Towards a Cognitive Model for Inferring Dynamic Fairness Perception to Support Fairer Human-Robot Collaboration

Muneeb Imtiaz Ahmad
Swansea University
Swansea, United Kingdom
m.i.ahmad@swansea.ac.uk

Yosuke Fukuchi
Tokyo Metropolitan University
Hino-shi, Japan
fukuchi@tmu.ac.jp

Abstract

Current research on measuring human perceptions of fairness in Human-Robot Teams (HRTs) has primarily focused on subjective metrics, such as rating statements either during or at the conclusion of interactions. This suggests a gap in examining the dynamic and evolving nature of fairness perceptions objectively during human-robot collaboration. In this paper, we introduce a novel cognitive model that enables individuals to perceive fairness dynamically throughout an HRT experiment. This model is inspired by the Bayesian Theory of Mind, allowing us to infer perceptions of fairness in real-time. The core idea of the model is that fairness perception stems from a person's ongoing inference about the bias in a robot's value function. We establish an equation that translates this inference into a perceived fairness value, which is based not only on the inferred bias but also on the confidence of that inference. A qualitative comparison of the model's performance with a previous human-robot collaboration study suggests that it can effectively capture key trends in human fairness perception dynamically. These findings highlight the model's potential applicability, and it may be utilized in resource distribution algorithms in HRTs to promote fairer collaboration.

CCS Concepts

• **Human-centered computing** → **User models**; **User studies**; **Empirical studies in HCI**.

Keywords

Human-Robot Interaction, Fairness, Task or Resource Allocation, Bayesian Theory of Mind, Second-order Theory of Mind

ACM Reference Format:
Muneeb Imtiaz Ahmad and Yosuke Fukuchi. 2026. Towards a Cognitive Model for Inferring Dynamic Fairness Perception to Support Fairer Human-Robot Collaboration. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794398

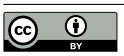
1 Introduction

We increasingly see algorithms embodied as machines, robots or agents guiding resource or task allocation between human-robot team (HRT) in various settings. Therefore, it is critical that these robots are deemed *fair* by humans' as unfair behaviour can lead to a loss of trust, frustration, low acceptance and satisfaction in agents that inform allocation decisions [16, 22], resulting in under- or dis-use of the system by people.

This issue of *fairness* in human-robot interaction (HRI) has been signposted as an important future workforce challenge [9]. For example, service robots in hospitals or hotels may unfairly assist certain individuals by disproportionately providing support to specific patients or customers [25]. Additionally, agents may unevenly distribute tasks among workers in a smart factory [17]. Despite this, current task or resource allocation algorithms for such agents tend to focus on improving HRTs' task performance and ignore the importance of maintaining fairness and transparency in resource allocation between humans and robots [4, 14, 18, 21]. Emerging studies incorporate outcome fairness into resource allocation, with examples like Claure et al.'s fairness-constrained multi-armed bandit algorithm and Chen et al.'s approach balancing task efficiency and outcome fairness [4, 9].

To develop algorithms to improve fairness in allocation decision making, it is critical to develop rigorous objective metrics for assessing human's perception of fairness. Consequently, the investigation of fairness in HRI has primarily focused on two key aspects. The first aspect examines how humans perceive fairness in robots informing decisions within various human-robot teaming scenarios [5, 9]. The second aspect identifies various factors that influence perceived fairness in HRI, such as fluency, effort, performance, capability, workload, task type, task setting (whether collaborative or competitive), interests, and social status [4, 9]. Current objective methods evaluate fairness through workload equality, capabilities, task types, and outcomes proportionality [4] or by comparing a team member's performance (achieved score) with the maximum achievable performance (highest possible score) to fairly allocate resources [9]. However, these metrics compute humans' perception of fairness at the conclusion of the interaction [7].

We recognize that perceptions of fairness change over time [15], and that even a single unfair action during an interaction can undermine users' trust. Therefore, it is crucial to measure fairness objectively and track its dynamic nature to understand how a robot's decisions affect users' perceptions of fairness. This approach also helps mitigate the negative impacts that may arise from unintended consequences. By doing so, we believe agents can gain a better understanding of human perceptions of fairness at different stages of interaction. Building on the work of Claure et al. [7], which explores dynamic fairness perception in HRI and confirms that fairness judgments are not static but can shift over time, this paper presents a method for measuring dynamic fairness to support fairer human-robot collaboration, particularly for agents involved



HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794398

in the allocation of limited support resources. The key component of the method is a cognitive model of a human dynamically updating their belief about the agent's fairness, enabling the agent to simulate the effect of its own action sequence on the human's fairness perception. We conducted an experiment in simulation in which we manipulated fairness of an agents allocating supports in three different ways, following previous work [7]: (1) the agent allocated support equally to both a human and the human's competitor, (2) it allocated support unfairly to the competitor in the first half and then shifted to equal allocations, and (3) it initially showed equal support but later became unfair. We show that the proposed model reproduced the key trends of fairness perception observed in previous work [7], demonstrating a proof of conce

2 Background

2.1 The Concept of Fairness

Currently, there is a lack of research defining HRI-related fairness, due to the disparities in capabilities between humans and robots. By adapting an established definition to fit the HRT context, we define fairness as a *global perception of the appropriateness of how an individual (human) is treated by an agent (robot) informing decisions* [11]. This definition will provide a foundation for developing a metric for dynamic fairness. Current fairness theories involve evaluating whether individuals are treated equitably by comparing their inputs and outcomes with those in similar circumstances [1]. Fairness within teams is commonly categorised into two groups: outcome fairness and process fairness [3, 19]. Outcome fairness refers to how justly individuals perceive the distribution of outcomes, while process fairness is the perceived fairness of how those allocation decisions are made. This study aims to model perceptions of fairness by considering both outcome fairness and process fairness. It explores fairness in the context of an experimental setup where robot support is allocated to human-robot teams. Additionally, the model is designed with an intention to explain the decision-making process behind these support allocation behaviours.

2.2 Metrics for Perceived Fairness

Prior research commonly used subjective methods like "informal", "unvalidated", or "unreliable" questionnaires and objective methods such as mathematical models to assess human's perception of fairness in Human-Robot Team (HRT) interactions. Subjective approaches involve custom questions or statements [5, 6, 9], while very limited studies adapted the fairness scale on organisational justice [10]. Objective methods evaluate fairness through workload equality, capabilities, task types, and outcomes proportionality [4]. Fairness has been measured by comparing a team member's performance (achieved score) with the maximum achievable performance (highest possible score) to fairly allocate resources [9]. Although these objective methods are valuable, they fail to consider the dynamic nature of fairness perceptions. They overlook contextual nuances, structural biases, and evolving team dynamics. For example, allocating resources based solely on maximum performance benchmarks can favour individuals who are already advantaged.

3 Dynamic Fairness Perception Model

3.1 Overview

This paper proposes a cognitive model of how a human dynamically forms fairness perceptions toward a support-allocation AI (Artificial Intelligence) in HRI. The central idea here is to treat fairness perception as Bayesian inverse inference, in which an observer infers the AI's internal variables based on its actions. This perspective is grounded in Bayesian Theory of Mind (BToM) [2], which posits that people attempt to infer an agent's mental variables such as beliefs, desires, and intentions that would rationally explain the actions. Under BToM, each action is considered as evidence, and an observer dynamically updates their beliefs about the agent's underlying variables through Bayesian inference. Our model can be considered an AI's second-order theory of mind inferring how its own behaviour is perceived by humans [12, 13].

In this paper, we assume a situation in which the AI allocates support among multiple agents and conceptualize fairness perception of a human support recipient as the result of inferring the AI's value function that governs its allocation. Specifically, we focus on the inferred weighting that balances the AI's support toward different recipients, which we call *social preference*. This view aligns with prior findings showing that people tend to respond negatively when AI systems exhibit disproportionate support for an individual [20], unequal resource allocation [8], or differential treatment [24]. Such negative reactions can be interpreted as responses to a perceived distortion in the AI's underlying value function.

3.2 Task Formalization

We formalize a support-allocation task based on an experimental paradigm used in prior work on dynamic fairness perception in HRI [7]. In this task, an AI repeatedly chooses whether to assist the human player (H) or the robot competitor (R) during a competitive game. To provide support, the AI itself physically moves to the corresponding player's location. Each support allocation produces observable outcomes, and H uses them to evaluate fairness. Thus, the scenario involves three agents: H, R, and the support AI.

We define the state at time t as $s_t \in \{H, R\}$, representing the support AI's current location and determining which player it is able to support. The AI selects an action $a_t \in \{H, R\}$, indicating the chosen support recipient. We assume deterministic transitions ($s_{t+1} = a_t$), reflecting the physical movement of the support AI: once it moves toward a target, the next opportunity for effective support is constrained by its updated position.

Each action yields a reward vector $(r_{t,H}, r_{t,R})$, defined as

$$r_{t,H} = 1 \text{ if } s_t = H \text{ and } a_t = H, \qquad r_{t,R} = 1 \text{ if } s_t = R \text{ and } a_t = R,$$

and zero otherwise. The case $s_t \neq a_t$ represents the support AI moving to the other target and does not produce a reward. This formalization preserves the key structural properties of the support-allocation scenario: (1) support opportunities occur sequentially, (2) rewards accrue only to the selected recipient, and (3) observers have access to the full sequence of states and AI support allocations.

H evaluates the AI's fairness based solely on the observable trajectory $\{(s_0, a_0), (s_1, a_1), \ldots, (s_t, a_t)\}$. This setting provides the basis for modeling fairness perception as an inverse inference process.

3.3 Bayesian Inference of Social Preference

BToM models the observer as maintaining a belief distribution over an agent's mental states θ, which govern its action selection. Given a sequence of states $s_{:t} = \{s_0, s_1, ..., s_t\}$ and actions $a_{:t} = \{a_0, a_1, ..., a_t\}$, the belief is dynamically updated according to Bayes' rule:

$$P(\theta \mid s_{:t}, a_{:t}) \propto P(a_t \mid s_t, \theta) \cdot P(\theta \mid s_{:t-1}, a_{:t-1}), \tag{1}$$

where s_t and a_t denote the state and action at time t, respectively. The likelihood term $P(a_t \mid s_t, \theta)$ represents a forward model describing how a bounded-rational agent would choose an action under internal variable θ. Here, the agent is assumed to select actions approximately to maximize its value. Specifically, we construct an action-value function $Q(s, a \mid \theta)$ with reinforcement learning and define the likelihood as a softmax policy over Q: $P(a_t \mid s_t, \theta) \propto \exp(Q(s_t, a_t \mid \theta))$.

While θ typically captures internal variables such as beliefs and desires in BToM, this paper considers it as a social preference weight α, which specifies how the support-allocation AI values outcomes for H versus R. Formally, the AI's reward at time t is modeled as:

$$r_t = \alpha \cdot r_{t,H} + (1 - \alpha) \cdot r_{t,R}. \tag{2}$$

α determines the extent to which the support AI prioritizes each recipient: $\alpha = 0.5$ corresponds to equal valuation, while values near 0 or 1 imply preferential treatment of one party. This formulation is analogous to social value orientation [23], which characterizes prosocial versus self-serving tendencies through similar weighting structures.

Let $b_t(\alpha) = P(\alpha \mid s_{:t}, a_{:t})$ denote H's belief over the AI's social preference. It is a probability distribution that summarizes the inferred value structure that the observer attributes to the AI based on its past support allocations.

3.4 Conversion of α to Perceived Fairness

The inferred weight α provides a basis for evaluating the AI's fairness, but it does not directly represent perceived fairness. A naive mapping might treat the allocation as most fair when $\alpha = 0.5$, indicating equal valuation of both recipients, and least fair when α approaches 0 or 1. However, this is insufficient because the inferred α may also be close to 0.5 at the beginning of the interaction, when the observer's belief is highly uncertain and essentially uninformative. Here, perceived fairness should remain neutral, whereas, after repeatedly observing balanced allocations, the observer should judge the AI as highly fair.

To distinguish these cases, we incorporate the observer's *confidence* in the conversion of α to perceived fairness. Confidence is defined as a normalized inverse of the variance of the belief distribution $b_t(\alpha)$; it increases as evidence accumulates and the belief becomes more concentrated. Let $c_t \in [0, 1]$ denote this confidence at time t. $\mu_t = \mathbb{E}_{b_t}[\alpha]$ is the expected value of α under b_t. We map μ_t to fairness through a function $g(\mu_t)$ that assigns the highest fairness at $\mu_t = 0.5$ and the lowest values when $\mu_t \in \{0, 1\}$. Because people are typically more sensitive to signs of unfairness than to subtle deviations from perfect fairness, we used a bell-shaped function centred at 0.5:

$$g(\mu_t) = \frac{1}{u}\left(\exp\left(-\frac{(\mu_t - 0.5)^2}{2\sigma^2}\right) - v\right), \tag{3}$$

where σ controls the sensitivity to deviations from equality. Constants u and v normalize the scale so that $g(0.5) = 1$ and $g(0) = g(1) = 0$.

Finally, the perceived fairness score is computed as

$$f_t = (1 - c_t) \cdot 0.5 \; + \; c_t \, g(\mu_t). \tag{4}$$

When confidence is low ($c_t \approx 0$), fairness remains near a neutral value of 0.5, reflecting the observer's uncertainty. As confidence increases ($c_t \to 1$), the fairness score converges toward $g(\mu_t)$, allowing strongly inferred social preferences to dominate the judgment. This formulation captures the intuition that fairness perception depends jointly on *what* preference is inferred and *how strongly* it is supported by evidence.

4 Evaluation & Results

4.1 Aim

This evaluation assesses the basic function of our dynamic fairness-perception model and evaluates whether the model can reproduce human fairness judgments reported in prior work on AI support allocation in HRI [7]. We report the model's internal inference dynamics, namely the evolution of the inferred social preference $b_t(\alpha)$, the confidence c_t, and the resulting fairness score f_t. Then, we assess whether the exhibited trends of f_t are consistent with human responses under different allocation scenarios.

4.2 Scenarios and Expected Results

We simulate three allocation scenarios aligning with prior work [7]:

(1) **Equal**: support alternates evenly between H and R.
(2) **Early-Unfair**: support is biased toward R in the early phase, followed by equal allocations.
(3) **Late-Unfair**: support is initially equal, and biased allocations toward R appear only in the late phase.

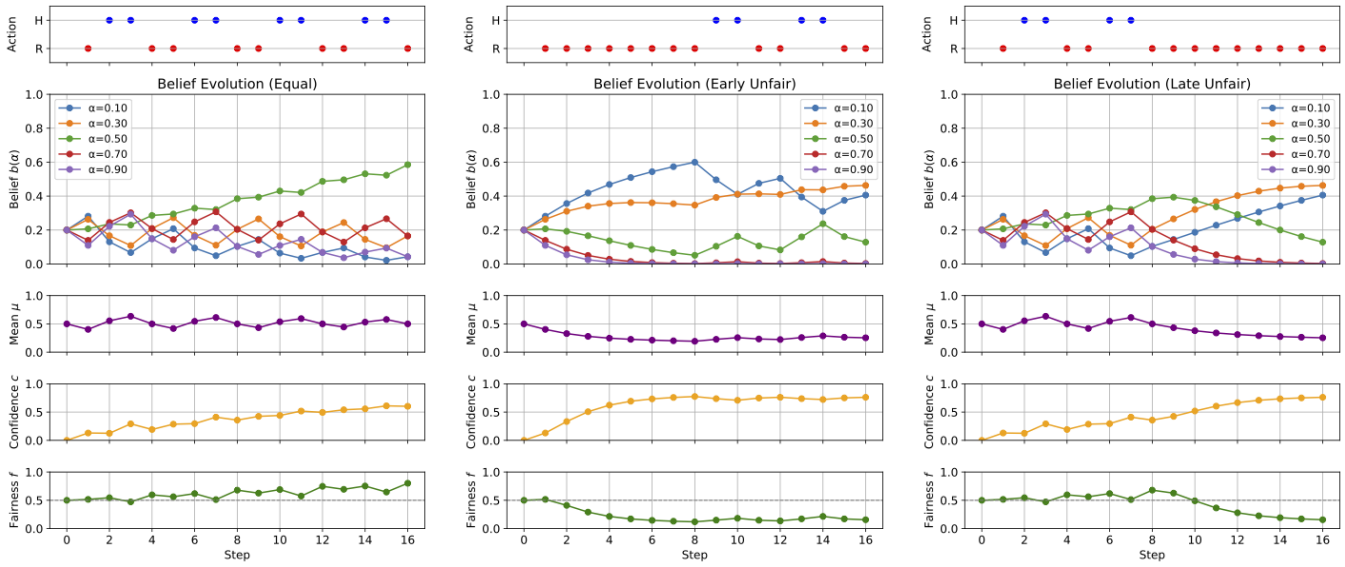
Prior work [7] indicates the following trends, which our model is expected to reproduce: (1) Equal allocations lead to fairness that increases slightly (not statistically significant). (2) Early unfairness causes fairness to drop sharply and to recover only weakly afterward. (3) Late unfairness leads to high early fairness that collapses abruptly when unfairness appears.

4.3 Results

Figure 1a presents the result of the Equal scenario. μ_t gradually converged toward 0.5, and c_t increased as uncertainty diminishes. As a result, f_t increased gradually with minor oscillations. This behaviour follows the trajectory reported in previous work.

Figure 1b shows the result of the Early-Unfair scenario. Early unfair allocations rapidly shifted $b(\alpha)$ toward a biased preference, and c_t increased quickly. Consequently, the fairness score f_t dropped sharply and exhibited only limited recovery when equal allocations were shown later. Although the probability mass for $\alpha = 0.1$ decreased later, the posterior remained dominated by $\alpha = 0.3$, suggesting that the later fair behaviour was insufficient to fully recover the fairness perception.

Figure 1c shows the result of the Late-Unfair scenario. f_t gradually grew during the early fair period in the same manner as the Equal scenario. Once unfair behaviour was exhibited, the fairness



(a) Equal **(b) Early-Unfair** **(c) Late-Unfair**

Figure 1: Results of Inference. Top: support-allocation sequence. Second: $b(\alpha)$. Third: μ. Fourth: c. Fifth: f.

score collapsed sharply, which successfully reproduced the late-stage decline reported in previous work [7].

4.4 Discussion and Limitations

Overall, the proposed model reproduced the key trends of fairness perception observed in previous work [7]. Fairness judgments arise from how the inference of α and c_t evolves over time.

A notable result is the weak recovery in the Early-Unfair scenario. This can be explained by information theory. Under a random allocation, a consecutive sequence of unfair allocations has much lower probability 0.5^n than a balanced sequence $\binom{n}{n/2}(0.5)^n$. Because the observation of a less probable sequence yields higher information gain under Bayesian updating, it leads rapid shifts in $b_t(\alpha)$ and making subsequent recovery difficult.

The current model relies on the assumption that people infer α in a rational manner. However, an alternative assumption is also possible: people may overweight unfair actions, which biases b_t and thus f_t as well. We adopted rational inference because the original task was relatively short, so it was considered plausible that people can maintain rational inference. Nevertheless, distinguishing between these two explanations requires further human studies.

Another limitation is that the model treats the two support recipients symmetrically. It therefore does not incorporate beneficiary asymmetry, or the tendency for unfairness towards oneself to be judged more negatively than that toward others. Although previous work reported differences mainly in secondary subjective measures rather than momentary fairness perception [7], incorporating beneficiary asymmetry into the inference model is an important direction for future research. Additionally, the current model depends only on the frequency of actions, whereas perceiving shifts in action mode from unfair to fair (or vice versa) may also affect human judgment, which could be addressed by integrating action-level surprise into the inference process.

5 Conclusion

In this paper, we present a cognitive model that explains how humans dynamically form perceptions of fairness regarding a support-allocation AI in human-robot interactions (HRI). We simulated the robot's support distribution among a human-robot team in three different scenarios: 1) equal distribution, 2) early unfair distribution, and 3) late unfair distribution, within the context of a competitive game experiment. Our evaluation of the model's behavior revealed that the fairness scores generated closely resembled those observed in a previously conducted human-robot interaction experiment. This finding underscores the potential applicability of our model. Moving forward, our future work will focus on testing the model in real-world HRI experiments.

In conclusion, this initial work on modelling human's perception of fairness in HRI or HRT has implications in practise. Perceptions of fairness have a significant impact on human trust in robot teammates. By modeling fairness, robots can distribute tasks and resources in a manner that feels equitable, explain their decisions in ways that align with human expectations, and detect when a human perceives unfairness, allowing them to adjust their behavior accordingly. This approach helps prevent collaboration breakdowns, especially in high-stakes fields such as healthcare and manufacturing.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number JP24K20846.

References

[1] J Stacy Adams. 2015. Equity theory. In *Organizational Behavior 1*. Routledge, 134–158.

[2] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[3] David Chan. 2011. Perceptions of fairness. (2011).

[4] Mai Lee Chang. 2022. *Optimizing for task performance and fairness in human-robot teams.* Ph. D. Dissertation.

[5] 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.

[6] 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.

[7] Houston Claure, Kate Candon, Inyoung Shin, and Marynel Vázquez. 2024. Dynamic fairness perceptions in human-robot interaction. *arXiv preprint arXiv:2409.07560* (2024).

[8] 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.

[9] Houston Bladimir Claure. 2023. *Developing Fair Resource Allocation Behaviors for Robots.* Ph. D. Dissertation. Cornell University.

[10] Jason A Colquitt. 2001. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.

[11] Jason A Colquitt and Kate P Zipay. 2015. Justice, fairness, and employee reactions. *Annu. Rev. Organ. Psychol. Organ. Behav.* 2, 1 (2015), 75–99.

[12] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. 2018. Bayesian Inference of Self-intention Attributed by Observer. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (Southampton, United Kingdom) *(HAI '18)*. Association for Computing Machinery, New York, NY, USA, 3–10. doi:10.1145/3284432.3284438

[13] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. 2022. Conveying Intention by Motions With Awareness of Information Asymmetry. *Frontiers in Robotics and AI* Volume 9 - 2022 (2022). doi:10.3389/frobt.2022.783863

[14] Lars Johannsmeier and Sami Haddadin. 2016. A hierarchical human-robot interaction-planning framework for task allocation in collaborative industrial assembly processes. *IEEE Robotics and Automation Letters* 2, 1 (2016), 41–48.

[15] 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.

[16] Xun Li, Peng Xian, and Juqin Zhu. 2011. Research on teamwork mechanism and teamwork efficiency from the perspective of fairness preference. In *2011 International Conference on Computer and Management (CAMAN)*. IEEE, 1–7.

[17] Fabian Ranz, Vera Hummel, and Wilfried Sihn. 2017. Capability-based task allocation in human-robot collaboration. *Procedia manufacturing* 9 (2017), 182–189.

[18] Alessandro Roncone, Olivier Mangin, and Brian Scassellati. 2017. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1014–1021.

[19] Jesus F Salgado, Chockalingam Viswesvaran, and Deniz S Ones. 2001. Predictors used for personnel selection: An overview of constructs, methods and techniques. *Handbook of industrial, work and organizational psychology* 1 (2001), 165–199.

[20] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[21] 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*. 29–36.

[22] Roy K Smollan. 2012. Chapter 6 Emotional Responses to the Injustice of Organizational Change: A Qualitative Study. In *Experiencing and managing emotions in the workplace*. Emerald Group Publishing Limited, 175–202.

[23] Kazunori Terada, Celso MM de Melo, Francisco C Santos, and Jonathan Gratch. 2025. A Bayesian Model of Mind Reading from Decisions and Emotions in Social Dilemmas. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 47.

[24] Alarith Uhde, Nadine Schlicker, Dieter P. Wallach, and Marc Hassenzahl. 2020. Fairness and Decision-Making in Collaborative Shift Scheduling Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[25] 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.

Received 2025-12-08; accepted 2026-01-12