intentional harm, it prompts strong expectation violations and counterfactual thinking, leading to greater blame and emotional reactions [54], [55]. In contrast, accidental harm is viewed as uncontrollable, resulting in less blame on a robot. This assumption of harmful intent, however, typically assumes the actor has independent agency [40]. This may not always be the case with robots. Some people may view robots as being controlled by third parties, like developers or organizations, and attribute the intentions of the robot's actions to these entities [56]. This in turn can reduce blame on the robot itself [4]. We thus anticipate that fairness perceptions—measured through single items, conduct, and moral transgression—will be influenced by the perceived intentions behind harmful actions and interact with perceived agency.

H3. Intention and Agency on Unfairness Perceptions: a) single-item unfairness ratings will be significantly higher when the harmful action is perceived as intentional b) with this effect being stronger when the robot is perceived to have high agency compared to low agency.

H4. Intention and Agency on Conduct: a) conduct, one of the Fairness Theory components, will be higher when the harmful action is perceived is intentional b) with this effect being stronger when the robot is perceived to have high agency compared to low agency.

H5. Intention and Agency on Moral Transgression: a) moral transgression, one of Fairness Theory components, will be higher when the harmful action is perceived is intentional, b) with this effect being stronger when the robot is perceived to have high agency compared to low agency.

As discussed earlier, reduced welfare, distributive, and procedural justice are measures based on experiences of negative events or outcomes. Following the same reasoning we applied to perceived agency, fairness perceptions based on these measures are likely to remain consistent regardless of whether the robot caused harm intentionally or accidentally. Therefore, we do not propose any hypotheses regarding these three measures and the perceived intention as well.

III. Study 1

A. Methods

To address the first research question and test the corresponding hypotheses (H1 and H2), we designed an online vignette study that presented two possible social contexts in which a robot played a role that emotionally, economically, or physically influenced a person in a work dynamic. In one scenario, a robot was responsible for assigning tasks to employees in the workplace. In the other scenario, the robot managed production goals and workflow while monitoring employee conditions in a warehouse. Each participant observed only one scenario. In each case, participants were asked to imagine themselves as the person influenced by the robot. All scenarios were set "in the near future" to help participants envision the events as plausible and likely to occur (given the popularity and rise of AI systems).

We crafted a 2x2 experiment with conditions that varied the fairness of the robot's actions (fair vs. unfair) and its level of perceived agency (high and low). A high agency robot was based on Trafton et al. [42] and presented as being driven by its own internal thoughts and feelings and less by the specific environment. A low-agency robot strictly followed programmed instructions, with any deviations caused by system errors. In the **high-agency** conditions, the human-like robot either upheld fairness (**fair condition**) or engaged in favoritism and self-serving bias (**unfair condition**). In the **low-agency** conditions, the programmed robot either followed expected procedures (**fair condition**) or caused negative outcomes due to system errors (**unfair condition**). To assist in manipulating participants' perceptions, we included images used in Figure 1 and prompts from Table 2. These images were obtained from Adobe Stock. We also modified these images using Adobe Photoshop's Generative Fill.

B. Procedure

With IRB approval, we administrated an online survey featuring one scenario followed by corresponding questions. After reading a brief description of the task, participants were randomly assigned to one of 8 scenarios where they were asked to imagine themselves as the person in the situation. They then completed an attention check to ensure they had accurately understood key details from the scenario. Next, they rated various items related to Fairness Theory (i.e., Reduced welfare, Conduct, Moral transgressions), perceived agency, negative intention behind the actions, fairness, distributive and procedural justice. Throughout the survey, participants were instructed to respond to those questions as if they were the person in the scenario. At the end of the survey, participants completed demographic questions.

C. Measures

Perceived Agency: To capture participants' perceived agency of a robot, we used the scale developed by Trafton et al. [42]. Participants rated four items on a 5-point Likert scale, with 5 representing "strongly agree". For the analysis, we calculated the mean score of those five items (Cronbach's $\alpha = .846$).

Fairness Theory Components: To our knowledge, there are no established measures to assess the three components of Fairness Theory in human-to-human or HRI contexts. Therefore, we developed five items for each component (see Table 3). Participants responded to all items on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), and we calculated the mean score for each component for the analysis. The internal reliability of the five items for each factor, measured using Cronbach's alpha and omega, was satisfactory ($> .75$). An exploratory factor analysis conducted on the 15 measurement items further supported a general three-factor loading structure.

Unfairness Perception: To capture perceptions of unfairness, we asked participants to rate the extent to which they perceived the robot's actions toward the person in the scenario as fair. Participants responded using a 5-point scale, where 1 indicated "fair" and 5 indicated "unfair".

Distributive and Procedural Justice: We adapted organizational justice scales from Colquitt et al. [57] to fit the HRI context. For distributive and procedural justice, participants rated items on a 5-point scale ranging from 1 (to a very small extent) to 5 (to a very large extent). The distributive justice scale had four questions while the procedural justice scale had seven questions. For analysis, the mean score for each scale was calculated (Cronbach's $\alpha = .917$ for distributive justice; Cronbach's $\alpha = .872$ for procedural justice). See the Supplementary Material for the complete list of items for each survey.

Table 2. Study 1: left (<>) is high-agency, right is low-agency.

	Scenario 1: Work Schedule	Scenario 2: Warehouse Worker
Prompt	You work at a company with a robot that manages work schedules and assigns tasks to employees.	You suffer an injury while working under a managerial robot.
Fair Context	The robot assigns a manageable workload to each employee because it <**decides**/follows its programming> to account for individual differences among coworkers.	The robot <**decides**/follows its programming> to follow appropriate injury protocol instead of prioritizing the company's productivity. You learn that the immediate action from the robot prevented your injury from getting worse.
Unfair Context	The robot <**decides**/follows its programming> to assign more tasks to you than to your coworker because your coworker interacts with the robot more frequently. As a result, you end up working overtime and appearing unproductive to upper management.	The robot <**decides**/follows its programming> to prioritize the company's productivity. The robot insists that you return to work and later your injury gets worse.

Table 3. Fairness Theory Survey Items.

Fairness Theory	Items
Reduced Welfare $\alpha = .95$	The robot's behavior made you feel frustrated
	The robot's behavior made you feel upset
	The robot's behavior resulted in a loss of your resources (e.g., money, time, effort, etc.)
	The robot's behavior prevented you from gaining benefits you deserved
	The robot's behavior made you feel unsafe
Conduct $\alpha = .91$	The robot acted poorly despite having better options
	The robot ignored alternative actions it could have taken
	The robot had the power to act more for your benefit
	The robot had the capability to help you more
	You expected the robot to help you more
Moral Transgression $\alpha = .95$	The robot's behavior was morally wrong
	The robot's actions were unethical
	The robot's actions were corrupted
	The robot violated social norms through its actions
	The robot's behavior showed bias against you

D. Study Participants

We recruited 225 participants from the United States on Prolific but 26 of them failed to pass attention check. The final sample of 199 (77 male, 120 female, 1 non-binary, 1 other) had an average age of 35 years. The participant was compensated 1.60 for participating in our study. The study lasted 6 minutes and participants were randomly distributed across four experimental conditions: 50 in the High Agency-Fair condition, 49 in the High Agency-Unfair condition, 48 in the Low Agency-unfair condition, and 52 in the Low Agency-fair condition.

E. Results

For each analysis, we used a mixed linear model estimated with Restricted Maximum Likelihood (REML) to analyze the survey responses. The model considered the fairness of the context, Fair Type (Fair or Unfair), and the Perceived Agency (High or Low) as the main effects. We also included scenario ID (i.e., the variable indicating 8 different scenarios) as a random effect. For posthoc tests, we used Tukey HSD tests.

1) Manipulation Check: Participants rated scenarios framed as fair significantly lower in unfairness scores ($M = 1.97$, $SE = .20$) compared to scenarios framed as unfair ($M = 4.24$, $SE = .20$; $F(1, 196) = 328.3$, $p < .0001$). Participants also rated robots with human-like decision-making capabilities as having significantly more agency ($M = 3.99$, $SE = .13$) compared to robots described as following third-party programming ($M = 2.58$, $SE = .12$; $F(1, 196) = 174.1$, $p < .0001$), reflecting a clear distinction in perceived agency.

2) Fairness $\times$ Agency on Unfairness Perceptions: We found evidence to support H1. We found an interaction effect between Perceived Agency and Fair Type on the single-item fairness measure ($F(1, 194) = 7.6$, $p = .006$). We observed

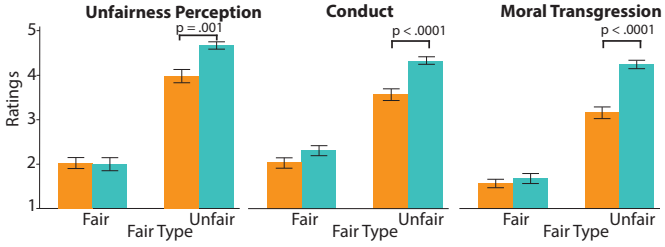

Fig. 2. Average ratings for Unfairness Perceptions (left), Conduct (middle), and Moral Transgression (right). Perceived agency is low ( ) or high ( ). Error bars are standard error.

that in Unfair Type, participants rated a harmful action as significantly more unfair when the robot had higher perceived agency ($M = 4.56$, $SE = .22$) compared to when it had lower perceived agency ($M = 3.90$, $SE = .22$, $p = .0012$). There was no difference in unfairness scores across agency in Fair Type. See Figure 2.

3) Fairness × Agency on Fairness Theory: We found evidence to support H2. There was a significant interaction effect between Perceived Agency and Fair Type on conduct ($F(1, 194) = 4.79$, $p = .02$) and moral transgression ($F(1, 194) = 19.8p < .0001$). When the context was perceived to be unfair, participants reported higher conduct ($M = 4.28$, $SE = .13$) and moral transgression ratings ($M = 4.24$, $SE = .10$) for a harmful action by a robot with higher perceived agency than the robot with perceived low agency ($M = 3.90$, $SE = .22$, $p = .0012$ for Conduct; $M = 3.16$, $SE = .11, p < .0001$ for Moral Transgression). See Figure 2. Although we did not include the reduced welfare component in our hypotheses, our analysis did not find significant effects of Perceived Agency on reduced welfare ($F(1, 194) = .18p = .67$).

4) Fairness × Agency on Organizational Justice Measures: We did not expect Perceived Agency to interact with distributive justice or procedural justice. Nonetheless, we explored how agency influenced perceptions of Organizational Justice, particularly in unfair contexts. We did not find significant effects of agency on either distributive ($F(1, 196) = .38, p = .53$) or procedural justice ($F(1, 196) = .29, p = .59$).

F. Study 1 Discussion

Study 1 confirmed our hypotheses (H1 and H2), addressing our first research question on how perceived agency influences fairness judgments toward robots. We found that only the single-item and Fairness Theory measures demonstrated differences in feelings of unfairness based on the perceived agency of the robot. Participants rated harmful actions by robots with higher perceived agency as more unfair using the single-item measure. Fairness Theory offered a similar but more nuanced result. When humans observed a robot with high perceived agency harm someone, they reported greater feelings that the robot could have acted differently and more ethically. This result likely stems from the heightened expectancy violations associated with a high-agency robot causing harm to a human [58]. This finding was further supported by the finding that outcome-based measures of unfairness (i.e., reduced welfare, distributive justice, or procedural justice) did not vary by the robot's perceived agency. While we expected this lack of difference, we did not formally hypothesize it due to concerns about Type II error. Additionally, participants reported the expected variations in fairness or unfairness perceptions, regardless of how these perceptions were measured. All fairness measures were found to distinguish successfully between fair and unfair a robot's actions.

IV. Study 2

A. Study Design

Study 2 aims to answer the second and third research questions around the perceived intention of a robot's harmful actions and its interaction with perceived agency. This experiment followed a design, measures, and procedures similar to our first experiment, with a few exceptions. First, we added two new scenarios (Lego Challenge and Raise) in addition to the two conditions we already had, resulting in a total of four scenarios. Using the four contexts, we deliberately crafted 2x2 conditions by varying the robot's level of agency and the intention behind its actions. Agency levels were adjusted in a similar matter to Study 1. The harmful action was described as either intentional— aimed at benefiting itself or favoring others in a biased way— or the robot action came as a result of a mistake or system error (unintentional). To measure intention, we asked a single item 5-point Likert scale question: "To what extent do you think the negative outcomes were intentionally caused, regardless of who was responsible?" (1 = Not Intentional at All, 5 = Completely Intentional).

In the **high agency** conditions, negative outcomes were either caused by the robot's independent, biased decisions aimed at benefiting itself or others (**intentional** condition) or resulted from the robot's independent mistakes, without self-serving intentions (**unintentional** condition). In the **low agency** conditions, the robot's programmed actions, designed by a third party, either caused harm to benefit the third party (**intentional** condition) or resulted from system errors unrelated to any intentional design (**unintentional** condition). See Table 3 for a summary of the prompts used.

B. Participants

We recruited 321 participants from the United States on Prolific, but 14 failed the attention check. The final sample included 307 participants who passed (113 male, 182 female, 10 nonbinary, 2 other), with an average age of 38 years. Participants were compensated $1.60 for completing our study. The study lasted 6 minutes and participants who were part of the first experiment were ineligible to participate in this experiment. Participants were distributed across four experimental conditions: 78 in the High Agency-Intentional condition, 78 in the High Agency-Unintentional condition, 72 in the Low Agency-Intentional condition, and 79 in the Low Agency-Unintentional condition.

Table 3. Study 2 prompt summary: left (<>) is high-agency, right is low-agency.

	Scenario 1: Work Schedule	Scenario 2: Warehouse Worker	Scenario 3: Lego Challenge	Scenario 4: Raise
Prompt	You work at a company with a robot that manages work schedules and assigns tasks to employees.	You suffer an injury while working under a managerial robot.	You are competing against a friend in a Lego building challenge. A robot distributes blocks and its assistance amongst both of you.	You are up for a promotion and are being evaluated by a managerial robot.
Intentional Harm	The robot <**decides**/is intentionally programmed> to assign more tasks to you than to your coworker. As a result, you end up working overtime and appearing unproductive to upper management.	The robot <**decides**/is intentionally programmed>to prioritize the company's productivity. The robot insists that you return to work and later your injury gets worse.	The robot <**intentionally decides**/follows its programming> to interrupt you frequently during the challenge, causing you to lose.	The robot <**decides**/follows its programming> to downplay your achievements in the domestic market and prioritizes the international market. As a result, you are denied the raise.
Unintentional Harm	The robot <**mistakenly decides**/follows its faulty programming> to assign more tasks to you than to your coworker. As a result, you end up working overtime and appearing unproductive to upper management.	The robot <**mistakenly decides**/follows its faulty programming> to prioritize the company's productivity. The robot insists that you return to work and later your injury gets worse.	The robot <**makes a mistake** /follows its faulty programming> and interrupts you frequently during the challenge, causing you to lose.	The robot <**mistakenly decides**/follows its faulty programming> to downplay your achievements in the domestic market and prioritizes the international market. As a result, you are denied the raise.

C. Results

For each analysis, we used a mixed linear model estimated with Restricted Maximum Likelihood (REML) to analyze the survey responses. The model considered the Intention (Intentional or Unintentional) and Perceived Agency (High or Low) as main effects. We also included the Scenario ID as a random effect. For posthoc tests, we used Tukey HSD tests.

1) Manipulation Check: As expected, participants rated robots with human-like decision-making capabilities as having significantly higher agency ($M = 3.78, SE = .06$) compared to robots described as following third-party programming ($M = 2.75, SE = .06; F(1, 302) = 126.8, p < .0001$). When a harmful action was presented as intentional, people rated the action as having a higher negative intention ($M = 3.72, SE = .17$) compared to an action presented as a mistake or error ($M = 2.29, SE = .17; (F(1, 302) = 114.49, p < .0001)$).

2) Intention on Unfairness Perception: There was no significant support for H3a, although a marginal effect of intention on the single-item unfairness measure was observed ($F(1, 302.3) = 3.36, p = .06$), indicating a trend but not reaching statistical significance

3) Intention on Fairness Theory: We did not find evidence to support H4a. There were no significant differences in the conduct ratings across the Intention conditions. On the other hand, we found evidence to support H5a. People reported higher moral transgression scores when the harm was perceived as intentional ($M = 3.6$, $SE = .15$) than when it was perceived as unintentional ($M = 3.1$, $SE = .15$; $F(1, 302) = 18.7$, $p < .0001$). Although not hypothesized, we explored the effect of intention on reduced welfare, another component of Fairness Theory. We did not find significant effects of agency on reduced welfare.

4) Intention × Agency on Unfairness Perception: There was evidence to support H3b. We found an interaction between Perceived Agency and Intention on unfairness perception measured through the single item ($F(1, 300) = 5.5$, $p = .01$). Specifically, the harmful action of a robot with a high perceived agency was seen as more unfair if the action was seen as intentional ($M = 4.41$, $SE = .13$) compared to unintentional ($M = 3.97$, $SE = .13$, $p = .01$). There were no differences in unfairness perception across the intention condition when a robot was perceived to have low agency. See Figure 3.

5) Intention × Agency on Fairness Theory: We did not find evidence to support H4b. Contrary to our expectations,

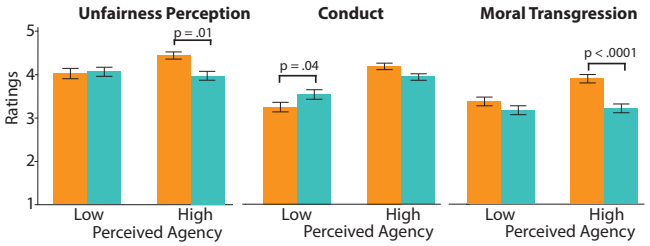

Fig. 3. Average ratings for Unfairness Perceptions (left), Conduct (middle), and Moral Transgression (right). The robot's actions were presented as intentional (■) or unintentional (■). Error bars are standard error.

we found interaction effects on Conduct when the robot's agency was perceived as low ($F(1, 300) = 11.14, p = .0009$). A robot with low perceived agency but whose harmful actions happened due to unintentional program errors received higher conduct ratings ($M = 3.54, SE = .17$) than a robot with low perceived agency but whose harmful actions were intentionally programmed by a third party ($M = 3.22, SE = .17, p = .04$). No significant differences were observed when the robot was perceived to have high agency. However, H5b was supported. There was an interaction effect between Perceived Agency and Intention on ratings of moral transgressions ($F(1, 300) = 6.9, p = .008$). When a robot was perceived to have high agency and committed a harmful action intentionally ($M = 3.92, SE = .17$), it received higher moral transgression scores compared to when it committed a harmful action by mistakes ($M = 3.22, SE = .17, p < .0001$). No significant differences were observed when the robot was perceived to have low agency. We did not find significant interaction effects of agency and intention on reduced welfare.

6) Intention, Agency and Their Interaction on Organizational Justice Measures: With the identical reasoning in Study 1, we did not provide any hypotheses regarding distributive justice and procedural justice. Similar to the findings of Study 1, our explanatory analyses showed that there was no evidence supporting that the intention itself and its interaction with agency affect distributive justice and procedural justice.

D. Study 2 Discussion

Our second question focused on how the perceived intention of a robot's action influences fairness perceptions. We found evidence to support that people tended to judge robots more harshly for moral transgressions when harmful acts were seen as intentional (H5a), consistent with previous studies showing

that conscious decisions are perceived as more morally wrong [45]. Our single-item measure showed a marginal difference in unfairness ratings depending on the perceived intention (H3a), suggesting it also captures nuances in fairness related to deliberate choices. Yet, we did not find evidence for H4a, as humans perceive similar conduct levels whether harm is intentional or not.

The analysis of interaction effects between the perceived agency and the perceived intention, addressing our final research question, offered deeper insights into how some fairness perceptions captured the nuances in the perceived intentions. Ratings of unfairness, as measured through a single-item question, were significantly affected by the perceived intention only when high agency was involved (H3b). Similarly, the perceived intention had a greater impact on moral transgression for robots with higher perceived agency (H5b). These findings suggest that people tend to consider the intentions behind actions after perceiving the robot's agency. It is consistent with our understanding that perceived agency reflects the robot's ability to act intentionally [59], while the expression of intentionality is shaped by contextual factors. Interestingly, the perceived intention significantly affected the conduct scores only when the robot was perceived to have low agency, rejecting H4b. It appears that participants viewed system errors as more controllable factors than a third party's malicious intentions, which resulted in higher conduct scores that reflect stronger expectation violation.

Similar to the findings of Study 1, we did not find evidence that distributive justice, procedural justice, or reduced welfare measures were affected by the perceived agency or the intention behind the harmful action. Given that these measures are focused on outcomes, they may overlook subtle nuances in the robot's actions.

V. General Discussion

Although several frameworks have been proposed in HRI to assess fairness perceptions, their differences have not been thoroughly explored. Our study examined how fairness judgments, as measured by different frameworks, varied depending on the robot's perceived level of agency and intention.

Both study 1 and study 2 demonstrated that only some fairness measures captured variations in fairness related to perceived agency and intentions. Specifically, we found that fairness perceptions designed for accountability judgments (such as the single-item question and the conduct and moral transgression assessments from Fairness Theory [4]) were influenced by the robot's perceived agency and action intentions. These results may be explained by the role of expectation violation [58] in the accountability process [40]. When a human-like robot causes harm intentionally, people may sense stronger expectation violations, leading to greater blame. These findings have implications for future research on how to apply fairness measures.

First, the single-item measure reflected differences in fairness based on agency and intention, but did not clarify which aspects of fairness were affected. This measure would be suitable for research focused on the overall fairness perception rather than its specific components.

Second, our findings showed that while reduced welfare focuses on harmful outcomes, conduct and moral transgression capture expectation violations based on the robot's agency and intentions. This indicates that three components of Fairness Theory are interconnected but capture different aspects of fairness judgment. HRI researchers can use the Fairness Theory-based measures to explore how different aspects of fairness judgments influence broader outcomes, such as trust towards a robot, in distinct ways.

Lastly, we found that the organizational justice-based measures (distributive and procedural justice) did not capture nuances related to agency and intention. If researchers are simply interested in distinguishing fair from unfair outcomes or procedures, regardless of the robot's capabilities, these measures would be appropriate.

VI. Limitations

We acknowledge certain limitations to our approach. First, we used online studies with prompts of scenarios where participants did not directly experience the harm from robots, which may raise questions about whether these results would replicate in in-person studies. However, we chose this method to collect large-scale data on fairness judgments, following previous work in both psychology [60] and HRI [29], [43].

Another limitation is the lack of well-established measures for the components of Fairness Theory. Building on the work of Claure et al. [12] and our review of Fairness Theory, we developed new measures, confirmed satisfactory reliability and validity, and observed consistent effects of agency and intention on fairness perceptions. Nevertheless, further validation procedures would enhance their robustness.

VII. Conclusion

Our paper has three main contributions. First, our study shows that fairness perceptions in HRI shift depending on whether people perceive a robot as acting with intentional harm (perceived intention) or as following predetermined programming versus acting independently (perceived agency). Second, we demonstrate that, while various fairness measures have been proposed, only certain ones capture these subtleties in fairness judgments. Third, we present a recently developed measure based on Fairness Theory—designed to assess fairness through three components: reduced welfare, conduct, and moral transgression that captures the subtleties in fairness judgments other frameworks on fairness judgment overlook.

VIII. Acknowledgments

This work was partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) under the Young Investigator Program (Award No. FA9550-24-1-0085). J. G. Trafton was supported in part by ONR and NRL. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. AFSOR or U.S. Navy.

REFERENCES

[1] J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. Porter, and K. Y. Ng, "Justice at the millennium: a meta-analytic review of 25 years of organizational justice research." *Journal of applied psychology*, vol. 86, no. 3, p. 425, 2001.

[2] K. T. Dirks and D. L. Ferrin, "The role of trust in organizational settings," *Organization science*, vol. 12, no. 4, pp. 450–467, 2001.

[3] M. A. Korsgaard, D. M. Schweiger, and H. J. Sapienza, "Building commitment, attachment, and trust in strategic decision-making teams: The role of procedural justice," *Academy of Management journal*, vol. 38, no. 1, pp. 60–84, 1995.

[4] R. Folger and R. Cropanzano, "Fairness theory: Justice as accountability," *Advances in organizational justice*, vol. 1, no. 1-55, p. 12, 2001.

[5] R. Folger and M. A. Konovsky, "Effects of procedural and distributive justice on reactions to pay raise decisions," *Academy of Management journal*, vol. 32, no. 1, pp. 115–130, 1989.

[6] R. Folger, S. Gilliland, D. Steiner, and D. Skarlicki, "Fairness as deonance," *Theoretical and cultural perspectives on organizational justice*, pp. 3–33, 2001.

[7] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 299–308.

[8] M. F. Jung, D. DiFranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–23, 2020.

[9] M. L. Chang, G. Trafton, J. M. McCurry, and A. L. Thomaz, "Unfair! Perceptions of Fairness in Human-Robot Teams," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 905–912.

[10] M. L. Chang, Z. Pope, E. S. Short, and A. L. Thomaz, "Defining fairness in human-robot teams," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1251–1258.

[11] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! An interaction with a cheating robot," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 219–226.

[12] H. Claure, K. Candon, I. Shin, and M. Vázquez, "Dynamic fairness perceptions in human-robot interaction," *arXiv preprint arXiv:2409.07560*, 2024.

[13] O. Ayalon, H. Hok, A. Shaw, and G. Gordon, "When it is ok to give the robot less: Children's fairness intuitions towards robots," *International Journal of Social Robotics*, vol. 15, no. 9, pp. 1581–1601, 2023.

[14] H. Claure, M. L. Chang, S. Kim, D. Omeiza, M. Brandao, M. K. Lee, and M. Jung, "Fairness and transparency in human-robot interaction," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 1244–1246.

[15] M. Brandao, M. Jirotka, H. Webb, and P. Luff, "Fair navigation planning: a resource for characterizing and designing fairness in mobile robots," *Artificial Intelligence*, vol. 282, p. 103259, 2020.

[16] R. Folger and J. Greenberg, "Procedural justice: An interpretive analysis of personnel systems," *Research in personnel and human resources management*, vol. 3, no. 1, pp. 141–183, 1985.

[17] T. R. Tyler, "The psychology of procedural justice: A test of the group-value model." *Journal of personality and social psychology*, vol. 57, no. 5, p. 830, 1989.

[18] Y. Mu and M. Karasawa, "Blame attribution and intentionality perception of human versus robot drivers: Implications for judgments about autonomous vehicles in moral dilemma contexts," *Cogent Psychology*, vol. 11, no. 1, p. 2384298, 2024.

[19] K. Gray and D. M. Wegner, "The sting of intentional pain," *Psychological science*, vol. 19, no. 12, pp. 1260–1262, 2008.

[20] M. C. Gombolay, R. A. Gutierrez, S. G. Clarke, G. F. Sturla, and J. A. Shah, "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams," *Autonomous Robots*, vol. 39, no. 3, pp. 293–312, 2015.

[21] M. Vázquez, E. J. Carter, B. Mcdorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze." [Online]. Available: https://s3-us-west-1.amazonaws.com/disneyresearch/wp-content/uploads/20170228142926/Towards-Robot-Autonomy-in-Group-Conversations-Understanding-the-Effects-of-Body-Orientation-and-Gaze-Paper.pdf

[22] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 61–68.

[23] M. L. Chang, "Optimizing for task performance and fairness in human-robot teams," Ph.D. dissertation, 2022.

[24] D. Doyle-Burke and K. S. Haring, "Robots are moral actors: Unpacking current moral hri research through a moral foundations lens," in *International Conference on Social Robotics*. Springer, 2020, pp. 170–181.

[25] K. Haring, K. Nye, R. Darby, E. Phillips, E. d. Visser, and C. Tossell, "I'm not playing anymore! A study comparing perceptions of robot and human cheating behavior," in *International Conference on Social Robotics*. Springer, 2019, pp. 410–419.

[26] J. Banks, "Good robots, bad robots: morally valenced behavior effects on perceived mind, morality, and trust," *International Journal of Social Robotics*, pp. 1–18, 2020.

[27] A. Litoiu, D. Ullman, J. Kim, and B. Scassellati, "Evidence that robots trigger a cheating detector in humans," in *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, 2015, pp. 165–172.

[28] L. Tanqueray, T. Paulsson, M. Zhong, S. Larsson, and G. Castellano, "Gender fairness in social robotics: Exploring a future care of peripartum depression." in *HRI*, 2022, pp. 598–607.

[29] T. Arnold and M. Scheutz, "Observing robot touch in context: How does touch and attitude affect perceptions of a robot's social qualities?" in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2018, pp. 352–360.

[30] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *arXiv preprint arXiv:1902.11097*, 2019.

[31] J. A. Colquitt and J. B. Rodell, "Measuring justice and fairness," *The Oxford handbook of justice in the workplace*, vol. 1, pp. 187–202, 2015.

[32] S. K. Ötting and G. W. Maier, "The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations," *Computers in Human Behavior*, vol. 89, pp. 27–39, 2018.

[33] L. Londoño, J. V. Hurtado, N. Hertz, P. Kellmeyer, S. Voeneky, and A. Valada, "Fairness and bias in robot learning," *Proceedings of the IEEE*, 2024.

[34] K. Y. Törnblom and R. Vermunt, "An integrative perspective on social justice: Distributive and procedural fairness evaluations of positive and negative outcome allocations," *Social Justice Research*, vol. 12, pp. 39–64, 1999.

[35] T. R. Tyler, "Using procedures to justify outcomes: Testing the viability of a procedural justice strategy for managing conflict and allocating resources in work organizations," *Basic and Applied Social Psychology*, vol. 12, no. 3, pp. 259–279, 1991.

[36] J. Lamont, *Distributive justice*. Routledge, 2017.

[37] E. A. Lind and T. R. Tyler, *The social psychology of procedural justice*. Springer Science & Business Media, 2013.

[38] 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 *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 125–132.

[39] B. F. Malle, "Moral judgments," *Annual Review of Psychology*, vol. 72, no. 1, pp. 293–318, 2021.

[40] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychological Inquiry*, vol. 25, no. 2, pp. 147–186, 2014.

[41] 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)*. IEEE, 2016, pp. 775–780.

[42] J. G. Trafton, J. M. McCurry, K. Zish, and C. R. Frazier, "The perception of agency," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–23, 2024.

[43] 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 *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 117–124.

[44] E. Phillips, B. F. Malle, A. Rosero, M. J. Kim, B. Kim, L. Melles, and V. B. Chi, "Systematic methods for Moral HRI: Studying human responses to robot norm conflicts," 2023.

[45] Y. W. Sullivan and S. Fosso Wamba, "Moral judgments in the age of artificial intelligence," *Journal of Business Ethics*, vol. 178, no. 4, pp. 917–943, 2022.
[46] F. Cushman, "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment," *Cognition*, vol. 108, no. 2, pp. 353–380, 2008.
[47] D. G. Johnson, "Computer systems: Moral entities but not moral agents," *Ethics and information technology*, vol. 8, pp. 195–204, 2006.
[48] T. Maninger and D. B. Shank, "Perceptions of violations by artificial and human actors across moral foundations," *Computers in Human Behavior Reports*, vol. 5, p. 100154, 2022.
[49] 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.
[50] B. F. Malle and M. Scheutz, "Inevitable psychological mechanisms triggered by robot appearance: morality included?" in *2016 AAAI Spring symposium series*, 2016.
[51] H. Claure, S. Kim, R. F. Kizilcec, and M. Jung, "The social consequences of machine allocation behavior: Fairness, interpersonal perceptions and performance," *Computers in human behavior*, vol. 146, p. 107628, 2023.
[52] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *Machine ethics and robot ethics*. Routledge, 2020, pp. 225–230.
[53] S. Nørskov, M. F. Damholdt, J. P. Ulhøi, M. B. Jensen, C. Ess, and J. Seibt, "Applicant fairness perceptions of a robot-mediated job interview: a video vignette-based experimental survey," *Frontiers in Robotics and AI*, vol. 7, p. 586263, 2020.
[54] J. K. Swim, E. D. Scott, G. B. Sechrist, B. Campbell, and C. Stangor, "The role of intent and harm in judgments of prejudice and discrimination." *Journal of personality and social psychology*, vol. 84, no. 5, p. 944, 2003.
[55] F. Schwartz, H. Djeriouat, and B. Trémolière, "Judging accidental harm: reasoning style modulates the weight of intention and harm severity," *Quarterly Journal of Experimental Psychology*, vol. 75, no. 12, pp. 2366–2381, 2022.
[56] G. Lima, N. Grgić-Hlača, and M. Cha, "Blaming humans and machines: What shapes people's reactions to algorithmic harm," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–26.
[57] J. A. Colquitt, "On the dimensionality of organizational justice: a construct validation of a measure." *Journal of applied psychology*, vol. 86, no. 3, p. 386, 2001.
[58] J. A. Bonito, J. K. Burgoon, and B. Bengtsson, "The role of expectations in human-computer interaction," in *Proceedings of the 1999 ACM International Conference on Supporting Group Work*, 1999, pp. 229–238.
[59] J. G. Trafton, C. R. Frazier, K. Zish, B. J. Bio, and J. M. McCurry, "The perception of agency: Scale reduction and construct validity," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 936–942.
[60] T. K. Tomova Shakur and L. T. Phillips, "What counts as discrimination? How principles of merit shape fairness of demographic decisions." *Journal of Personality and Social Psychology*, vol. 123, no. 5, p. 957, 2022.