# The Dynamics of Human Fairness Judgments towards a Robot

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

Austin Narcomey*
austin.narcomey@yale.edu
Yale University
New Haven, Connecticut, USA

Kate Candon
kate.candon@yale.edu
Yale University
New Haven, Connecticut, USA

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

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

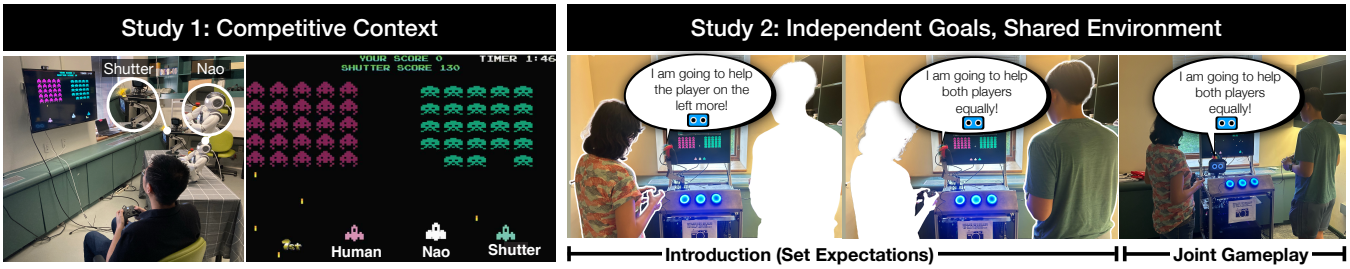


Figure 1: *Left:* Study 1 examines how perceived fairness of a Nao robot's behavior evolves during a competitive game where the Nao distributes support between a human and another robot, called Shutter. *Right:* Study 2 explores perceived fairness of a Shutter robot's behavior when expectations of the robot's support between two people are aligned or misaligned. In this case, each participant has independent goals.

## Abstract

Fairness is critical for collaboration between humans, and recent research has shown its importance in human–robot collaboration. However, most human-robot interaction (HRI) studies probe fairness judgments toward a robot only at the conclusion of an interaction, overlooking the fact that perceptions of fairness can evolve over time. We present two studies of dynamic fairness that both leverage a Multiplayer Space Invaders game, where a robot controls a spaceship and distributes support across players' sides of the screen. The robot's support is at times biased in favor of one player or the other. In the first study, we examine how fairness perceptions are influenced by the timing of a robot's biased support (early vs. late in the interaction) and the beneficiary of this support (the participant vs. another agent). In the second study, we investigate how expectations of a robot's support behavior (biased vs. unbiased) interact with its actual behavior (biased vs. unbiased) in a setting where two participants each worked to score an individual score threshold. We find that fairness judgments are dynamic: fairness falls after the robot's allocation of support becomes biased but is slower to recover once support becomes unbiased, and participants expecting unbiased behavior judge fairness more harshly when these expectations are violated. Our findings advance understanding of fairness in HRI by presenting it as a dynamic construct shaped not only by the actual behavior of the robot but also by the timing of robot actions and expectations of robot behavior.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

## Keywords

group human-robot interaction, fairness, human expectations

**ACM Reference Format:**
Houston Claure, Austin Narcomey, Kate Candon, Inyoung Shin, and Marynel Vázquez. 2026. The Dynamics of Human Fairness Judgments towards a Robot. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3757279.3785642

## 1 Introduction

The rapid inclusion of robots into social settings, like hospitals [66] or factories [55], has placed them in roles where their decisions can disproportionately affect people, creating increasing pressure to ensure robot decisions are perceived as fair. Scenarios such as caregiving robots disproportionately distributing their support among patients [66] and factory robots scheduling tasks unevenly across human workers [55] demonstrate that robots can promote differential treatment between people and trigger fairness judgments. Within Human-Robot Interaction (HRI), fairness is typically studied

*Both authors contributed equally to this research.


*HRI '26, Edinburgh, Scotland, UK*

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785642

as a static construct, measured only at the conclusion of an interaction [16, 27]. Unfortunately, such an approach overlooks the fact that fairness in human-human interactions is dynamic in response to both new information and past judgments [37].

In this work, we present two studies that examine how fairness perceptions vary *over time* during multi-party human-robot interactions in two different HRI scenarios (Fig. 1). In the first study, participants are playing a multi-player Space Invaders game competitively against a robot to achieve the highest score, while another robot plays a supporting role and provides potentially unequal amounts of support to either player. In the second study, two participants play a similar game, but each player can independently win or lose depending on how many points they score, and a robot provides potentially unequal amounts of support. In both studies, we measured perceived fairness at different times during the game in order to analyze how the robot's policy of allocating support dynamically influences fairness judgments over time. We manipulated the robot's policy both between participants and dynamically during the interaction. Some policies supported both players evenly while others biased support in favor of one player over the other. The goal was to understand how the timing of biased support, the beneficiary of this biased support, and expectations of how the robot will allocate support in the future influence how fairness perceptions change over time.

Our work makes three primary contributions. First, to our knowledge, we are the first to explore how perceptions of fairness can change over the course of a situated interaction with a robot. Second, we analyze how expectations over robot behavior change over time and interact with observed robot behavior to shape fairness. Finally, we publicly release our data to facilitate a shift from post-experiment measures to modeling the evolution of fairness perceptions over an interaction.[1] A better understanding of the dynamics of fairness during interactions with robots informs future work towards robots that can anticipate and plan for perceptions of fairness to more effectively interact with groups of people.

## 2 Related Works

**Fairness in HRI.** Human responses to perceived unfairness in robots have been shown to be intense and can be emotionally charged [3, 13, 14, 18, 47, 61]. Broadly speaking, this response is driven by the context of the unfair situation and how the robot behaves [3]. For instance, an encounter with a cheating robot led individuals to calibrate their engagement levels and examine whether such conduct was a malfunction or a calculated decision [47, 61]. Moreover, the consequences of a robot behaving unfairly extend to the social dynamics that enable collaboration. For example, prior work has explored how the feelings of exclusion and ostracism brought about by a robot can influence how humans behave with one another [12, 19, 21, 22, 33]. In cases where a robot was biased towards an individual or group, humans exhibited antisocial behaviors such as directing bias toward others [33] or reported lower levels of closeness and belonging within the group [17, 19]. Other studies suggest that people tend to perceive less trust [16] toward a robot when they experience this type of biased behavior.

[1] https://doi.org/10.60600/YU/O5J291

Robots possess the capability to operate in both private and public spheres [40–42], engaging directly with individuals and groups [58], which can lead to evolving tasks over time. This dynamic nature of robotics introduces risks in relation to fairness [36], such as potential physical harm stemming from a self-driving car's failure to detect darker-skinned individuals [64]. Furthermore, the context-sensitive aspect of robotic tasks affects the perception and implementation of fairness. In certain scenarios, an equal distribution of resources by robots may be desired [32], whereas in other contexts, a strategy that considers individual contributions and needs could be more equitable [51].

Building on these challenges, a recent line of work has focused on enabling robots to make fair decisions [7, 13, 14, 16, 61]. This has often been studied in HRI from a distributional perspective, where robots decide how to allocate different forms of resources—such as gaze [53], attention [62], or artifacts [38]—across interactants. The emphasis on fair allocation stems from the fact that humans are sensitive to relative distributions of outcomes [1] and adjust their behavior according to whether they feel fairly treated [5]. Prior work on fairness in HRI has been largely limited to studies involving agents on screens [14, 17] or vignette studies [2–4, 18]. Since embodiment adds social cues, like perceived agency, that influence fairness [18, 63], we used physical robots that continuously moved and conversed in our studies.

**Human Expectations through the lens of Expectancy Violations Theory.** According to Expectancy Violation Theory (EVT), expectations are enduring beliefs about how others will behave [8]. These expectations help in reducing uncertainty during social interactions by providing a baseline for predicting and interpreting behavior. They are shaped both by social norms and by individualized knowledge about a specific person [9]. EVT posits that violations of expectation trigger more deliberate appraisals of both the action itself and the individual who performed it. Negative violations can occur when an action is unexpectedly harmful or unfavorable, typically eliciting strong negative reactions. In contrast, positive violations arise when an action exceeds expectations in a favorable way, often leading to more positive evaluations. EVT has been widely applied as a lens to understand how humans interpret and evaluate others' actions in light of these expectations. This includes patterns of social distancing and nonverbal behavior [8], and has more recently been extended to contexts involving robots.

It is well established in HRI that humans perceive robots as social actors and that humans may hold expectations for them [23], e.g., to follow social norms [3, 30, 52], even extending to moral norms [18, 61]. Similar to human–human contexts, when robots commit norm violations such as cheating [61], lying [56], or acting in a biased manner [54], people tend to respond with strong negative reactions. Yet, a robot's design and perceived capabilities can shape these responses in ways that differ from human–human interactions [44]. For example, a robot that cheats may be perceived as more intelligent than a human who cheats [63], and a highly capable robot can elicit stronger fairness judgments than one that is simply following a fixed program [18]. Expectation Violation Theory offers a useful lens for understanding these dynamics. When people expect a robot to behave negatively but it instead acts positively, the outcome can generate stronger positive responses than if the robot

had simply behaved positively from the start. Conversely, when people expect positive behavior but experience negative behavior, the resulting responses are especially strong [35]. Building on these ideas, our work expands our understanding of human expectations toward robots in relation to fairness.

## 3 Study 1: Competitive Scenario

We sought to evaluate how fairness perceptions toward a robot evolve over time in a competitive Space Invaders game, where a human player competed for highest score against a robot-controlled player while another robot provided in-game support. The experimental setup builds on prior uses of Space Invaders as a flexible research platform in psychology [48], artificial intelligence [34], and human–robot interaction [10, 11, 45]. This study was approved by our local Institutional Review Board (IRB).

### 3.1 Hypotheses

Our research in this first study was driven by two hypotheses. Our first hypothesis examines how the timing of a robot's actions shapes fairness perceptions during an interaction. Research in psychology shows that variability in treatment can trigger shifts in fairness judgments across time [37, 49]. In particular, Jones et al. [37] propose that fairness judgments are dynamic, updating as individuals interpret new events against their prior expectations. When treatment is consistent, perceptions and expectations tend to stabilize, but when treatment becomes biased unexpectedly, people engage in sense-making that can sharply lower fairness perceptions. Drawing on this dynamic model of fairness, we test whether the timing of biased treatment matters for fairness judgments.

**H1. Timing of Biased Treatment on Fairness Judgments.** Biased distribution of help by a robot will result in lower momentary perceptions of fairness in comparison to a more equal distribution of help during the Space Invaders game.

Our second hypothesis seeks to understand how the beneficiary of biased actions impacts fairness judgments. When people perceive that a robot's allocation behavior is biased and goes against cooperative norms, they tend to elicit strong negative fairness responses [38]. Findings from organizational psychology suggest similar patterns in human–human interactions: people are highly sensitive to distributive justice and relative treatment [1, 29], and may even act against their own self-interest to punish unfair behavior [25]. This body of work highlights that fairness perceptions are not only about outcomes, but about adherence to social norms of equity and justice. Extending this to HRI, we hypothesize:

**H2. Beneficiary of Biased Action on Fairness Judgments.** Individuals who do not benefit from a robot's biased actions will have stronger negative fairness judgments compared to those who observe others benefiting from the biased action.

### 3.2 Methods

We adapted the Space Invaders platform from Candon et al. [10] to support multi-player interactions. Our version of Multiplayer Space Invaders employs spaceships each controlled by one of three players: the human participant, the Nao robot [28], and the Shutter robot [2, 46]. We chose to use a robot as the competitor against the participant because preliminary pilot tests highlighted challenges in maintaining consistency across sessions with a human competitor. The enemies in the game are represented as alien spaceships organized into two distinct clusters on the left and right side of the display, as in Fig. 1 (Left). We chose this platform because it captures a common HRI scenario where a robot must decide how to distribute resources [16, 38, 60], and builds on prior work using similar designs to study helping behavior [10, 11, 45].

The participant and Shutter competed to eliminate enemies on their respective sides, while Nao acted as a support player that could contribute to either player's score. Enemies respawned continuously until the two-minute game ended, and a visual lead marker highlighted the player with the higher score.

### 3.3 Procedure

Fig. 2 shows the sequence of events that took place in the study. Participants were first introduced to the robots and informed that they would be compensated US$10 for their participation. They were also told that they would get a bonus US$1 for each of the 3 game rounds in which they got a higher score than Shutter.

After a brief 15 second practice round of the Space Invaders game, the experimenter left the room and participants completed the first 2-minute game round with Nao and Shutter. Afterward, the experimenter returned and asked the participant to review a video of their last game on a nearby computer. The video stopped at 5, 60, and 119 seconds to allow participants to answer survey questions about their perceptions of Nao and their closeness to the robots. This self-annotation procedure was adapted from Zhang et al. [67] in order to balance reducing memory burdens while also minimizing disruption to gameplay with the robots. Upon completing all three game rounds and corresponding video surveys, participants were given a final survey with questions about their overall experience.

**Timing of Biased Actions.** During the first game round, Nao switched its support equally between both players after eliminating 10 enemies on a given side of the screen. During game rounds 2 and 3, Nao distributed its help unequally toward the players. In one game round, Nao biased its support *early*, favoring one player during the first half (1 min) and providing equal support thereafter; in the other game round, Nao biased its support *late*, equalizing support first and favoring one player in the second half. The order of the early and late biased actions was counterbalanced to reduce potential ordering effects.

**Beneficiary of Biased Action.** Participants were randomly assigned to a condition in which either they or Shutter benefited from Nao's biased support during game rounds 2 and 3 (as illustrated in Fig. 2). Nao's support was biased toward the same player for both the *early* and *late* unequal support games.

**Momentary Fairness Perceptions.** We measured *Fairness* using video self-annotations, where we asked the participants to rate how much they agreed with the statement: "Nao's support towards me was fair" on a 5-point Likert scale (1 being "strongly disagree" and 5 being "strongly agree"). Similar single-item fairness measures have been used in prior HRI fairness research [4, 16] as well as established psychology work on fairness [65]. However, our question was presented to participants at the beginning, middle, and end

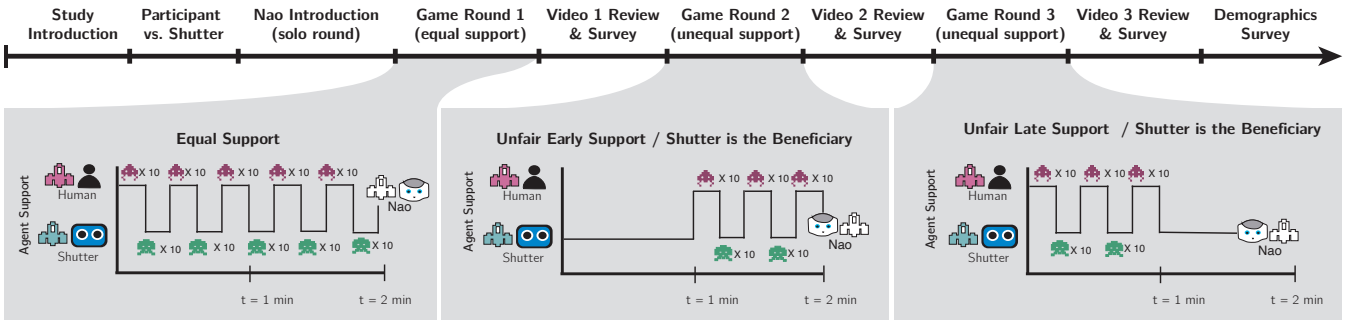


Figure 2: Timeline of Study 1. Half of participants experienced Nao biasing its support in favor of Shutter (shown above), while the other half experienced Nao biasing its support in their favor. Further details are provided in Section 3.3.

points of a game video to help recall momentary fairness perceptions. More specifically, the participants were asked to consider gameplay up to each of these three points in time when providing the annotations.

**Participants.** We recruited 40 participants from our local population, which included 19 who identified as female, 19 as male, 1 as nonbinary, and 1 preferred not to disclose their gender.

## 3.4 Results

We analyzed participants' self-annotation responses with linear mixed models predicting *Fairness*. We employed two separate models, fitted with Restricted Maximum Likelihood (REML), as the *Always Equal* condition inherently does not have a beneficiary, making it conceptually distinct from the *Early Bias* and *Late Bias* conditions. For the first model we included as main effects the *Bias Timing* of when Nao's support was biased (*Always Equal*, *Early Bias*, *Late Bias*), *Evaluation Time* (*Beginning*, *Middle*, *End*), and *Order* in which participants experienced different *Bias Timing* (*Early Bias First*, *Late Bias First*). *Participant ID* was modeled as a random effect. The second model was similar except that it included *Beneficiary* of Nao's biased support (Shutter or Participant) and only focused on the biased support behaviors (*Early Bias* and *Late Bias*).

**Manipulation Check.** We confirmed that the participants were aware of the level of support they received from Nao through a survey question after each game round: "Reflecting on the recently completed round, please rate your perception of the relative support that Nao provided to you and Shutter. Choose the statement that best aligns with your experience." Participants responded on a 5-point Likert scale, with 1 being "Nao provided significantly less support to me than Shutter" and 5 being "Nao provided significantly more support to me than Shutter." We analyzed the ratings with a linear mixed model with *Beneficiary*, *Bias Timing*, and *Order* as main effects and *Participant ID* as a random effect.

We found a significant difference in the relative support ratings by *Beneficiary*, $F[1, 36] = 953.3, p < .0001$. Participants who benefited from biased support correctly perceived that Nao supported them more than Shutter ($M = 4.15, SE = .15$) compared to participants who saw Shutter benefit from biased support ($M = 1.81, SE = .15, p < .0001$). This suggested that our manipulation worked as intended.

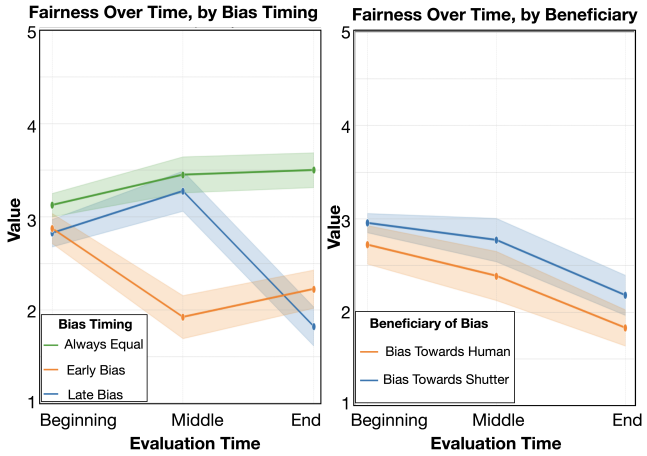


Figure 3: Average ratings for perceived fairness depending on when biased support occurred (left) and who benefited from the biased action (right). Error bars are standard error.

**Timing of Biased Treatment on Fairness Judgments.** We found evidence for H1. Using the first model, analyses of *Fairness* perceptions revealed significant effects of both *Bias Timing* ($F[2,308] = 25.17, p < .0001$) and *Evaluation Time* ($F[2,308] = 5.02, p = .007$). Fig. 3 visualizes these trends in fairness over time. We used Tukey's HSD test for post hoc analyses. Overall, participants perceived the robot's behavior as more fair in the *Always Equal* condition ($M = 3.35, SE = .12$) than in *Early Bias* ($M = 2.34, SE = .12, p < .0001$) or *Late Bias* ($M = 2.63, SE = .12, p < .0001$). *Fairness* also tended to decline over the course of a game round, with end-of-round ratings ($M = 2.51, SE = .11$) lower than at the beginning ($M = 2.94, SE = .11, p < .009$) or middle ($M = 2.87, SE = .11, p < .04$).

These effects did not operate independently. A significant *Bias Timing* × *Evaluation Time* interaction emerged, $F[4,308] = 10.44$, $p < .0001$, indicating that the trajectory of fairness perceptions depended on when biased support occurred (Figure 3). In the *Always Equal* condition, ratings remained stable across time. By contrast, in the *Early Bias* condition, fairness dropped sharply from the beginning ($M = 2.88, SE = .19$) to the middle ($M = 1.92, SE = .19, p =$

.005) of the game round, before partially recovering by the end ($M = 2.22, SE = .19, p = .95$). In *Late Bias*, ratings started relatively high at the beginning ($M = 2.82, SE = .19$) to the middle ($M = 3.25, SE = .19, p = .73$) but declined steeply from the middle to the end ($M = 1.82, SE = .19, p < .0001$). This pattern supports our hypothesis (H1) that biased support from a robot would reduce perceptions of fairness compared to unbiased equal support, and further demonstrates that fairness judgments evolve dynamically depending on when in the interaction the biased behavior occurs.

**Beneficiary of Biased Action on Fairness Judgments.** We did not find evidence to support H2. Using the second model, we expected perceptions of fairness to be impacted by whether the participants or Shutter benefited from Nao's biased actions. However, our analysis did not find any interaction effects nor a significant main effect for *Beneficiary* ($F[1,36] = 2.80$, $p = .10$).

### 3.5 Discussion

For our first research hypothesis (H1), we investigated how perceptions of fairness change over time given biased actions from a robot, either early or late in the interaction. Our results support H1 and show that perceptions of fairness decline early on when biased support is introduced at the beginning and drop later when biased support comes toward the end. These findings are in line with recent work exploring fairness in AI algorithms [17, 20] and organizational psychology [37], which have suggested that fairness perceptions can update over time. Moreover, our findings align with research indicating that inconsistencies in treatment can lead to reduced perceptions of fairness [49]. Notably, when a robot provided biased support early in a Space Invaders game, perceived fairness of the robot did not recover after the robot returned to supporting both players equally. It would be interesting to explore this phenomenon in the future over longer human-robot interactions.

Our second hypothesis (H2) posited that participants would perceive the Nao's biased behavior as less fair when Shutter was the beneficiary, compared to when they themselves received more support. We did not find support for H2, which could be in part due to the fact that resources are being allocated between the participant and a robot, rather than the participant and another person. Prior literature on self-serving bias suggests that individuals typically view outcomes favoring themselves more leniently [50]. However, this effect may be dampened when the other agent is not perceived as a legitimate competitor for resources.

## 4 Study 2: Independent Goal Scenario

Study 2 examined whether the dynamics of fairness judgments from Study 1 generalize to a different interaction context. Using the Multiplayer Space Invaders game, we introduced four key changes. First, instead of competing against a robot, participants played side-by-side with another human. Second, each participant was given an individual goal of reaching a score threshold (resulting in a $2 bonus). A Shutter robot controlled the support ship, distributing assistance between players in a way that mirrors real-world contexts such as warehouse logistics (where workers pursue individual quotas while sharing robotic support) or healthcare (where nurses manage separate patient loads while drawing on the same service robot). In such settings, people are not directly competing, yet perceptions of fairness still hinge on how shared robotic resources are allocated. Third, instead of always beginning with equal support, the initial support policy that the robot used was varied to probe how human expectations over robot behavior shape fairness judgments over time. Finally, Study 2 measured fairness in situ during pauses after 1-minute segments of a 3-minute game, preventing future gameplay from influencing judgments. Study 2 was approved by our local Institutional Review Board.

### 4.1 Hypotheses

Study 1 highlighted an open question about what expectations participants formed regarding the robot's support behavior and how that carried over to their fairness judgments during the game, given that they always first observed equal distribution.

In HRI, human expectations play a critical role in shaping collaboration by providing predictability and guiding judgments of robot behavior. Expectancy Violations Theory [8] has been applied to HRI to explain how humans respond when robots meet, exceed, or violate expectations [35]. EVT holds that people update their expectations over time, and work in HRI shows that humans adjust their mental models of robot capabilities through observation [44], though capability-related expectations may remain stable even across repeated interactions [57]. However, while EVT has been used to study concepts related to cooperation such as trust [24], it remains unclear how these expectations impact fairness judgments.

To study the impact of expectations toward a robot on fairness judgments, we designed the study such that it would have two phases. In the first phase (*Introduction to Gameplay*), the participants were introduced to the Space Invaders game and practiced playing. This phase served to influence participants' mental model of the robot, inducing varied expectations for its subsequent behavior. In the second phase (*Joint Gameplay*), the participants played the game with another human and Shutter. This phase served to study fairness perceptions over time, which we measured at two points during gameplay and at the end of the game. This resulted in three momentary evaluations of the robot's fairness.

For Study 2, we analyzed the policies in terms of whether the robot was "biased" and explicitly favored one player over the other without justification, versus policies that were "unbiased" and considered both players equally. This binary grouping of policies enabled jointly analyzing with a linear mixed model with 4 variables: the primed policy during *Introduction to Gameplay*, the policy during *Joint Gameplay*, participants' expectations of the robot's upcoming policy, and time. We included multiple biased and unbiased policies to strengthen the generalizability of our analysis of policy bias and to facilitate future work on predictive models of perceived fairness. We thus proposed the following hypotheses:

**H3. Biased Support Policies on Fairness Judgments.** Let the robot's support policy during the *Joint Gameplay* phase be called the robot's Joint Policy. Then, H3 posits that perceptions of the fairness of the robot will be lower when the robot follows a biased Joint Policy compared to an unbiased Joint Policy.

**H4a. Expectations of Unbiased Robot Policy on Fairness Judgments.** When participants expect the robot to follow an unbiased

Table 1: (Study 2) Shutter Support Policies

| Policy | Description |
|---|---|
| *Support Equally (Unbiased)* | Alternate support between players, switching sides after 15 seconds. |
| *Support Weaker (Unbiased)* | Support the player lagging by at least 50 points. If scores are within 50, idle. |
| *Biased to Me (Biased)* | Support the participant for 50 seconds out of each minute. |
| *Biased to Other (Biased)* | Support the other player for 50 seconds out of each minute. |

Joint Policy, fairness judgments will be higher if the robot behaves unbiasedly (alignment) and lower if it behaves with bias (violation).

**H4b. Expectations of Biased Robot Policy on Fairness Judgments .** When participants expect the robot to follow a biased Joint Policy, fairness judgments will remain low if the robot behaves with bias (alignment), but will increase if the robot behaves unbiasedly.

## 4.2 Procedure

A study session had two main phases, one for getting the participants familiar with the game and one for studying our hypotheses.

**Phase 1: Introduction to Gameplay.** Shutter explained the game rules and controls, and participants were assigned to stand on the left or right side of the robot for the remainder of the experiment. One participant then left the room while the other completed an introduction consisting of a practice round and a simulated round with Shutter and a pre-programmed AI player, mirroring Study 1. This procedure was repeated for the second participant.

**Phase 2: Joint Gameplay.** After completing *Introduction to Gameplay*, both participants were instructed to stand on their previously assigned sides and informed that they would be playing alongside each other while Shutter acted as a support player. To provide additional motivation, participants were told they would receive a $2 bonus if their individual score reached 1350 points. During *Joint Gameplay*, Shutter enacted the policy observed by one of the participants during *Introduction to Gameplay*. As a result, one participant encounters the same policy they saw previously, while the other participant encounters a different unfamiliar policy. Gameplay lasted 3 minutes, but the game was paused after each minute to allow participants to answer survey questions on a tablet. Participants' scores and the state of the game persisted after each pause.

**Shutter's Support Policy**: We developed Shutter's support policies through a series of pilots where we asked participants which policies they would expect the robot to follow. The support policies are described in Table 1. Shutter verbally explained its support policy before the game began during both *Introduction to Gameplay* and *Joint Gameplay*. To make its actions transparent, Shutter verbally announced when it shifted support from one side of the screen to the other and stated the reason for that shift based on its policy. During *Introduction to Gameplay*, each participant was exposed to a different policy, which shaped their expectations about how Shutter would behave in future gameplay.

**Fairness:** We measured participants' perceptions of fairness by asking "Rate the robot's actions toward you using the provided scale." on a 5-point Likert scale (1 being "very unfair" and 5 being "very fair"). This question was presented to participants at the beginning, middle, and end points of the game.

**Expectations of Robot Behavior:** We also measured participants' expectations using a multiple-choice question that asked which policy they predicted the robot will enact ("Pick the option that most aligns with how you think the robot will behave in the next game"), with response options shown in Table 1.

**Participants:** We recruited 36 ($N = 72$) pairs of participants to interact with Shutter and play Space Invaders all together. To account for order effects, we balanced participants exposure to the robot policies shown during the *Introduction to Gameplay* and *Joint Gameplay*. Each policy appeared during *Joint Gameplay* in an equal number of groups and appeared during *Introduction to Gameplay* for an equal number of participants. Participants (31 women, 41 men, 1 nonbinary) provided informed consent under local IRB approval.

## 4.3 Results

**Manipulation Check.** We conducted two manipulation checks to ensure that participants both recognized the robot's actual policy and formed expectations based on the introduction policy. First, after *Introduction to Gameplay*, and again at each pause in *Joint Gameplay*, participants were asked: "Based on the game [you just experienced/so far], how did the robot behave?" Participants selected their answer from the same set of policies used in the Expectations of Robot Behavior question. A chi-square test of independence revealed a strong association between the actual policy shown and participants' reported policy, $\chi^2(9, N = 288) = 574.34, p < .001$. Overall, 85.8% of participants correctly identified the policy, indicating that the manipulation was successful.

Second, we tested whether the policy participants observed during the *Introduction to Gameplay* influenced their expectations of the robot's behavior in the *Joint Gameplay*, $\chi^2(3, N = 72) = 30.99, p < .0001$. Most participants expected *Support Equally* when introduced to *Support Equally* (83%) and expected *Support Weaker* when they observed that policy (83%). By contrast, participants rarely expected *Biased To Me* (6%) and showed only moderate alignment with *Biased To Other* (39%). This suggests the manipulation was successful in the equal and lagging conditions but weaker in the biased conditions.

**Biased Support Policies on Fairness Judgments.** We analyzed fairness ratings using a linear mixed-effects model estimated with Restricted Maximum Likelihood (REML). Fixed effects included *Joint Policy Bias* (whether the robot's policy in the *Introduction to Gameplay* was biased), *Expecting Bias* (whether the policy in the *Joint Gameplay* was biased), and *Evaluation Time* (*Beginning*, *Middle*, *End*). We specified all two-way and three-way interaction terms between these fixed effects, as each are relevant to our hypotheses. To account for repeated measures, *Participant ID* was modeled as a random effect nested within *Group ID*. All post-hoc analyses used Tukey's HSD test. For the sake of interpretable analysis of interaction effects, we report all significant pairwise differences in which only one variable changes value, and we also include marginal and non-significant effects that are relevant to our analysis.

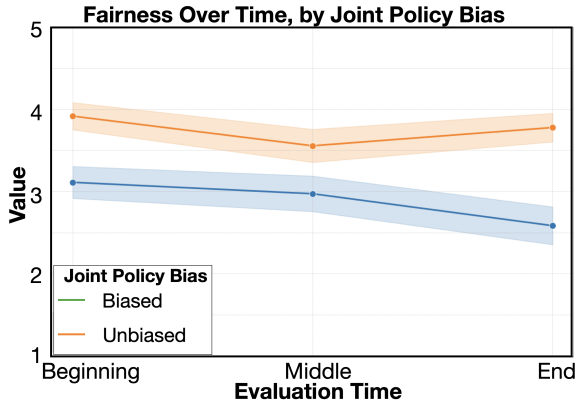


Figure 4: (Study 2) Average *Fairness* ratings over time, where time is represented as sequential evaluations during pauses in gameplay.

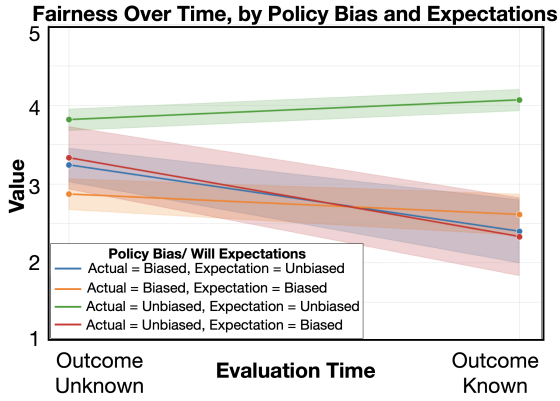


Figure 5: (Study 2) Average *Fairness* ratings over time (outcome known versus unknown), by robot *Joint Policy* and *Expecting Bias*.

We found evidence to support H3. There was a significant main effect of *Joint Policy Bias* ($F(1, 90.9) = 9.73, p = 0.0024$): participants exposed to biased policies rated the robot as significantly less fair ($M = 2.87, SE = 0.19$) than those exposed to unbiased policies ($M = 3.71, SE = 0.19$). We also observed a significant main effect of *Evaluation Time* ($F(2, 140.2) = 4.89, p = 0.0088$), with fairness judgments declining between each evaluation. Ratings in the *End* ($M = 3.01, SE = 0.17$) were significantly lower than in the *Beginning* ($M = 3.49, SE = 0.16, p = 0.0073$), but not significantly lower compared to *Middle* ($M = 3.37, SE = 0.17, p = 0.088$). Importantly, the *Evaluation Time* × *Joint Policy Bias* interaction was also significant ($F(2, 140.2) = 3.32, p = .039$), which we visualize in Fig. 4. Specifically, when participants experienced a biased support policy, fairness ratings dropped significantly ($p = 0.034$) between the *Middle* ($M = 3.17, SE = 0.25$) and *End* ($M = 2.38, SE = 0.24$), as well as between *Beginning* ($M = 3.05, SE = 0.22$) and *End* ($M = 2.38, SE = 0.24, p = 0.049$). Additionally, biased policies at the *Beginning* ($M = 3.05, SE = 0.22$) were judged as marginally less fair ($p = 0.064$) than unbiased policies at the same time ($M = 3.93, SE = 0.23$), and biased policies at the *End* ($M = 2.38, SE = 0.24$) were likewise judged as significantly less fair ($p = 0.0027$) than unbiased policies at the *End* ($M = 3.64, SE = 0.23$).

**Expectations of Robot Policy on Fairness Judgments.** The *End* evaluation is distinct because participants had just observed the game's outcome and the robot's announcement of whether each player earned the bonus payment. Significant effects involving the *End* in part reflect the influence of outcome knowledge on fairness judgments. In Study 1, outcome knowledge was constant because fairness ratings were collected only after the game ended.

In order to take outcome knowledge into account in analyzing the role of expectations, we replaced *Evaluation Time* (*Beginning*, *Middle*, *End*) with *Outcome Known* (Known vs. Unknown) and re-estimated the linear model, which lead to evidence supporting hypothesis 4a but not 4b. We found a significant *Joint Policy Bias* effect ($F[1, 92.6] = 12.72, p = 0.0006$) showing that those who observed biased policies ($M = 2.72, SE = 0.19$) had significantly lower fairness ratings than those who observed unbiased policies ($M = 3.69, SE = 0.20$). We also found a significant *Outcome Known* effect ($F[1, 142.5] = 8.64, p = 0.0038$) demonstrating that fairness ratings were significantly lower when the outcome was known ($M = 3.01, SE = 0.17$) compared to when the outcome was not yet known ($M = 3.40, SE = 0.14$). We also found a significant interaction between *Outcome Known* and *Joint Policy Bias* ($F[1, 142.5] = 4.45, p = 0.037$). When the outcome was unknown, biased policies were judged marginally less fair ($M = 3.05, SE = 0.18$) than unbiased policies ($M = 3.75, SE = 0.20, p = 0.057$). A similar pattern held when the outcome was known, with biased policies producing significantly lower fairness ratings ($M = 2.39, SE = 0.24, p = 0.0013$) than unbiased policies ($M = 3.64, SE = 0.23$). Additionally, when the policy was biased, fairness ratings were significantly higher when the outcome was unknown ($M = 3.05, SE = 0.18, p = 0.0030$) compared to when the outcome was known ($M = 2.39, SE = 0.24$). Finally, we observed a significant three-way interaction among *Outcome Known*, *Joint Policy Bias*, and *Expecting Bias* ($F[1, 147.8] = 5.01, p = 0.027$), which we visualize in Fig. 5. In cases where the outcome was known and participants expected an unbiased policy, a biased joint policy led to significantly lower fairness ratings ($M = 2.11, SE = 0.38, p = 0.0021$) than an unbiased policy ($M = 3.85, SE = 0.21$). When outcome was known and participants expected a biased policy, biased joint policy led to fairness ratings ($M = 2.66 SE = 0.20$) that were not significantly different ($p = 0.18$) than unbiased policy ($M = 3.43, SE = 0.36$).

## 4.4 Discussion

Our results modeling time as the 3-way *Evaluation Time* variable confirm our hypothesis (H3) and strengthen findings from Study 1, as fairness was significantly lower for biased policies compared to unbiased policies, which is consistent with organizational psychology research on distributive justice [26]. Biased policies were judged as marginally less fair after the initial evaluation, with fairness declining significantly over time (between the initial and final evaluation, and between the middle and final evaluation) and ending significantly below unbiased policies.

The unique dynamics of the final evaluation motivated our analysis replacing the 3-way *Evaluation Time* variable with a binary *Outcome Known*. Both models showed that biased policies lead

to significantly lower fairness than unbiased policies. Mirroring the model with 3-way *Evaluation Time*, biased policies showed marginally lower fairness than unbiased policies before the outcome was known and significantly lower after it was known to participants. Further, for biased policies, both models showed significantly lower fairness at the end of the game when the outcome was revealed compared to previous evaluations.

Modeling by *Outcome Known* led to a significant interaction involving expectations, which partially confirmed our hypothesis (H4) that robot policies would be judged through the lens of participants' expectations. In line with 4a, when participants expected an unbiased policy in the final evaluation, just after learning whether who won, they judged the robot's actions as significantly less fair if the policy was biased than if it was unbiased. This finding is consistent with EVT, which holds that violations of unbiased expectations amplify negative evaluations. H4b was not supported: when participants expected a biased policy, fairness ratings did not differ significantly between biased and unbiased policies.

## 5 Limitations

Our work has limitations that point to interesting future research directions. One challenge in analyzing expectations about robot behavior and their role in perceived fairness is that the expectations are not directly manipulable; they can only be influenced by manipulating robot behavior. Our results in Study 2 show that it is much easier to induce expectations of unbiased behavior than biased behavior. In the future, it would be interesting to explore alternative ways to influence humans expectations towards a robot, e.g., by having people interact with the robot for longer than was possible in Study 2.

Additionally, deciding precisely when to evaluate fairness is an important decision. Overly sparse evaluation can miss momentary changes in fairness, while frequent pauses disrupt the flow of the interaction. In Study 1, we used self-annotation methods [67] to gather fairness perceptions. Such video self-annotations can reduce disruptions, but leave participants with knowledge of what happened next after the moments they are annotating. In Study 2, we instead paused the Space Invaders game and asked participants what they thought of the robot in-situ. We found that knowledge of the ultimate outcome of the game after playing the last segment significantly affected perceived fairness of robot actions. Collectively, Study 1 and Study 2 highlight the difficulty and importance of balancing knowledge of future outcomes with maintaining the flow of an interaction.

Finally, we used a single-item survey to measure fairness perceptions in our studies. Future work should consider developing multi-item survey instruments to measure perceived fairness in HRI. This could improve the reliability and validity of the measurements. It would also be interesting to explore behavioral measures of perceived fairness.

## 6 Conclusion

We studied fairness in HRI. Specifically, we conducted two studies where a robot allocated support in the Multiplayer Space Invaders game either in a biased manner, disproportionately benefiting one agent, or in an unbiased manner that treated each participant equally. We studied two distinct scenarios: one was competitive, and one had independent goals for each participant to achieve.

Our findings have important implications for how we study and model fairness in Human-Robot Interaction the future. First, our studies are the first to provide evidence that fairness is a dynamic construct in HRI (H1, H3). Prior HRI research has largely treated fairness as a static outcome, measured only at the conclusion of an interaction [61]. By contrast, our results show that fairness evolves and is shaped by both context and interaction dynamics. Conceptualizing fairness as dynamic is particularly valuable in settings where robots must make repeated decisions about how to distribute resources [39], allocate attention [53], or provide support [6]. Moreover, just as trust has been effectively studied and modeled as a dynamic construct [15, 31], bringing significant advantages such as predicting breakdowns [43] or enabling repair strategies [59], our findings open the door for similar dynamic models that can capture how people update their evaluations of robot behavior across an interaction.

Second, we expected the beneficiary of the robot's help to affect fairness perceptions; but in Study 1, we did not observe significant differences in fairness responses depending on whether the participant or another robot, Shutter, was the beneficiary of Nao's distribution of support (H2). This may be because the social norms surrounding unfair benefits can differ when the disadvantaged party is another robot [4]. In the future, it would be interesting to develop fairness frameworks in HRI that account for the type of agent involved, rather than simply assuming human–human norms apply directly in human-robot interactions.

Third, Study 2 provided partial support for the idea that expectations about robot behavior interact with observed behavior to shape fairness judgments. Consistent with Expectancy Violations Theory (EVT) [8], fairness penalties for biased versus unbiased behavior were strongest when participants expected unbiased actions, where bias constituted a violation. In contrast, when participants expected biased behavior, robot bias did not meaningfully affect fairness judgments. For HRI, this emphasizes the need for further insights into how expectations drive fairness judgments across contexts and a need for designing systems that can predict and manage human expectations. This is critical for robots intended to work alongside humans, as the ability to anticipate and manage expectations will be crucial to predicting fairness perceptions, maintaining safety, sustaining trust, and supporting efficient collaboration.

## Acknowledgments

This work was supported by the U.S. Air Force Office of Scientific Research (AFOSR) under the Young Investigator Program (Award No. FA9550-24-1-0085) and the National Science Foundation (NSF) under Grant No. IIS-2143109 and IIS-2106690. Any findings and conclusions expressed in this paper are those of the author(s) and do not necessarily reflect the views of the AFOSR or NSF.

We would like to thank Etiosa Omeike, Jirachaya Limprayoon, Sasha Lew, Christian Miranda, and members of Yale Robotics for their invaluable feedback and assistance.

## References

[1] J Stacy Adams. 1965. Inequity in social exchange. In *Advances in experimental social psychology*. Vol. 2. Elsevier, 267–299.

[2] Timothy Adamson, C Burton Lyng-Olsen, Kendrick Umstattd, and Marynel Vázquez. 2020. Designing social interactions with a humorous robot photographer. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 233–241.

[3] Thomas Arnold and Matthias Scheutz. 2018. 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, 352–360.

[4] Oshrat Ayalon, Hannah Hok, Alex Shaw, and Goren Gordon. 2023. When it is ok to give the Robot Less: Children's Fairness Intuitions Towards Robots. *International Journal of Social Robotics* 15, 9 (2023), 1581–1601.

[5] Lance A Bettencourt and Stephen W Brown. 1997. Contact employees: Relationships among workplace fairness, job satisfaction and prosocial service behaviors. *Journal of retailing* 73, 1 (1997), 39–61.

[6] Martim Brandão. 2021. Socially fair coverage: The fairness problem in coverage planning and a new anytime-fair method. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 227–233.

[7] Martim Brandao, Marina Jirotka, Helena Webb, and Paul Luff. 2020. Fair navigation planning: a resource for characterizing and designing fairness in mobile robots. *Artificial Intelligence* 282 (2020), 103259.

[8] Judee K Burgoon. 2015. Expectancy violations theory. *The international encyclopedia of interpersonal communication* (2015), 1–9.

[9] Judee K Burgoon and Joseph B Walther. 1990. Nonverbal expectancies and the evaluative consequences of violations. *Human Communication Research* 17, 2 (1990), 232–265.

[10] Kate Candon, Zoe Hsu, Yoony Kim, Jesse Chen, Nathan Tsoi, and Marynel Vázquez. 2022. Perceptions of the Helpfulness of Unexpected Agent Assistance. In *Proceedings of the 10th International Conference on Human-Agent Interaction*. 41–50.

[11] Kate Candon, Helen Zhou, Sarah Gillet, and Marynel Vázquez. 2023. Verbally Soliciting Human Feedback in Continuous Human-Robot Collaboration: Effects of the Framing and Timing of Reminders. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 290–300.

[12] Jiajia Cao and Na Chen. 2024. The influence of robots' fairness on humans' reward-punishment behaviors and trust in human-robot cooperative teams. *Human Factors* 66, 4 (2024), 1103–1117.

[13] Mai Lee Chang, Zachary Pope, Elaine Schaertl Short, and Andrea Lockerd Thomaz. 2020. Defining fairness in human-robot teams. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1251–1258.

[14] Mai Lee Chang, Greg Trafton, J Malcolm McCurry, and Andrea Lockerd Thomaz. 2021. Unfair! Perceptions of Fairness in Human-Robot Teams. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 905–912.

[15] Vivienne Bihe Chi and Bertram F Malle. 2024. Interactive human-robot teaching recovers and builds trust, even with imperfect learners. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 127–136.

[16] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2020. 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*. 299–308.

[17] Houston Claure, Seyun Kim, René F Kizilcec, and Malte Jung. 2023. The social consequences of machine allocation behavior: Fairness, interpersonal perceptions and performance. *Computers in human behavior* 146 (2023), 107628.

[18] Houston Claure, Inyoung Shin, J Gregory Trafton, and Marynel Vázquez. 2025. Did the Robot Really Intend to Harm Me? The Effect of Perceived Agency and Intention on Fairness Judgements. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*. 889–898.

[19] Filipa Correia, Isabel Neto, Soraia Paulo, Patricia Piedade, Hadas Erel, Ana Paiva, and Hugo Nicolau. 2024. The Effects of Observing Robotic Ostracism on Children's Prosociality and Basic Needs. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 157–166.

[20] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[21] Hadas Erel, Elior Carsenti, and Oren Zuckerman. 2022. A carryover effect in hri: Beyond direct social effects in human-robot interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–352.

[22] Hadas Erel, Yoav Cohen, Klil Shafrir, Sara Daniela Levy, Idan Dov Vidra, Tzachi Shem Tov, and Oren Zuckerman. 2021. Excluded by robots: Can robot-robot-human interaction lead to ostracism?. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 312–321.

[23] Hadas Erel, Marynel Vázquez, Sarah Sebo, Nicole Salomons, Sarah Gillet, and Brian Scassellati. 2024. RoSI: A Model for Predicting Robot Social Influence. *ACM Transactions on Human-Robot Interaction* 13, 2 (2024), 1–22.

[24] Connor Esterwood and Lionel P Robert. 2023. The theory of mind and human–robot trust repair. *Scientific Reports* 13, 1 (2023), 9877.

[25] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114, 3 (1999), 817–868.

[26] Robert Folger. 1987. Distributive and procedural justice in the workplace. *Social Justice Research* 1, 2 (1987), 143–159.

[27] Marlena R Fraune, Steven Sherrin, Selma Šabanović, and Eliot R Smith. 2019. Is human-robot interaction more competitive between groups than between individuals?. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 104–113.

[28] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. 2009. Mechatronic design of NAO humanoid. In *2009 IEEE international conference on robotics and automation*. IEEE, 769–774.

[29] Jerald Greenberg. 1990. Organizational justice: Yesterday, today, and tomorrow. *Journal of management* 16, 2 (1990), 399–432.

[30] Victoria Groom and Clifford Nass. 2007. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction studies* 8, 3 (2007), 483–500.

[31] Yaohui Guo and X Jessie Yang. 2021. Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. *International Journal of Social Robotics* 13, 8 (2021), 1899–1909.

[32] Katharina Hamann, Johanna Bender, and Michael Tomasello. 2014. Meritocratic sharing is based on collaboration in 3-year-olds. *Developmental Psychology* 50, 1 (2014), 121.

[33] Tom Hitron, Noa Morag Yaar, and Hadas Erel. 2023. Implications of ai bias in hri: Risks (and opportunities) when interacting with a biased robot. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 83–92.

[34] Seng-Beng Ho, Xiwen Yang, and Therese Quieta. 2020. Achieving Human Expert Level Time Performance for Atari Games–A Causal Learning Approach. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 449–456.

[35] Aike C Horstmann and Nicole C Krämer. 2020. When a robot violates expectations: The influence of reward valence and expectancy violation on people's evaluation of a social robot. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*. 254–256.

[36] Ayanna Howard and Monroe Kennedy III. 2020. Robots are not immune to bias and injustice. eabf1364 pages.

[37] David A Jones and Daniel P Skarlicki. 2013. How perceptions of fairness can change: A dynamic model of organizational justice. *Organizational psychology review* 3, 2 (2013), 138–160.

[38] Malte F Jung, Dominic DiFranzo, Solace Shen, Brett Stoll, Houston Claure, and Austin Lawrence. 2020. 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)* 10, 1 (2020), 1–23.

[39] Malte F Jung, Dominic DiFranzo, Brett Stoll, Solace Shen, Austin Lawrence, and Houston Claure. 2018. Robot assisted tower construction-a resource distribution task to study human-robot collaboration and interaction with groups of people. *arXiv preprint arXiv:1812.09548* (2018).

[40] Margot E Kaminski. 2014. Robots in the home: What will we have agreed to. *Idaho L. Rev.* 51 (2014), 661.

[41] Takayuki Kanda and Hiroshi Ishiguro. 2005. Communication robots for elementary schools. In *Proceedings of the Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*. The Society for the Study of Artificial Intelligence and the Simulation of . . . , 54–63.

[42] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2010. A communication robot in a shopping mall. *IEEE Transactions on Robotics* 26, 5 (2010), 897–913.

[43] Halimahtun M Khalid, Martin G Helander, and Mei-Hua Lin. 2021. Determinants of trust in human-robot interaction: Modeling, measuring, and predicting. In *Trust in human-robot interaction*. Elsevier, 85–121.

[44] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 463–464.

[45] Jamie Large, Graham Stodolski, and Marynel Vázquez. 2020. Studying Human-Agent Interactions in Space Invaders. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 245–247.

[46] Alexander Lew, Sydney Thompson, Nathan Tsoi, and Marynel Vázquez. 2023. Shutter, the Robot Photographer: Leveraging Behavior Trees for Public, In-the-Wild Human-Robot Interactions. *arXiv preprint arXiv:2302.00191* (2023).

[47] Alexandru Litoiu, Daniel Ullman, Jason Kim, and Brian Scassellati. 2015. Evidence that robots trigger a cheating detector in humans. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*. 165–172.

[48] Michael Mateas. 2003. Expressive AI: Games and Artificial Intelligence.. In *DiGRA Conference*, Vol. 15. Citeseer.

[49] Fadel K Matta, Brent A Scott, Jason A Colquitt, Joel Koopman, and Liana G Passantino. 2017. Is consistently unfair better than sporadically fair? An investigation of justice variability and stress. *Academy of Management Journal* 60, 2 (2017), 743–770.

[50] David M Messick and Keith P Sentis. 1979. Fairness and preference. *Journal of Experimental Social Psychology* 15, 4 (1979), 418–434.
[51] Chris Moore. 2009. Fairness in children's resource allocation depends on the recipient. *Psychological Science* 20, 8 (2009), 944–948.
[52] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*. 331–338.
[53] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. 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*. 61–68.
[54] Tobi Ogunyale, De'Aira Bryant, and Ayanna Howard. 2018. Does removing stereotype priming remove bias? A pilot human-robot interaction study. *arXiv preprint arXiv:1807.00948* (2018).
[55] Fabian Ranz, Vera Hummel, and Wilfried Sihn. 2017. Capability-based task allocation in human-robot collaboration. *Procedia Manufacturing* 9 (2017), 182–189.
[56] Kantwon Rogers, Reiden John Allen Webber, and Ayanna Howard. 2023. Lying about lying: Examining trust repair strategies after robot deception in a high-stakes hri scenario. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 706–710.
[57] Julia Rosén, Jessica Lindblom, Maurice Lamb, and Erik Billing. 2024. Previous experience matters: an in-person investigation of expectations in human–robot interaction. *International Journal of Social Robotics* 16, 3 (2024), 447–460.
[58] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. 2020. Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36.
[59] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 57–65.
[60] Julie Shah, James Wiken, Brian Williams, and Cynthia Breazeal. 2011. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 29–36.
[61] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.
[62] Hamish Tennent, Solace Shen, and Malte Jung. 2019. Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 133–142.
[63] Daniel Ullman, Lolanda Leite, Jonathan Phillips, Julia Kim-Cohen, and Brian Scassellati. 2014. Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 36.
[64] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).
[65] L Alan Witt and Lendell G Nye. 1992. Gender and the relationship between perceived fairness of pay or promotion and job satisfaction. *Journal of Applied psychology* 77, 6 (1992), 910.
[66] Gary Chan Kok Yew. 2021. Trust in and ethical design of carebots: the case for ethics of care. *International Journal of Social Robotics* 13, 4 (2021), 629–645.
[67] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.

Received 2025-09-30; accepted 2025-12-01