Unfair! Perceptions of Fairness in Human-Robot Teams

Mai Lee Chang[1], Greg Trafton[2], J. Malcolm McCurry[3], Andrea Lockerd Thomaz[1]

***Abstract*— How team members are treated influences their performance in the team and their desire to be a part of the team in the future. Prior research in human-robot teamwork proposes fairness definitions for human-robot teaming that are based on the work completed by each team member. However, metrics that properly capture people's perception of fairness in human-robot teaming remains a research gap. We present work on assessing how well objective metrics capture people's perception of fairness. First, we extend prior fairness metrics based on team members' capabilities and workload to a bigger team. We also develop a new metric to quantify the amount of time that the robot spends working on the same task as each person. We conduct an online user study (n=95) and show that these metrics align with perceived fairness. Importantly, we discover that there are bleed-over effects in people's assessment of fairness. When asked to rate fairness based on the amount of time that the robot spends working with each person, participants used two factors (fairness based on the robot's time and teammates' capabilities). This bleed-over effect is stronger when people are asked to assess fairness based on capability. From these insights, we propose design guidelines for algorithms to enable robotic teammates to consider fairness in its decision-making to maintain positive team social dynamics and team task performance.**

I. INTRODUCTION

With the wide-spread use of artificial intelligence (AI) and machine learning (ML) in various applications around us, fairness has become an important focus for researchers ([1], [2], [3]). Both AI algorithms and robotic teammates make decisions that can impact groups of people. Inherent in human group dynamics is fairness ([4], [5], [6]). Prior research shows that fairness is the foundation of trust and team effectiveness [7]. In particular, fairness may significantly affect the team's efficiency [8]. For example, a teammate who feels he is being treated unfairly is more likely to perform poorly, not engage in the task, or treat other teammates badly.

It is only recently that human-robot interaction (HRI) researchers have started to explore the concept of fairness in human-robot teaming. Most of the previous human-robot teaming algorithms focus on solely minimizing objective metrics of task performance including the human's idle time [9], task completion time [10], and the number of actions to reach the goal state [11]. Claure et al. [12] defined fairness as a constraint on the minimum rate that each human teammate is selected to play and showed that poor performers trusted the robot more when given more chances to play. Chang et al. [13] investigated the effects of a robot's behavior that is characterized by fluency (i.e., coordination of actions) and effort (i.e., tasks that take the most time) on participants' perception of fairness. They showed that a robot's effortful behavior significantly increased participants' perception of fairness and proposed objective fairness metrics based on workload, teammate capabilities, and task type for a single-robot single-human team. Each of these three metrics assume that people will perceive a fair situation when different aspects are balanced (equal): workload is balanced across teammates; teammates' abilities are balanced across tasks; and all teammates work an equal amount on each task.

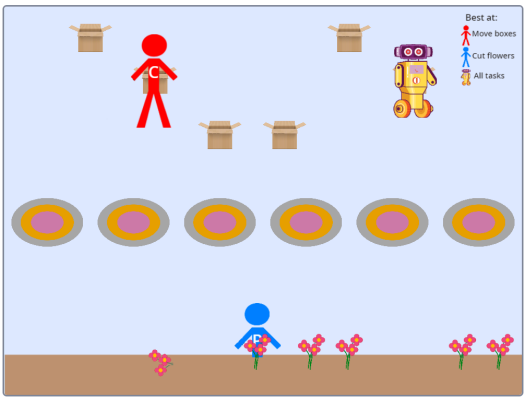


Fig. 1. Team scenario used in the user study.

However, a critical next step towards understanding the impacts of fairness on team performance is to first ensure that fairness metrics are capturing people's perception of fairness. Achieving this is challenging because fairness is preference-based [8] and context-dependent ([14], [15]). Working towards this challenge, our approach includes expanding prior fairness metrics for a single-robot single-human team [13] to a bigger team. This is based on the literature showing that people assess fairness by comparing outcomes with others who are in a similar situation ([14], [15]). In a bigger human-robot team, humans are more likely to compare themselves to other humans than to a robot [16]. In addition, within a larger team, the amount of time that the robot spends working on the same task as one person versus the other may influence people's perception of fairness, which led to the creation of a new fairness metric. We conduct a user study via Amazon Mechanical Turk to evaluate how well these fairness metrics based on workload, capability, and time capture people's perception of fairness (Figure 1) and also discuss implications for robotic teammate algorithm design.

This material is based upon work supported by the Naval Research Laboratory and Intel.

[1]Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78705, USA mlchang@utexas.edu, athomaz@ece.utexas.edu

[2] Naval Research Laboratory, Washington, DC, 20375, USA greg.trafton@nrl.navy.mil

[3] Peraton, Alexandria, VA 22314, USA jmccurry@peraton.com

II. RELATED WORK

Fairness is a hallmark of cooperative relationships [17]. Robotic teammates should consider fairness to achieve lasting partnerships. The social science literature shows that an individual's assessment of fairness depends on the comparison of their outcome to others in a similar situation ([14], [15]). In addition to caring about one's own payoffs, individuals also care about others' payoffs ([15], [17]).

When an algorithm interacts with or considers more than one person, such as ML algorithms that are used to inform the decisions made by humans, fairness comes into play ([1], [2], [3]). ML algorithms and robotic teammate algorithms are both making decisions about groups of people and fairness is an inherent, critical aspect of human group interactions ([4], [5], [6], [18]). However, research about fairness in human-robot teaming remains an open challenge. In the following subsections, we highlight existing work on fairness in machine learning (Section II-A) and human-robot teams (Section II-B) along with the role that perspective-taking plays in fairness (Section II-C).

A. Fairness in Machine Learning

In recent years, fairness in ML has gained prominence as ML algorithms are increasingly being used to inform critical decisions in the areas of financial lending, criminal justice, healthcare, and beyond ([1], [2], [3]). Researchers have proposed a rich set of fairness definitions that can be used to design and evaluate ML systems to ensure that biases in the data and model inaccuracies do not lead to models that discriminate against people based on sensitive attributes (e.g., race, gender, age) ([1], [2], [3], [19], [20], [21]).

These objective fairness definitions are developed by ML researchers and require some knowledge of ML so it is unclear if the general public agree and understand them. This is important because they are the ones who will be impacted by these systems. Saha et al. [22] conducted an online study to understand non-experts' comprehension and perception of three standard fairness notions (demographic parity, equal opportunity, and equalized odds). They found that people had the most difficulty understanding equal opportunity which involves understanding false negative rate and false positive rate. Another study investigated how users collaborate with risk assessment tools, focusing on a tool that aides judges' pretrail release decisions ([23], [24]). They found that participants could not effectively evaluate the risk assessment's performance and thus unable to appropriately rely on the risk assessment.

This rich body of prior work of bias in ML inspires our work. In our work, we quantify fairness based on each team member's contribution to the overall teamwork such as the extent to which they contribute their skills and time working with other team members. We seek to understand how well our metrics match people's perception of fairness.

B. Fairness in Human-Robot Teams

Many of the human-robot teamwork algorithms optimize for the team's task performance and ignore fairness ([9], [10], [11], [25]), perhaps assuming that optimal task performance will lead to a high level of perceived fairness (i.e., speed of finishing the task is "fair" to everyone). Jung et al. [26] created a collaborative tower construction task where the human teammates' roles are to build the tallest tower using blocks and the robot's role is to allocate the blocks. They studied two conditions (equal vs. unequal distribution) and found that participants rated the unequal condition to be significantly lower in terms of team relationship satisfaction. In a similar resource allocation study where the robot selects one of two human teammates to play the Tetris game, Claure et al. [27] designed a multi-arm bandit algorithm with fairness constraints.

Chang et al. [13] explored the influence of a robot's effortful and fluent behavior on people's perception of fairness. Their results revealed that a robot who displays effort significantly increased participants' fairness ratings. They also proposed three fairness metrics for a single-robot single-human team: Equality of Workload, Equality of Capability, and Equality of Task Type. In this paper, we expand Equality of Capability and Equality of Workload to a bigger team and propose a new metric based on the robot's time working with each human teammate.

C. Perspective-Taking

During interactions, people benefit from their ability to assume another person's point of view which is known as perspective-taking. Prior work in the social sciences show that social perspective-taking can help achieve fairness [28], [29]. Heck et al. [30] demonstrated that through affective perspective-taking, children were able to make fair resource allocations in both the first-party (participant had a stake in the outcomes) and third-party scenarios (participant did not have a stake in the outcomes).

In HRI, one of the earliest works of perspective-taking is from Trafton et al. [31]. They show that a robot that takes on the human teammate's perspective results in successful collaboration. Zhao et al. [32] showed that people's tendency to take on the robot's visual perspective is dependent on the robot's behavior, in particular object-directed gaze and goal-directed reaching. We leverage these findings about people's ability to also take on the robot's perspective and the role that perspective-taking plays in people's assessment of fairness in the design of our user study.

III. FAIRNESS DEFINITIONS

Chang et al. [13] propose three fairness definitions for a singe-robot single-human team: Equality of Capability (E_c), Equality of Task Type (E_t), and Equality of Workload (E_w) In general, these metrics compare how balanced the work completed by each team member is. In this paper, we are interested in expanding E_c and E_w to a single-robot two-human team. When we add another human to the team, we now compare the work done by one human to the other human. This is based on the literature showing that people assess fairness by comparing outcomes with others who are in a similar situation ([15], [17]). For most multi-human

robot teams, the humans will compare themselves to other human teammates because people's capabilities are more similar to other people than to a robot (e.g., people are better at manipulation and creativity but worse at mental math)[16]. Equality of Capability is based on the insight that people prefer to work on tasks they are good at rather than tasks they are poor at. Note that preference is an implicit aspect to E_c, but may be separated out in later work. Formally, E_c (Equation 1) is defined as equalizing the number of completed tasks that the human team members are most skilled at, i.e., strength. Note that we use the terms capability and strength interchangeably.

$$E_c = \frac{\#(C_{H1} \cap S_{H1})}{\#S_{H1}} - \frac{\#(C_{H2} \cap S_{H2})}{\#S_{H2}} \quad (1)$$

We denote by $\#(C_{Hi} \cap S_{Hi})$ the number of tasks completed by human i that is his strength and $\#S_{Hi}$ is the total number of possible tasks that are human i's strength. This means that only the humans can directly influence E_c. The robot can indirectly influence E_c by its actions because they impact the tasks that are available for the humans.

E_w (Equation 2) is defined as equalizing the number of tasks completed by each human.

$$E_w = \frac{\#C_{H1} - \#C_{H2}}{\#A} \quad (2)$$

$\#C_{Hi}$ is the number of tasks completed by human i and $\#A$ is the total number of possible tasks for the team divided by the total number of team members.

Another way that human teammates may make a comparison is by the amount of time that the robot works on the same type of task as them versus other human teammates. Thus, we propose a new fairness notion called Equality of Time, E_{time}. E_{time} is defined as equalizing the amount of time that the robot spends working on the same task as one of the human teammates versus the other human (Equation 3). E_{time} does not consider the time when all team members work on the same task.

$$E_{time} = \frac{t_{H1} - t_{H2}}{t_{H1} + t_{H2}} \quad (3)$$

t_{Hi} denotes the total amount of time that the robot spends working on the same task as human i. Note that for E_{time}, the robot directly influences this metric unlike in the E_c and E_p metrics.

These fairness metrics are in the range $[-1, 1]$ where -1 means unfairness towards $H2$, i.e.,$H2$ completed all the work. An equality value of 1 means unfairness towards $H1$ and 0 means the work was equally shared. These metrics fluctuate throughout the interaction and are task agnostic.

IV. USER STUDY

We conducted a user study with two goals: 1) Assess how well E_c and E_{time} capture people's perception of fairness and 2) Extract algorithm design guidelines that factor fairness. The study was between-subjects with $|E_c|$ and $|E_{time}|$ as factors and conducted via Amazon Mechanical Turk.

A. Task Description

Participants watched a video that was about 1 minute long showing animations of a team of one robot and two humans working together on three types of tasks related to cleaning and setting up tables in a restaurant (Figure 1). We used Scratch 3.0 [33] to create the animations. The three task types were: move empty boxes, vacuum rugs, and cut flowers (to be used as table decoration). There were a total of six boxes that need to be moved north (towards the top of the screen), six rugs that need to be vacuumed, and six bunches of flowers that need to be cut. When a rug is vacuumed, its colors change transparency to become lighter. The flower cutting task involved moving the flowers to the cutting location and then cutting them.

At the start of the video, there was a description of the team's tasks and each team member introduced themselves including which task they are best at. The human team members were Chris and Pat. We selected these names because they are gender-neutral. Chris' strength was moving boxes and Pat's strength was cutting flowers. At the start of the video, we displayed the following text, "Chris, Pat, and the robot are on the same team. They need to get all these jobs done: move boxes, cut flowers, and vacuum rugs. Each team member chooses what they want to work on." Next was Chris' introduction, "Hi, I am Chris! I am best at moving empty boxes." Then it was Pat's introduction, "Hi, I am Pat! I am best at cutting flowers." Finally, Poli introduced itself, "Hi, I am Poli! I can do all tasks equally well." Thus, each person had a separate and different strength, but the robot could perform all tasks equally well. After the introductions, we displayed each team member's strengths at the top right corner (Figure 1). We controlled for E_w by having each team member complete 6 tasks, assuming that each task had similar difficulty. Another control was setting each task to take the same amount of time, i.e., unit time.

This scenario emphasizes team members having different strengths, a common situation in real-world environments. The robot is not allocating tasks to the humans. Note that each teammate's action influences the tasks that are available for the other teammates. After watching the video, participants completed a questionnaire.

B. Independent Variables

The independent variables (IVs) are $|E_c|$ and $|E_{time}|$. We denote these IVs as Capability and Time. While all fairness metrics are likely to vary over time, we measured $|E_c|$ and $|E_{time}|$ at the end of the scenario, as close to our surveys as possible. The scenarios were constructed such that $|E_c|$ had two levels: $|E_c| = 0.0$ and $|E_c| = 0.5$. We denote these levels as Capability Balanced and Capability Unbalanced. With Capability Balanced, the robot selected the tasks that it works on in such a way that Chris and Pat can complete the same number of tasks that they are best at, i.e., strengths. With Capability Unbalanced, the robot moved more boxes which left Chris with less of the tasks that he's best at (boxes) so in the end, Chris contributed less of his strength to the team compared to Pat.

Capability Balanced & Time Balanced

Time	Chris	Pat	Robot	$\lvert E_c \rvert$	$\lvert E_{time} \rvert$	$\lvert E_w \rvert$
1				0.0	1.0	0.0
2				0.0	1.0	0.0
3				0.0	0.3	0.0
4				0.0	0.0	0.0
5				0.0	0.0	0.0
6				0.0	0.0	0.0

Capability Balanced & Time Unbalanced

Time	Chris	Pat	Robot	$\lvert E_c \rvert$	$\lvert E_{time} \rvert$	$\lvert E_w \rvert$
1				0.2	1.0	0.0
2				0.3	1.0	0.0
3				0.3	1.0	0.0
4				0.3	1.0	0.0
5				0.2	1.0	0.0
6				0.0	1.0	0.0

Capability Unbalanced & Time Balanced

Time	Chris	Pat	Robot	$\lvert E_c \rvert$	$\lvert E_{time} \rvert$	$\lvert E_w \rvert$
1				0.0	1.0	0.0
2				0.2	1.0	0.0
3				0.3	1.0	0.0
4				0.5	1.0	0.0
5				0.7	0.0	0.0
6				0.5	0.0	0.0

Capability Unbalanced & Time Unbalanced

Time	Chris	Pat	Robot	$\lvert E_c \rvert$	$\lvert E_{time} \rvert$	$\lvert E_w \rvert$
1				0.0	1.0	0.0
2				0.0	1.0	0.0
3				0.0	1.0	0.0
4				0.2	1.0	0.0
5				0.3	1.0	0.0
6				0.5	1.0	0.0

Fig. 2. A table represents a condition in the user study, showing the task that each teammate completed and equality value magnitudes at each time step. The final equality value magnitudes are highlighted in green. Chris' strength is moving boxes and Pat's strength is cutting flowers. The robot's strength is all the tasks.

We used a similar approach for the two constructed levels of $|E_{time}|$: $|E_{time}| = 0.0$ and $|E_{time}| = 1.0$ and are denoted as Time Balanced and Time Unbalanced. For Time Balanced, the robot spent the same amount of time working on the same task as Chris and Pat. For Time Unbalanced, the robot spent the entire time working on the same task as Chris, i.e., the robot did not spend any time working with Pat. This resulted in four conditions which are four team scenarios. Figure 2 details each condition at each time step, showing the task that each teammate completed and the equality value magnitude. Note that in the conditions with Capability Unbalanced and/or Time Unbalanced, unfairness is towards Pat. So, in the Capability Unbalanced and Time Unbalanced condition, Pat experienced two types of unfairness: 1) Pat completed more tasks that he is best at compared to Chris and 2) the robot spent more time working with Chris than Pat. We chose these values to provide large differences between conditions; future studies could examine intermediate or different values.

C. Hypotheses

Overall, our hypotheses are that balancing time and capability across human teammates will lead to higher perception of fairness, while an imbalance will lead to a perception of unfairness. Specifically, we predict that there will be separation effects and bleed-over effects.

Separation Effects: When queried about how fair the scenario was with respect to a single dimension of fairness (e.g., time), people will most likely be able to assess the specified fairness. For example, when a robot spends much more time working on the same task as one person than another person and we ask whether the robot was fair with respect to equality of time, we expect people to notice the difference and say that it was unfair.

Bleed-over Effects: When queried about how fair the scenario was with respect to a single dimension of fairness (e.g., time), people may also use multiple factors to assess the specified fairness. In particular, people may consider a situation where a person has been treated unfairly on multiple dimensions to have a cumulative effect, even when asked to judge unfairness on a single dimension. In other words, we are proposing that people will use multiple factors to gauge how unfair a scenario is **even when asked to focus on a specific aspect of fairness**. Consider the scenario where the robot spends much more time working on the same task as one person than another person (unbalanced time) and one person completes more tasks that he is best at than the other person (unbalanced capability). There are two dimensions of unfairness present. If we ask whether the robot was fair to both humans with respect to equality of time, we hypothesize that there will be a cumulative effect. That is, their assessment of unfairness based on time is influenced by unbalanced time and unbalanced capability. This would suggest that people are not well able to differentiate between multiple components of fairness.

D. Dependent Variables

Table I shows the subjective measures we used. For each of the fairness metrics, we asked participants about their level of agreement with the metric and their assessment of the teamwork based on that metric. The questionnaire also

included perspective-taking questions that ask participants to put themselves in Chris' and Pat's roles.

Equality of Workload:
1. Each team member completing the same amount of work is a fair way to contribute to the team. (Likert)
2. Poli was fair to both Chris and Pat with respect to the amount of work Chris and Pat were doing. (Likert)
3. Please elaborate on your answer to question #2 above.
Equality of Time:
1. The amount of time that Poli spends working on the same task as each human teammate should be the same. (Likert)
2. Poli was fair to both Chris and Pat with respect to the amount of time it spent working with them individually. (Likert)
3. Please elaborate on your answer to question #2 above.
Equality of Capability:
1. Working on tasks that you are best at is a fair way to contribute to the team. (Likert)
2. Poli was fair to both Chris and Pat with respect to them working on tasks that they are best at. (Likert)
3. Please elaborate on your answer to question #2 above.
Perspective-Taking
1. Pat was treated fairly by Poli. (Likert)
2. Chris was treated fairly by Poli. (Likert)
Overall Feedback
1. Please provide brief feedback. We are interested in your thoughts and also if you ran into any problems with any part of the experiment.

TABLE I
SUBJECTIVE MEASURES ADMINISTERED IN THE USER STUDY (7-POINT LIKERT ITEMS).

E. Participants

A total of 95 participants (30 females, 65 males, age: $Mean = 38.55$, $SD = 12.25$), participated in the study. Seventy-seven participants self-reported their race (Caucasian/White: 65, Black: 3, African American: 2, American: 1, Hispanic: 1, Hispanic/Latinx: 1, Latino: 1, Mexican: 1, Mixed: 2). The number of participants in each condition were: 24 in Capability Balanced and Time Balanced, 22 in Capability Balanced and Time Unbalanced, 25 in Capability Unbalanced and Time Balanced, and 24 in Capability Unbalanced and Time Unbalanced. The task took about 10 minutes and participants received $2 for compensation. This study was approved by the Institutional Review Board (IRB).

V. RESULTS

A statistical model based on the 2×2 between-subjects design with Capability (E_c) and Time (E_{time}) as factors was used in the analyses of variance (ANOVA). Tables II and III show a summary of the results. As a manipulation check of our control of E_w, we analyzed participants' agreement with E_w as a fairness metric. In general, they highly agreed that each team member completing the same amount of work is a fair way to contribute to the team as shown by the non significant results in Table II. For their ratings of E_w, our analysis also showed no significant results (Figure 3(a), Table II). That is, they perceived the robot to be fair when they factored in the amount of work that Chris and Pat completed. These results are expected since E_w was balanced (equivalent) across conditions.

A. Equality of Time

Participants reported high agreement with equality of time as a fairness metric, meaning that they agreed that the amount of time that Poli spends working on the same task as each human teammate should be the same (Table II). Our analysis showed significant main effects of Capability and Time on participants' judgment of E_{time} (Figure 3(b), Table II). As expected, they perceived the Time Balanced condition ($M = 5.59, SD = 1.43$) to be significantly fairer than the Time Unbalanced condition ($M = 4.17, SD = 2.29$) which supports our *separation effects* hypothesis. Additionally, their perception of E_{time} was also impacted by fairness based on capability. They rated the Capability Balanced condition ($M = 5.33, SD = 1.76$) to be significantly fairer than the Capability Unbalanced condition ($M = 4.51, SD = 2.17$) in terms of E_{time}. This suggests that people's assessment of fairness with respect to time is impacted by other factors such as capability in this case, providing support for our *bleed-over effects* hypothesis.

B. Equality of Capability

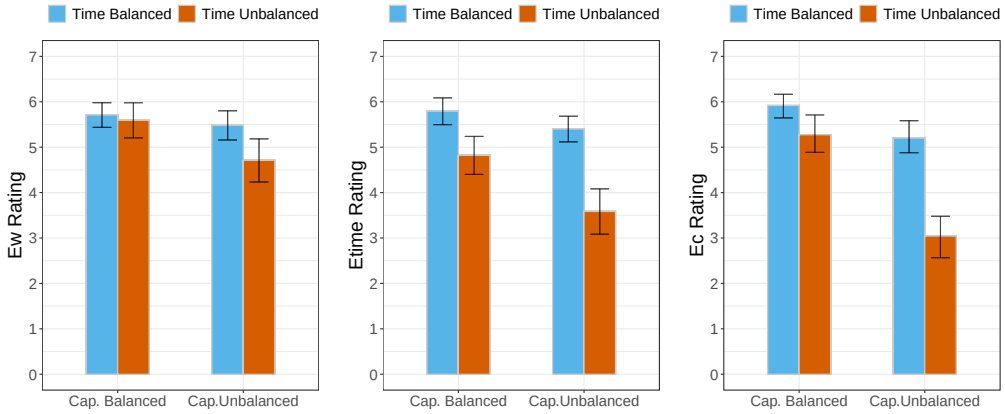
In general, participants highly agreed that equality of capability, working on tasks that you are best at, is a fair way to contribute to the team (Table II). Capability and Time significantly influenced participants' ratings of E_c as shown by the significant main effects and interaction (Figure 3(c), Table II). Participants felt that the Capability Balanced condition ($M = 5.61, SD = 1.68$) was significantly fairer than the Capability Unbalanced condition ($M = 4.14, SD = 2.28$), supporting our *separation effects* hypothesis.

In addition, the amount of time that the robot spent with each human also influenced participants' ratings of E_c. They reported higher E_c ratings when the robot spent an equal amount of time with each human ($M = 5.55, SD = 1.63$) in comparison to when the robot spent all its time with one human ($M = 4.11, SD = 2.36$). Also, the interaction is driven by both factors, Capability and Time. This suggests that when people were asked to rate fairness based on capability, their assessment is influenced by fairness based on both capability and time which supports our *bleed-over effects* hypothesis. That is, once one type of unfairness exists, additional types of unfairness increases people's perception of unfairness.

C. Perspective-Taking

When we asked participants to take Chris' perspective, they thought the robot treated them fairly as seen by the non significant results in Table III and Figure 4(a). However, when they were asked to take Pat's perspective, they felt that the robot treated them unfairly in terms of the amount of time it spent working with them (Table III, Figure 4(b)). Note that in the conditions with Capability Unbalanced and/or Time Unbalanced, unfairness is towards Pat.

They rated the Time Balanced condition ($M = 6.04, SD = 1.31$) to be significantly fairer than the Time Unbalanced condition ($M = 4.98, SD = 2.01$). These results show that people noticed a difference when the robot spent an equal



(a) Equality of Workload (b) Equality of Time (c) Equality of Capability

Fig. 3. Cap is the abbreviation of Capability. Ratings were on 7-point Likert items (higher ratings mean more fair).

	E_w Agreement		E_w Rating		E_{time} Agreement		E_{time} Rating		E_c Agreement		E_c Rating	
Predictor	$F(1,91)$	p	$F(1,91)$	p	$F(1,91)$	p	$F(1,91)$	p	$F(1,91)$	p	$F(1,91)$	p
Capability	0.20	0.65	2.20	0.14	1.83	0.18	4.56	< 0.05	2.80	0.10	14.72	< 0.001
Time	1.82	0.18	1.50	0.22	1.57	0.21	13.60	< 0.001	0.48	0.49	13.92	< 0.001
Capability x Time	0.15	0.70	0.78	0.38	0.29	0.59	1.22	0.27	0.20	0.66	3.92	0.05

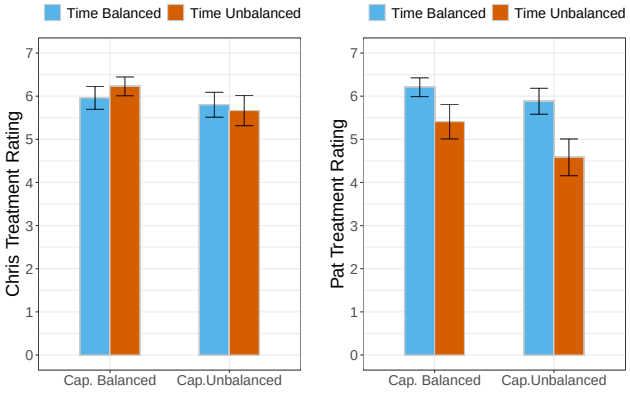
TABLE II

SUMMARY OF RESULTS FOR FAIRNESS METRICS.

	Chris Treatment		Pat Treatment	
Predictor	$F(1,91)$	p	$F(1,91)$	p
Capability	1.50	0.22	2.97	0.09
Time	0.05	0.83	9.50	< 0.01
Capability x Time	0.49	0.49	0.53	0.47

TABLE III

SUMMARY OF RESULTS FOR PERSPECTIVE-TAKING.



(a) Christ Treatment (b) Pat Treatment

Fig. 4. Perspective-taking results (higher ratings mean more fair).

amount of time working with Chris and Pat versus when it spent all its time working with only Chris. These results also support our *separation* hypothesis and show that our E_{time} metric does capture people's perception of fairness.

D. Qualitative Results

Participants' comments supported our findings of the bleed-over effects. Participants also factored in capability even though they were asked to rate fairness based on time. An example of this bleed-over effect is seen in one of the participant's elaboration on their E_{time} rating after watching the Capability Balanced and Time Balanced scenario: *"Because Poli (robot) is equally good at all the tasks, it was able to spend the same amount of time on the box moving as compared to Chris because Chris is only good at moving boxes. Since Pat is only good at cutting flowers, Poli would spend the same amount of time with Pat too because it is good at all tasks equally."*

The bleed-over effect is stronger when participants assessed the teamwork based on capability. There seems to be a cumulative effect when people assess fairness based on capability in the presence of another source of unfairness. For example, a participant who watched the Capability Unbalanced and Time Unbalanced scenario explained their E_c rating as follows, *"Pat worked solely on cutting flowers and that was what he was good at and he did not receive any help so he spent more time on that task than Chris who was good at moving boxes but he got help from Poli."*

When participants imagined themselves in Pat's and Chris' shoes, they felt that the robot treated Pat unfairly compared to Chris. A participant in the Capability Balanced and Time Unbalanced condition commented, *"I would not want to feel left out. I'm sure if this was a real life scenario, Pat would feel left out because Poli only worked closely with Chris."*

Moreover, from participants' overall feedback, they noticed the effect of the robot working on the tasks that are Chris' and Pat's strengths. For instance, a participant in the Capability Unbalanced and Time Unbalanced condition noted, *"If anything, I think Poli might have been unfair to the person he was helping. The more I think about it, the more I realize that by helping him with the boxes, Poli robbed Chris*

of his chance to do only what he was good at and forced him to help with a task he was less skilled in." Another participant in the same condition's remark was, *"There were also factors I didn't know if I should consider like resentment for someone else receiving help while you have to do your task alone. It's a hard thing to judge since every person would look at the situation differently."*

VI. IMPLICATIONS FOR ALGORITHM DESIGN

In this section, we present algorithm design guidelines to enable robotic teammates to consider fairness in its decision-making. Previous robotic teammate algorithms mostly optimize for the team's task performance and ignore fairness ([9], [10], [11]). Our fairness metrics aim to balance workload, teammate abilities, and the amount of time that the robot spends working on the same task as each person. Note that only the humans can directly influence E_w and E_c and the robot can indirectly influence these two metrics by its actions because they impact the tasks that are available to the humans. On the other hand, the E_{time} metric is directly influenced by the robot's actions. These equality values fluctuate throughout the interaction.

In designing robotic teammate algorithms, we recommend for the overall goal to focus on achieving equality value magnitudes as close to 0 as possible throughout the interaction. This recommendation is based on our results showing that for the E_c metric, people perceive equality value magnitudes equal to 0 to be fair and 0.5 to be unfair. Similarly, people perceive E_{time} value magnitudes equal to 0 to be fair and 1.0 to be unfair. Achieving these equality value magnitudes is challenging because each teammate's action impacts other teammates' behavior and thus how the interaction unfolds. For example, existing robotic teammate algorithms such as ([9], [10], [11]) can add fairness consideration by calculating the appropriate fairness metrics at each time step and estimating the future equality value magnitudes. This means that the algorithm will need to project future interaction trajectories based on the interaction thus far. The robot would want to take actions that are within the bounds of the trajectory that would arrive at the most desirable final equality value magnitudes.

Critically, the bleed-over effects suggest that algorithms should aim to be robust to other potential sources of unfairness. For instance, an algorithm may be designed to consider fairness in terms of E_c and E_{time} only but during the interaction, other sources of unfairness may occur. Consider the team scenario that we used in our study. In the situation where the robot works with different human teammates who have different capabilities, other potential sources of unfairness could be due to an unequal amount of human idle time, task difficulty, and task preference. Depending on the robot's role, team composition, and team goal, optimizing for one fairness metric could result in unintended unfair perceptions, thus it is important that the designers select the appropriate metrics. We observed this effect from the qualitative data about E_c when the robot performs the tasks of a team mate's strength to achieve fairness in terms of E_c. This could be perceived by that teammate as unfairness towards him. One potential solution is to add weights of importance for each fairness metric. In general, the algorithm would need a way to detect these other sources of unfairness and take actions to mitigate their effects. Finding and modeling sources of unfairness that are unknown *a priori* is a difficult but important future research area.

VII. DISCUSSION AND CONCLUSION

We extend Equality of Capability and Equality of Workload to a team of one robot and two humans and also propose a new metric called Equality of Time that quantifies the amount of time that the robot spends working on the same task as one human versus the other. We conducted an online user study to investigate how well E_c and E_{time} capture people's perception of fairness while controlling for E_w. We show that in general, people agree that fairness in teamwork can be assessed based on the human teammates' contribution of their strengths, the amount of time that the robot spends working with each human teammate, and workload. Most importantly, we show that our fairness metrics do capture people's perception of fairness in human-robot teaming.

When people judge fairness in the teamwork based on the amount of time that the robot spends working on the same task as each human teammate, they used two factors in our study. The first factor is the robot's time working with each teammate where they perceive balanced time to be significantly fairer than when the robot only worked solely with the same teammate which supports our *separation effects* hypothesis. The second factor is the amount of each human teammate's contribution based on their skill levels. When there was a balanced contribution of strengths, they rated fairness based on the robot's time working with each teammate to be significantly fairer than when one human completed all their strength tasks and the other human only completed half of their strength tasks. That is, they also factor in Equality of Capability even though they were asked to rate fairness based on Equality of Time, providing support for our *bleed-over effects* hypothesis.

Interestingly, the bleed-over effect increases when participants assess how fair the robot was to both humans with respect to them working on tasks that they are best at. They thought the robot was fairer when E_c was balanced, which supports our *separation effects* hypothesis. In addition, they thought the robot was fairer when E_{time} was balanced. However, when unfairness in terms of both capability and time is present, people perceived even greater unfairness, supporting our *bleed-over effects* hypothesis. This suggests that there is a cumulative effect when people assess fairness based on capability in the presence of other sources of unfairness. We expect that this result will generalize to other sources of (un)fairness because people are less able to isolate individual components of unfairness in a situation where multiple components of unfairness exist.

Even though this study examined fairness from a third perspective, research shows that people's assessment of fairness is not dependent on being an observer vs. stakeholder due to

perspective-taking [30]. Since people also care about others' payoffs, people can still be fair even as an observer. When people were asked to take Chris' and Pat's perspectives, they noticed that the robot treated Pat unfairly because it spent less time working with Pat compared to working with Chris. This result is expected since the conditions with unbalanced time and/or unbalanced capability were unfair towards Pat and the magnitude of unfairness in terms of time is greater than capability. Participants also voiced about the potential negative impacts of unfairness on team social dynamics such as the lack of inclusiveness.

Based on these insights, we provide design guidelines for robotic teammate algorithms to achieve fairness. In particular, we recommend that algorithms aim to achieve equality value magnitudes as close to 0 as possible throughout the interaction. Another suitable guideline is for robotic teammate algorithms to be robust to other sources of unfairness that it is not explicitly considering since people's assessment of one specific type of fairness can be influenced by the presence of other types of unfairness.

Overall, we show that our fairness metrics of Equality of Capability and Equality of Time do capture people's perception of fairness. Our key finding is that there are bleed-over effects in people's assessment of fairness. We also propose robotic teammate algorithm design guidelines to achieve fairness in human-robot teaming. This work is a step towards enabling robotic teammates to integrate within human teams easier and maintain long-lasting partnerships with human teammates.

REFERENCES

[1] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[2] L. Oneto and S. Chiappa, "Fairness in machine learning," in *Recent Trends in Learning From Data*. Springer, 2020, pp. 155–196.

[3] B. Hutchinson and M. Mitchell, "50 years of test (un) fairness: Lessons for machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 49–58.

[4] J. Rawls, *A theory of justice*. Harvard university press, 2009.

[5] T. R. Tyler and R. M. Dawes, "Fairness in groups: Comparing the self-interest and social identity perspectives," *Psychological perspectives on justice: Theory and applications*, pp. 87–108, 1993.

[6] M. Deutsch, *Distributive justice: A social-psychological perspective*. Yale University Press New Haven, CT, 1985.

[7] J. A. Colquitt, C. P. Zapata-Phelan, and Q. M. Roberson, "Justice in teams: A review of fairness effects in collective contexts," in *Research in personnel and human resources management*. Emerald Group Publishing Limited, 2005, pp. 53–94.

[8] X. Li, P. Xian, and J. Zhu, "Research on teamwork mechanism and teamwork efficiency from the perspective of fairness preference," in *2011 International Conference on Computer and Management (CAMAN)*. IEEE, 2011, pp. 1–7.

[9] J. Shah, J. Wiken, B. Williams, and C. Breazeal, "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, 2011, pp. 29–36.

[10] A. Roncone, O. Mangin, and B. Scassellati, "Transparent role assignment and task allocation in human robot collaboration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1014–1021.

[11] L. Johannsmeier and S. Haddadin, "A hierarchical human-robot interaction-planning framework for task allocation in collaborative industrial assembly processes," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 41–48, 2017.

[12] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Reinforcement learning with fairness constraints for resource distribution in human-robot teams," *arXiv preprint arXiv:1907.00313*, 2019.

[13] 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, pp. 1251–1258.

[14] W. G. Runciman and B. Runciman, *Relative deprivation and social justice: A study of attitudes to social inequality in twentieth-century England*. University of California Press Berkeley, 1966, vol. 13.

[15] C. Bicchieri, "Local fairness," *Philosophy and Phenomenological Research*, vol. 59, no. 1, pp. 229–236, 1999.

[16] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction Studies*, vol. 8, no. 3, pp. 483–500, 2007.

[17] K. McAuliffe, P. R. Blake, N. Steinbeis, and F. Warneken, "The developmental foundations of human fairness," *Nature Human Behaviour*, vol. 1, no. 2, pp. 1–9, 2017.

[18] J. Rawls, *Justice as fairness: A restatement*. Harvard University Press, 2001.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[20] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

[21] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in neural information processing systems*, 2017, pp. 4066–4076.

[22] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, and M. Tschantz, "Measuring non-expert comprehension of machine learning fairness metrics," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8377–8387.

[23] B. Green and Y. Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 90–99.

[24] ——, "The principles and limits of algorithm-in-the-loop decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[25] S. Nikolaidis and J. Shah, "Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy," in *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 2013, pp. 33–40.

[26] 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.

[27] 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.

[28] D. Brugman, C. Out, and J. C. Gibbs, "Fairness and trust in developmental psychology," in *Women and Children as Victims and Offenders: Background, Prevention, Reintegration*. Springer, 2016, pp. 265–289.

[29] A. D. Galinsky and G. B. Moskowitz, "Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism." *Journal of personality and social psychology*, vol. 78, no. 4, p. 708, 2000.

[30] I. A. Heck, N. Chernyak, and D. M. Sobel, "Preschoolers' compliance with others' violations of fairness norms: The roles of intentionality and affective perspective taking," *Journal of Cognition and Development*, vol. 19, no. 5, pp. 568–592, 2018.

[31] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, 2005.

[32] X. Zhao, C. Cusimano, and B. F. Malle, "Do people spontaneously take a robot's visual perspective?" in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 335–342.

[33] Scratch.mit.edu, "Scratch 3.0," 2020. [Online]. Available: https://scratch.mit.edu/