Evaluating Fluency in Human–Robot Collaboration

Guy Hoffman, *Member, IEEE*

***Abstract*—Collaborative fluency is the coordinated meshing of joint activities between members of a well-synchronized team. In recent years, researchers in human–robot collaboration have been developing robots to work alongside humans aiming not only at task efficiency, but also at human–robot fluency. As part of this effort, we have developed a number of metrics to evaluate the level of fluency in human–robot shared-location teamwork. While these metrics are being used in existing research, there has been no systematic discussion on how to measure fluency and how the commonly used metrics perform and compare. In this paper, we codify subjective and objective human–robot fluency metrics, provide an analytical model for four objective metrics, and assess their dynamics in a turn-taking framework. We also report on a user study linking objective and subjective fluency metrics and survey recent use of these metrics in the literature.**

***Index Terms*—Artificial intelligence, computational and artificial intelligence, cooperative systems, human-robot interaction, intelligent robots, intelligent systems, man-machine systems, systems, man, cybernetics, user interfaces.**

I. Introduction

WHEN humans collaborate on a shared activity, and especially when they are accustomed to the task and to each other, they can reach a high level of coordination, resulting in a well-synchronized meshing of their actions. Their timing is precise and efficient, they alter their plans and actions appropriately and dynamically, and this behavior emerges often without exchanging much verbal information.

We denote this quality of interaction the *fluency* of the shared activity. With the aim of using robots in the workforce, we are interested in how robotic teammates could similarly perform more fluently with their human counterparts. This paper provides tools to evaluate the level of fluency in a human–robot shared activity.

Fluency in human–robot collaboration has garnered interest over the past several years, as a large portion of the human–robot interaction (HRI) literature is working toward computational models of collaboration (e.g., [1]–[3]; for a survey, see [4]). Nevertheless, the majority of human–robot collaborative systems are structured in a stop-and-go fashion, following command and response patterns, and holding little of the fluent quality that is part of a satisfying collaboration.

A fluent teammate evokes appreciation and confidence. If robotic teammates are to be widely integrated in a variety of workplaces to collaborate with nonexpert humans, their acceptance may depend on the fluent coordination of their actions with that of their human counterparts.

The goal of this paper is to set the stage for a commonly accepted toolkit to evaluate fluency in human–robot collaboration. This could help benchmark advances in human–robot collaboration and lead to better-designed robotic teammates.

A. Fluency Metrics in Human–Robot Collaboration Research

Collaborative fluency metrics were introduced by the author and collaborators in the context of an anticipatory controller for shared-workspace Markov decision processes [5]. We noted that even in cases where task efficiency was not improved, people's sense of fluency was increased when the robot anticipated their actions. This finding suggested that fluency is a separate construct that does not simply track efficiency. Further investigation revealed differences in a number of objective task metrics that could explain this discrepancy.

In subsequent years, other researchers have used the proposed fluency metrics to evaluate aspects of human–robot collaboration. For example, Cakmak *et al.* have measured fluency of handovers from a robot to a human [6], Chao and Thomaz measured fluency to evaluate a multimodal turn-taking system based on timed Petri Nets [7], and Nikolaidis and Shah used fluency as a criterion to evaluate a system of human–robot cross training [8].

A survey of these works, including a list of metrics used thus far, was presented at the 2013 Robotics: Science and Systems Workshop on Human–Robot Collaboration [9]. However, that paper merely included a list of metrics, without further analysis, comparison, or in-depth discussion. In the past five years, interest in human–robot fluency has grown further, as evidenced by the recurrent use of some of the metrics listed in [9], detailed in the literature review in Section VII.

Given this growing interest, it makes sense to consider which of these metrics are most useful and to work toward agreed benchmarks for human–robot collaborative fluency. This paper is the first attempt to systematically present, map, and validate the measurement of fluency in human–robot collaboration. Its goals are threefold: to provide an archival inventory of subjective and objective fluency metrics, to suggest a first theoretical analysis of objective fluency metrics, and to systematically investigate the relationship between subjective and objective fluency metrics.

Following standard psychometric practice, we distinguish between two types of fluency metrics for human–robot collaboration: *subjective* metrics, measuring people's perception of the fluency of an interaction and related qualities of the robot, and *objective* metrics, which quantitatively estimate the degree of fluency in a given interaction.

Manuscript received June 20, 2018; revised December 7, 2018; accepted February 24, 2019. Date of publication April 1, 2019; date of current version May 15, 2019. This paper was recommended by Associate Editor R. Chavarriaga.

The author is with the Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: hoffman@cornell.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2904558

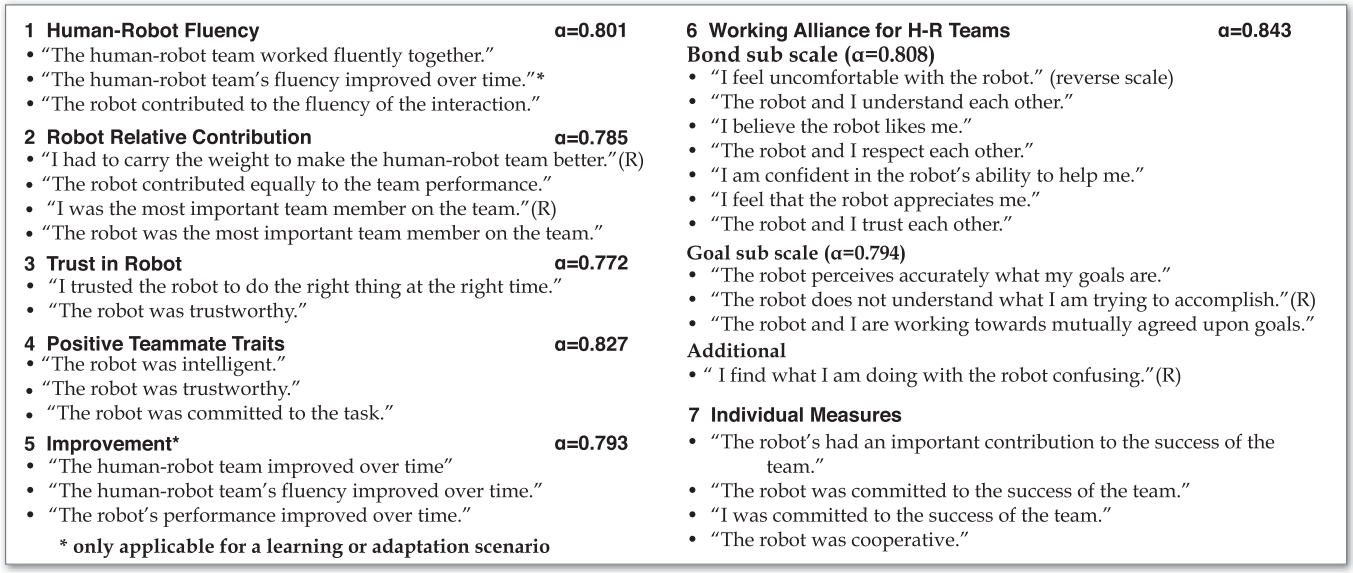

Fig. 1. Subjective fluency metric scales and items used in our studies. Cronbach's α is reported as measured in [10]. (R) indicates reverse scale.

II. Subjective Fluency Metrics

Subjective metrics include both direct measures of fluency that people attach to a collaboration and downstream outcomes of the perceived fluency, such as the trust human collaborators put in the robot, the perceived contribution of the robot, its positive teammate traits, or the human's sense that the robot is committed to the team.

In a number of human-subject studies, we and others have used questionnaires to rate agreement with fluency notions, including both single statements and composites of indicators related to the same measure. The list of questions used in our research is detailed in [9] and brought here in summary format alone in Fig. 1. We report on Cronbach's α for internal validity for each composite scale, as it was measured in [10]. Researchers have used additional subjective metrics in the past years alongside the measures presented here. We list those in detail in the survey in Section VII.

The scales in Fig. 1 include a three-item scale evaluating fluency directly (1) and six possible downstream outcomes of collaborative fluency (2–7). Scale 6 is an adaptation of an existing instrument, the "Working Alliance Inventory (WAI)" [11], adapted to human–robot teamwork.

III. Objective Fluency Metrics

While subjective metrics measure the sense of fluency perceived by a human in a human–robot team, it is also useful to codify objective measures. If we could reliably tie these metrics to the perceived fluency, they could serve as common benchmarks for evaluation. Such numerical metrics could also be used by machine learning or other optimization algorithms as part of their cost and reward functions.

This paper discusses four objective metrics that have been used in the context of human–robot fluency. Three of them were originally proposed in [5]: the percentage of concurrent activity (C-ACT), the human's idle time (H-IDLE), and the robot's functional delay (F-DEL). Later work added a fourth metric: the robot's idle time (R-IDLE).

These metrics were designed with some generality in mind and are agnostic to the specific content of the collaborative acts, relating only to periods of activity. Both the human and the robot are modeled as either active or inactive.

That said, the metrics were developed in a specific collaborative context, in which a human and a robot bring objects to a shared workspace and operate on them. These metrics were also evaluated later in the context of a repetitive handover task. While these are common human–robot collaboration scenarios, their formulation makes assumptions about the task that may not hold for other kinds of collaborative activities. For example, we make a simplifying assumption that the start and end time of the whole task is identical for both agents. Another assumption is that the team is made up of one human and one robot. The discussed metrics also do not cover certain aspects of cooperative motion and physically coupled collaboration. Other metrics may thus be more appropriate for different kinds of collaborative scenarios. These metrics include total task time, smoothness of trajectory features, or other temporal measures such as recurrence quantification metrics.

In the following section, we provide a description and motivation for the four metrics described above, along with an initial discussion of considerations when using these metrics.

A. Human Idle Time

This measure corresponds to the percentage of the total task time that the human was not active. Generally, humans have faster perceptual processing and more dexterous and faster manipulation and locomotion capabilities than the robot. Therefore, the human is often waiting for the robot to complete an action in order for them to perform the next step of the collaboration.

We postulate that H-IDLE relates to the subjective sense of fluency because it can be perceived as boredom, time wasted, or an imbalance between team members.

B. Robot Idle Time

Symmetrically, this measure corresponds to the percentage of the total task time that the robot was not perceivably active. R-IDLE can occur when the robot waits for input from the human, is processing input, is computing a decision, is waiting for additional sensory evidence, or is waiting for the human to complete an action.

Note that this measure could be interpreted in two ways, as the robot's apparent inactivity can be due to one of two cases: The robot could appear inactive to the human, but be internally active, for example, while processing data; conversely, the robot could be truly inactive, for example, when it is waiting for the human to complete an action. In the first case, the R-IDLE could increase at the same time that the H-IDLE increases and would be "dead time" in the collaboration. In the latter case, the R-IDLE increases while the human is active and might not be noticed by the human. Authors could interpret the R-IDLE to include or not include the robot's internally active, but seemingly inactive, time. In addition, the robot could mitigate its perceived idle time during processing by communicating its internal processing, as suggested in a number of HRI projects [12], [13].

We hypothesize that R-IDLE relates to the subjective sense of fluency because it can be perceived as a problem in the team coordination, be viewed as an inefficient use of the robot's resources, reflect poorly on the robot as a teammate, and indicate an imbalance between team members.

C. Concurrent Activity

While the previous two measures were computed on an individual agent level, this measure relates to both agents simultaneously and corresponds to the percentage of time out of the total task time, during which both agents have been active concurrently. In other words, C-ACT corresponds to the rate of action overlap between the agents.

Similar to the R-IDLE measure, C-ACT is also tied to the perceived activity of the robot and can be interpreted as including or excluding times, during which the robot is internally active, but seemingly inactive. It is also nontrivially related to the human's attention to the robot's level of activity. The human can dedicate more attention to the robot's activity while they are inactive, but conversely might be more sensitive to a robot being inactive while the human is active and "doing their part" of the collaboration.

We hypothesize that a high level of C-ACT is related to a subjective sense of fluency, as it could be seen as an indication for the team being well synchronized, the team members being similar to each other, and the work balance being fair.

D. Functional Delay

The fourth measure relates to the delay experienced by the agents immediately after completing an activity, as incurred by their teammate. F-DEL is defined as the accumulated time, as a ratio of the total task time, between the completion of one agent's action and the beginning of the other agent's action. F-DEL can be calculated for both agents together or for each agent separately. Realistically, human F-DEL is often negligible, making the total F-DEL roughly equal to that imposed by the robot.

F-DEL can be negative when actions are overlapping. Counter to intuition, the F-DEL ratio can also be larger than 1, meaning that the accumulated delay is longer than the total task time. This occurs if an agent completes a number of actions without response, each starting a "timer" on the F-DEL. F-DELs can, thus, be overlapping and accumulative.

We hypothesize that a low level of F-DEL is related to the subjective perception of human–robot fluency, as it indicates an efficient use of team members' time and a sense that the interface points between their activities are smooth and precise. As the human has just completed an action, we also postulate that there is heightened saliency to the robot's F-DEL. The importance of F-DEL is highly dependent on the nature of the collaboration. In turn-taking scenarios where an agent needs to wait for the other agents' action, this metric may be more salient than in scenarios where both agents mostly operate simultaneously.

E. Interaction Between Objective Metrics

The four metrics described above are interrelated, as they are all a function of the amount and timing of each agent's action. However, depending on the task, their dynamics along a collaborative period can be intricate, with one measure improving while another regresses. To illustrate this effect, Fig. 2 presents the interplay between the various measures in four common HRI scenarios: strict turn-taking with no processing delay, strict turn-taking with robot processing delay, robot processing delay with human anticipatory action, and fixed robot cycles with erratic human behavior.

While these examples are anecdotal, they illustrate that the four metrics are not trivially interchangeable. Between examples, H-IDLE and R-IDLE are similar, while CONC and F-DEL change independently. In real human–robot collaborations with more than three action cycles and varying timing parameters, we can expect more complex interrelations between the metrics, as discussed in the rest of this paper.

IV. Minimal Turn-Taking Model

To further explore the relationship between these objective metrics, we can look at a minimal model that captures some of the dynamics of a turn-taking human–robot collaboration.

In this model, we define a task instance as a tuple $< H, R >$ over an arbitrary time unit, with

$$H \equiv \{h_i\}_{i=1}^{n} \equiv \{(s_{hi}, d_{hi})\}_{i=1}^{n} \tag{1}$$

being the sequence of n human activity periods. Period i is denoted as h_i, which is a pair starting at s_{hi} and lasting d_{hi}. Similarly, we can denote the robot's actions

$$R \equiv \{r_i\}_{i=1}^{m} \equiv \{(s_{ri}, d_{ri})\}_{i=1}^{m}. \tag{2}$$

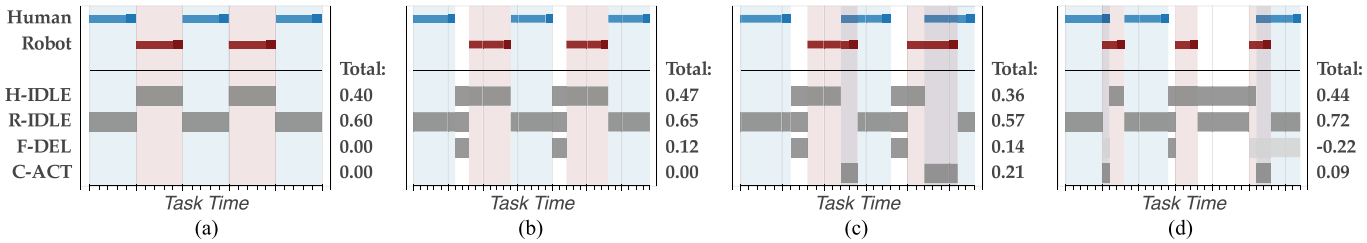

Fig. 2. Objective fluency metrics in four prototypical scenarios. The top sections indicate the agents' actions. From left to right: (a) Strict turn-taking, initiated by the human; each action is immediately followed by the next action of the other teammate. (b) Strict turn-taking, initiated by the human, but with some fixed processing delay on the robot's part, following fully completed human actions. (c) Fixed robot processing delay in a situation in which the human can start their action, while the robot is still working on its last action. (d) Fixed robot cycle with erratic human activity.

Without loss of generalization, the human starts the first activity at time $t = 0$ without any prior robot activity. The total task time T follows from the above definitions as

$$T = \max(s_{hn} + d_{hn}, s_{rm} + d_{rm}). \tag{3}$$

In Example (a) in Fig. 2, this would correspond to

$$H = \{(0, 5), (10, 5), (20, 5)\}, \quad n = 3 \tag{4}$$

$$R = \{(5, 5), (15, 5)\}, \qquad m = 2. \tag{5}$$

Given this model, we can analytically derive the following metric values. These can serve as an estimate of an idealized collaboration for benchmarking or prediction purposes

$$\texttt{H} - \texttt{IDLE} = 1 - \frac{1}{T}\sum_{i=1}^{n} d_{hi} \tag{6}$$

$$\texttt{R} - \texttt{IDLE} = 1 - \frac{1}{T}\sum_{i=1}^{m} d_{ri} \tag{7}$$

$$\begin{aligned}\texttt{C} - \texttt{ACT} = \frac{1}{T}\Big[&\max(0, s_{h1} + d_{h1} - s_{r1}) \\ &+ \sum_{i=2}^{n}\Big(\max(0, s_{ri-1} + d_{ri-1} - s_{hi}) \\ &+ \max(0, s_{hi} + d_{hi} - s_{ri})\Big)\Big]\end{aligned} \tag{8}$$

$$\texttt{F} - \texttt{DEL} = \frac{1}{T}\sum_{i=1}^{n}(s_{ri} - s_{hi} - d_{hi}). \tag{9}$$

Note that in the term for F-DEL, we make a simplifying assumption that actions, while being able to accumulate, do not accumulate more than once. In other words, we assume that an agent will start at most one new action before the other agent starts its next action. Given this assumption, we can also simplify (up to a constant) that $m = n$.

V. Simulations Using the Minimal Model

We characterize the temporal dynamics of the proposed objective fluency metrics by simulating the minimal turn-taking model.

The simulations are produced by generating a sequence of interleaved human and robot activity segments given the length of each agent's action and the length of the delay between actions.

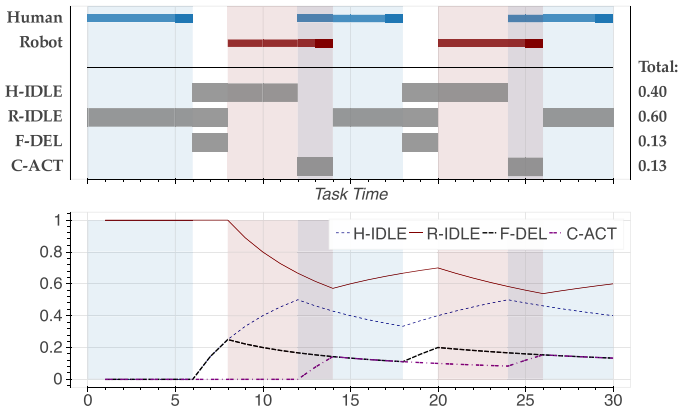

Fig. 3. Rates of objective fluency metrics oscillate throughout the task period, if n and m are low. In this example, $H = \{(0, 6), (12, 6), (24, 6)\}, R = \{(8, 6), (20, 6)\}$.

The delay can be specified as negative to generate overlapping actions. The simulator is given either fixed values or a mean and standard deviation for each parameter. In the latter case, the simulator picks a value based on a normal distribution, randomized per activity segment. Given a set of generated human and robot activity segments, the simulator then calculates and graphs a running value for each of the four metrics.

A. Instantaneous Dynamics

The examples in Fig. 2 and the derivations in Section IV present only summary statistics for each metric. This is how these measures have been used thus far in the HRI literature. However, examining the instantaneous dynamics for each metrics shows that there is additional temporal information that is useful to consider.

Fig. 3 shows the instantaneous dynamics of each metric and how they are related to periods of human and robot activity, overlap, and delay—as indicated by the shaded strips in the graph. Rates of H-IDLE and R-IDLE oscillate directly with the agent's activity. Rates of F-DEL and C-ACT oscillate similarly with periods of overlap and waiting periods.

Notably, the rates of all the objective metrics vary rapidly when dealing with low numbers of human and robot actions. If calculated in relation to the total task time, as is common in current use, there is thus a risk of misestimating the actual value of each metric. For low numbers of n and m, it is preferred

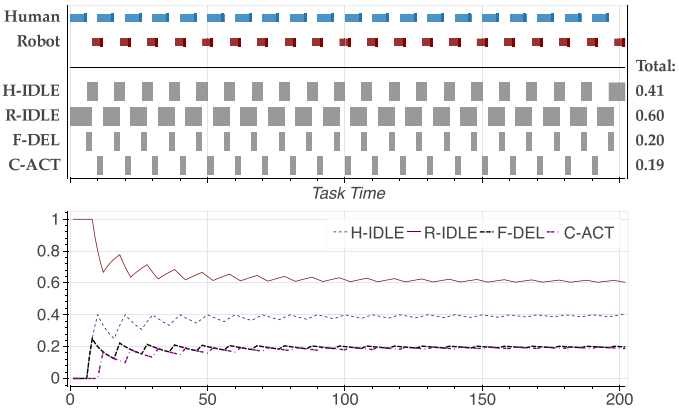

Fig. 4. Consistent human–robot team settles on mean values for objective metrics. Here, human actions take six time units, robot actions take four units, the robot's F-DEL is two units, and the human has an anticipation of two units.

to evaluate the metrics as a fraction of a single, or average, agent turn. However, due to the possible variance in the length of these turns, this approach is not always feasible. At the least, researchers should be aware of this issue and take measures to counteract it case by case.

B. Settling Period

Fig. 4 shows that the oscillations eventually settle on the average value as a fraction of task time. Thus, per HRI scenario, it makes sense to estimate the "ringing" or "settling" period for the specific collaborative dynamic and make sure to measure metrics after their values are likely to have settled.

C. Effects of Agent Variance

That said, there is a caveat with respect to settling and the subsequent extrapolation of metrics from the mean values at an early stage of the collaboration. In particular, the simulations in the previous example model both the human and the robot as invariant over time. In this case, the objective fluency metrics can be closely modeled by the analytical derivations given in Section IV, by estimating the mean over an initial period, or after the metrics have converged.

In contrast, adding variance to either or both of the agents notably undermines these assumptions. Fig. 5 shows results from simulations of a human turn-taking agent with the same parameters as in the previous example. The robot's duration, however, is not fixed but drawn from a normal distribution $d_{ri} \sim N(4.0, 0.25)$. We ran the same simulation twice.

While the metrics in Fig. 5(a) converge on what can be considered the mean value for each, Fig. 5(b) shows that metrics that appear to have settled undergo perturbations as late as three times the duration after which the initial turn-by-turn "ringing" has decayed.

The dynamics are even more pronounced if both agents are subject to variance in their turn-by-turn timing. Fig. 6 shows 50-turn simulations in which both turn durations are drawn from $d_{ri}, d_{hi} \sim N(6.0, 1.0)$. The strip graphs show that metrics accumulation is highly time dependent, and that in some cases, metrics converge to a fixed point (a), but in others (b), (c), changes occur late in the task execution.

D. Discussion

The objective metrics suggested herein have thus far been estimated exclusively as a single ratio value over the total task period. Simulations using a minimal model show that this use should be qualified: First, we see there is an initial "ringing" period, in which the metrics as part of total task time greatly oscillate before converging on their mean value. Researchers should evaluate the metrics with respect to a single turn instead of the total task time if the number of turns is low.

Second, even with relatively fixed human and robot behaviors, small variations can cause major dynamic shifts in metric values. These can occur late in the interaction. This emphasizes that fluency metrics need to be tracked and dynamically evaluated throughout the human–robot collaboration.

VI. Validating the Objective Metrics

As a final step, we set out to evaluate how objective fluency metrics relate to people's subjective sense of fluency. Both subjective and objective fluency metrics are increasingly used in human–robot collaboration research (see Section VII), with the assumption that the objective metrics capture something of people's sense of fluent collaboration. However, their relation has not yet been empirically grounded. To this end, we conducted a study relating the metrics discussed above.

We have developed a simple human–robot collaborative scenario simulator with a number of flexible timing parameters. The scenario is a joint workspace (see Fig. 7), in which the human and the robot transfer a number of objects from the right (human) end table of the workspace to the left (robot) end table. To do this, the human places objects on the shared (middle) table. This is a prototypical abstraction of shared-location human–robot collaboration, mapping to a factory stocking or home cleaning task. In our simulator, we can model a number of processing delays into each of the participants' action policies. Specifically, we can vary the time it takes each agent to pick up the objects, to detect the existence of an object in each workspace location, and to drop off the objects at the various workspaces.

A. Online Study

We collected data using Amazon's Mechanical Turk platform, which was found to be a reliable source for data gathering in similar contexts, given appropriate study controls [14]. A total of 143 people participated in the study. Of those, 104 participants were analyzed (age 21–68; M = 32.02 ± S = 10.43; 39 female), as determined by prerun criteria. These criteria included English proficiency, approval rate, filtering questions, and task time. Participants were paid $0.10, an amount on par with the going rate at the time for online surveys of approximately the length of the current study.

Participants were presented with two instruction pages and a form for informed consent. Then, each participant was presented with five short clips of the simulator. Each clip was approximately 40 s in length. The clips were counterbalanced and randomly assigned to the participants, chosen from a database of 50 clips, as described below. Each video was followed by an eight-item fluency scale. Then, participants filled out a

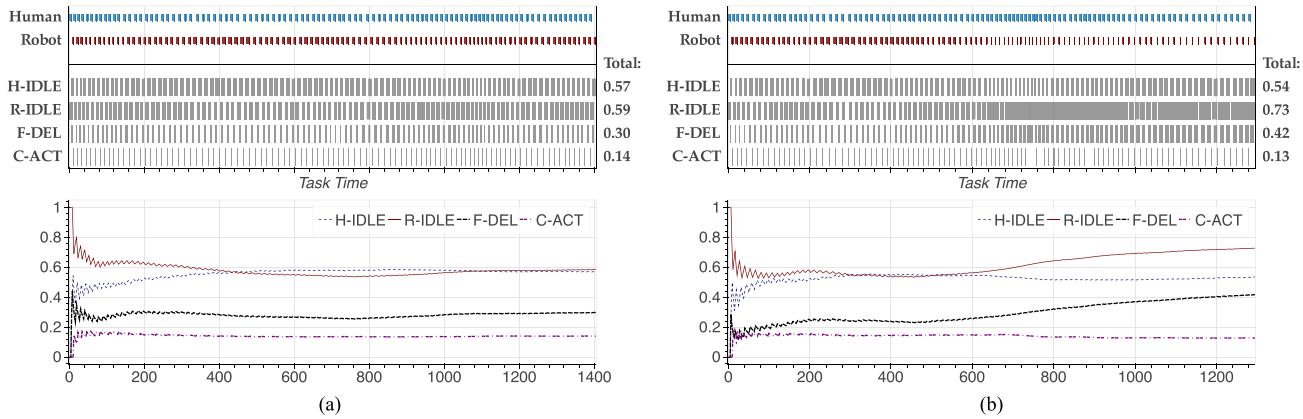

Fig. 5. Two simulation results using the minimal turn-taking model with robot-only variance. In some cases (a), the metrics converge after an initial settling period; in other cases (b), significant changes to metrics can occur late in the interaction.

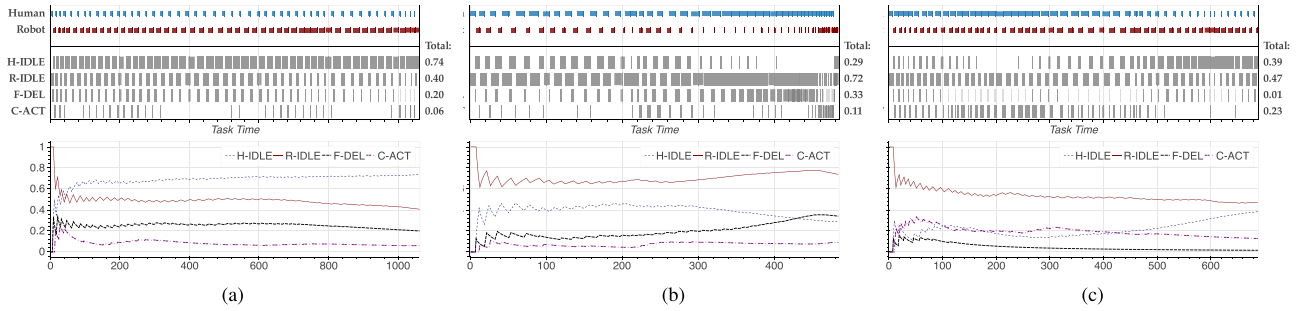

Fig. 6. (a)–(c) Simulation results using the minimal turn-taking model using the same basic human and robot parameters but with both human and robot variance, illustrating the dynamic nature of metrics over time.

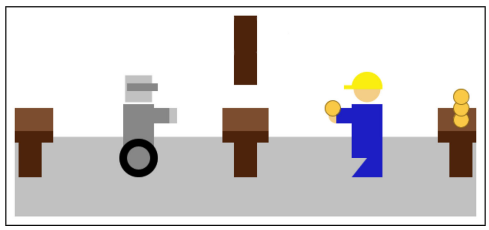

Fig. 7. Screenshot of the collaboration scenario used to evaluate the relation between objective and subjective fluency metrics.

demographic questionnaire, were debriefed, and were credited for their participation.

B. Scenario Generation

The study used 50 clips of various agent behaviors, covering a range of collaborative scenarios. To generate the clips, we randomized agent policy parameters, including the time it takes to detect and pick up the items, the time it takes to put them down, and additional processing delays. Initially, we generated 150 clips using random values for each of the above-mentioned agent parameters. Delays and processing times for both agents were uniformly selected between 0 and 4 seconds, with agents' travel time through their half of the workspace being approximately 3 seconds long. We then narrowed the 150 clips down to 50, striving to cover a broad representation of each of the four fluency metrics. We eliminated similar clips and collected those that would vary one fluency metric while leaving the others largely unchanged. To do so, we visually examined histograms of the clips along each metric and chose videos based on their location on each of the four graphs. To evaluate this procedure, we analyzed the histograms of the resulting 50 videos and confirmed that each metric was adequately represented, i.e., that there were low and high cases for each metric, and that the distribution for each metric was approximately normal, spherical, and spread across a similar number of clips for each metric.

We also conducted a qualitative manipulation check to verify that participants perceive the scenarios in the way we intended. Pilot study participants ($n = 12$) were presented with eight videos, in random order, each representing an extreme case of each objective metric. After each video, participants were interviewed along three open-ended questions: "How would you describe the interaction between the human and the robot?" "How would you describe the robot?" and "Would you describe the collaboration as fluent?" Subjective judgment of the responses suggested that participants perceived the extreme cases of our metrics as representing extremes of concepts related to collaborative fluency.

C. Dependent Measures

As this is an online study with a large participant population, we aimed to make the subjective questionnaire as concise as possible, especially given that each participant fills it out five times. We, therefore, used only eight of the above-mentioned indicators, selected as most closely related to a subjective sense of fluency. The indicators are listed in Table I, covering general fluency, robot contribution, commitment, and

TABLE I
CORRELATION BETWEEN OBJECTIVE METRICS AND INDICATORS OF SUBJECTIVE FLUENCY

Indicator	Robot Idle	Human Idle	Conc. Act.	Func. Delay
The human-robot team worked fluently together	-.074	**.098***	.042	**-.117****
The human was the most important member of the team	-.028	.050	.003	-.055
The robot was unintelligent (R)	-.088	-.006	**.089***	-.031
The robot was trustworthy	.004	.073	-.031	-.053
The robot was uncooperative (R)	-.015	.048	.023	-.032
The robot contributed to the fluency of the collaboration	-.043	**.108***	.027	**-.104***
The robot was committed to the success of the team	-.040	**.105***	.013	**-.096***
The robot had an important contribution to the success of the team	-.067	**.114***	.050	**-.120****
Composite Scale	-.074	**.119****	.047	**-.123****

*$p<0.05$, **$p<0.01$.

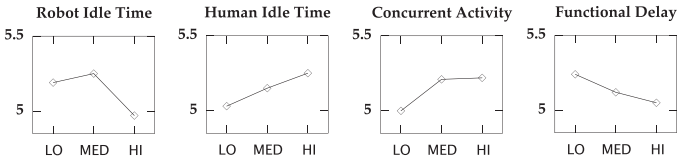

Fig. 8. Mean fluency rating per tercile of clips split by objective metric.

teammate traits. All scales were rated on a seven-point Likert scale from "*strongly disagree*" to "*strongly agree*." In a separate scale pretest ($n = 21$), we found high cross intercorrelations for this composite (Cronbach's $\alpha = 0.81$). In the study, Cronbach's α for this composite was 0.74.

D. Results

To investigate the relation between the objective and subjective metrics, we calculated Pearson correlations between each of the objective metrics and the eight-item composite, as well as for each indicator separately. The significance of correlation was evaluated using a two-tailed test with a $p < 0.05$ threshold.

R-IDLE was not significantly correlated with subjective fluency: $r(495) = -0.074$. H-IDLE was significantly correlated with subjective fluency: $r(495) = 0.119, p < 0.01$. C-ACT was not significantly correlated with subjective fluency: $r(495) = 0.047$. F-DEL was significantly reverse-correlated with subjective fluency: $r(495) = -0.123, p < 0.01$. Table I summarizes these findings and reports the correlation of each of the indicators with the four objective metrics.

Fig. 8 further illustrates the relationship between objective and subjective metrics. To explore the trends contributing to the discovered correlations, we split the clips into three equally sized populations—along each one of the objective metrics. The figure shows the mean rating on the composite fluency scale for each tercile.

E. Discussion

Our exploratory study shows significant correlation between two of the four objective metrics suggested and subjective observer fluency perception: H-IDLE and F-DEL (inversely correlated). R-IDLE is consistently inversely correlated with fluency perception, but not significantly so. C-ACT was not found correlated with the subjective metrics. Exploring the tercile trends shows, however, that very high R-IDLE does lead to a drop in fluency perception, as does very low C-ACT. Perhaps, the effects of these metrics plateau more quickly than the other two metrics. Examining the individual indicators, we see that the robot's personality traits were not strongly correlated with the objective fluency measures. General fluency, the robot's commitment, and team member contribution appear to be the more salient constructs in the composite scale.

Surprisingly, H-IDLE is positively correlated with fluency. A possible explanation is that people mainly considered the robot's contribution and thus saw the human's possibility to rest as a positive aspect of the collaboration.

The indicator "The human was the most important member" was the weakest indicator on the scale. Its inclusion in the scale reduces the scale's Cronbach's α from 0.84 to 0.74, and it is not correlated with any of the objective metrics that we evaluated. It should probably be eliminated in future studies.

VII. USE OF FLUENCY METRICS IN HUMAN–ROBOT COLLABORATION RESEARCH

As mentioned in Section I, the metrics proposed and discussed herein have been used in a growing number of human–robot collaboration studies. This section provides a current survey on their usages and related findings. This could shed light on which metrics were deemed more and less useful, although most authors did not explicitly consider this question. Table II shows a chronological list of metric usage.

The works fall into three broad categories: The majority of studies are *Shared Workspace* tasks, where agents bring objects to a shared area, such as a table, and can pick up or manipulate these objects at that workspace. This is akin to the simulation scenario used in the study in Section VI. Two smaller categories are *Handover* tasks, where objects are carried to a point where they are handed over from one agent to the other, and *Shared Manipulation* tasks, where agents concurrently manipulate objects, at least for some portion of the task time. This distinction is not perfectly sharp, as some tasks include elements of two or more collaborative activities.

A. Shared Workspace Tasks

The first suggestion of fluency evaluation was in a simulation workspace, where humans carried car parts to a shared workspace and an anticipatory robot assembled them [5]. Subjects were asked to rate a subset of five of the subjective metrics

TABLE II
CHRONOLOGY OF FLUENCY METRICS USED IN HUMAN–ROBOT COLLABORATION RESEARCH

Year	Authors	HRI Algorithm and Task	Objective Metrics Reported	Subjective Metrics Reported
2007	Hoffman & Breazeal [5]	Anticipatory action on simulated car assembly	H-IDLE, C-ACT, F-DEL	Robot contribution, fluency, human and robot commitment; human contribution; trust in robot
2010	Hoffman & Breazeal [10]	Perceptual simulation on physical "painting" task	H-IDLE, F-DEL	Full fluency questionnaire in Fig. 1
2011	Cakmak *et al.* [6]	Spatial and temporal contrast for human-robot handovers	Human F-DEL, Robot F-DEL	None
2012	Chao & Thomaz [7]	Timed Petri Nets for turn-taking interactions	Task time	Human and robot relative contribution, trust in robot, naturalness
2013	Nikolaidis & Shah [8]	Cross training for shared assembly task	C-ACT, R-IDLE, H-IDLE	Trust in robot, subset of of the WAI scale
2014	Unhelkar *et al.* [15]	Approach experiment in a delivery task	R-IDLE, H-IDLE and task time	Team "works well", robot's actions are "smooth", human "worked fluently" with robot
2015	Dragan *et al.* [16]	Motion planning for cup retrieval task	C-ACT, Task time, coordination time	Eight items adapted from Fig 1 relating to robot fluency and contribution, and trust in robot
2015	Nikolaidis *et al.* [2]	MOMDP for a hand-painting task	H-IDLE, Task time	Robot intelligence, accuracy, trustworthiness, and smoothness
2015, 2017	Gombolay *et al.* [17]–[19]	Decision-making roles in joint fetch-and-assembly tasks	H-IDLE, Task time, rescheduling time	21 items mostly adapted from Fig 1 relating to robot teammate traits, trust, Working Alliance, teamwork, commitment, and contribution
2015	Huang *et al.* [20]	Timing adjustments for human-robot handover	C-ACT, H-IDLE, R-IDLE, Task time	Undisclosed 5-item fluency questionnaire adapted from Fig 1
2016	Baraglia *et al.* [21]	Study of robot initiative for shared-workspace manipulation	C-ACT, H-IDLE, R-IDLE, Task time	Five-item interaction quality questionnaire including efficiency, fluency, and equal contribution
2016	Maniadakis *et al.* [22]	Multirobot collaboration planning	R-IDLE (separately for two robots), C-ACT	None
2017	Nikolaidis *et al.* [23]	Mutual adaptation in joint table-moving task	Other, non-fluency metrics	Ten-item scale with eight fluency questions selected from Fig. 1, including trust, Working Alliance, and positive robot traits
2017	Faria *et al.* [24]	Comparing trajectories in a liquid-pouring task	F-DEL	Undisclosed subset of questions in Fig 1 measuring perceived collaboration and fluency
2018	Rahman [25]	Virtual-physical conversation system for a collaborative search task	H-IDLE, R-IDLE, C-ACT, F-DEL	Trust in the system along with 11 non-fluency related scales

described above. Significant differences were found in the rating of the robot's contribution and commitment. In terms of objective metrics, the rate of concurrent motion was significantly higher in the anticipatory group, settling at approximately twice the rate compared to the reactive group. There was also a significantly lower F-DEL in the anticipatory group, but no difference in H-IDLE.

A first physical robot study evaluated fluency in the context of anticipatory perceptual simulation for collaborative robots [10]. In a human-subject study, there were significant differences in human–robot fluency, the improvement of the team, the robot's contribution, and the WAI goal subscale, as well as some individual measures. There was no significant difference in measures of the trust in the robot, the robot's character, or the WAI scale. Overall task efficiency was better in the experimental condition. In addition, two objective fluency metrics were measured: H-IDLE and the F-DEL incurred by the robot. Both were found to have been positively affected by the proposed algorithm.

Chao and Thomaz designed a system based on timed Petri Nets for multimodal turn-taking and joint action meshing. In a human-subject study, participants rated subjective fluency metrics relating to the relative contribution, trust, and naturalness of the interaction. Participants in the interruption condition rated their mental contribution higher and rated the interaction as less "awkward" than those in the baseline condition. Task efficiency was used as an objective metric of team fluency.

Nikolaidis and Shah proposed several algorithms to improve human–robot teaming including human–robot cross training [8] and mixed-observability Markov decision processes (MOMDP) [2]. In [8], authors used individual indicators from the "trust in robot" measure and adapted two indicators from the WAI "goal" subscale. The study also evaluated three objective metrics: concurrent motion, H-IDLE, and R-IDLE. There were significant improvements in all of these measures. In [2], researchers measured H-IDLE as an objective metric, and they did not find significant differences between the proposed algorithm and manual robot control. Subjective measures of the robot's intelligence, accuracy, trustworthiness, and smoothness were also measured and found to be similar in two variants of the algorithm.

Unhelkar *et al.* compared the performance of mobile robots with human assistants delivering parts to human workers [15]. Objective fluency metrics were interaction time and idle time, and subjective fluency metrics included overall performance, and how well the subject perceived the fluency and smoothness of the interaction. The results suggested no significant subjective difference between human and robot assistants. Increased idle time with the robot assistant indicated that human–human collaboration was more fluent.

Dragan *et al.* measured fluency for three different paths of motion in a coffee-making scenario [16]. Subjective metrics included perceived fluency, safety, comfort, trust, and contribution, whereas objective metrics included coordination time, total task time, and concurrent motion time. The results showed less coordination time with legible motion than predictable motion, with effects on the subjective view of the interaction.

Gombolay *et al.* investigated human–robot teams with varying degrees of robot control [17]. In related work, they tested how a robot should best incorporate human teammates' preferences into the team's schedule [18], [19]. In their work, objective metrics include assembly time, rescheduling time, and H-IDLE. The authors used most of the above-presented subjective fluency metrics along with satisfaction with the robot's performance, perceived productivity, and necessity of both agents in the interaction. Both objective and subjective metrics improved with complete robot control, and a positive correlation was found between team fluency and a human subject's willingness to collaborate with robots.

Baraglia *et al.* investigated when a robot should take initiative in a collaborative scenario. Three scenarios were tested, including a proactive robot, a reactive robot, and a human-initiated robot [21]. Objective metrics analyzed were concurrent motion, zero motion, H-IDLE and R-IDLE, human-only movement time, and robot-only movement time; subjective metrics included the robot's helpfulness, its awareness of human and task progress, its contribution to the task, its overall fluency, efficiency, and the naturalness of the interaction. The findings suggested proactive robots led to lower R-IDLEs without a significant change in H-IDLE, while subjective ratings were not significantly different between both proactive and human-initiated robot conditions.

Maniadakis *et al.* analyzed a proposed system for collaboration planning using fuzzy time intervals and constraints [22]. They demonstrated it in a joint cooking task between two robots. Objective fluency metrics included idle times and concurrent action. The planner was shown to have lower idle times, facilitating C-ACT.

Finally, Rahman studied the collaboration of a virtual human and a humanoid robot through a cyber-physical system, in an experiment where both work together to find a hidden object [25]. Objective metrics were leader idle time, follower idle time, nonconcurrent activity, and F-DEL. The results indicated more fluent collaboration when the humanoid robot acted as the master agent, correlated with less idle time, nonconcurrent activity, and F-DEL.

B. Handover Tasks

Cakmak *et al.* developed methods using spatial and temporal contrast to enable more fluent handovers from a robot to a human [6]. A survey was used to estimate the readability of handovers, and in an experimental human-subject study, two objective measures of fluency were evaluated, including human F-DEL and robot F-DEL. The researchers found that temporal contrast positively affects human F-DEL.

Huang *et al.* assessed human–robot handovers in unloading a dish rack [20]. Objective fluency metrics included C-ACT, completion time, H-IDLE, and R-IDLE; a subjective measure on a five-item scale was adopted from the questions in Fig. 1. The results suggested that proactive coordination improved performance in terms of C-ACT and idle time but impaired a user's perceived fluency, while reactive coordination did the inverse, suggesting a complex relationship between subjective and objective fluency.

C. Shared Manipulation Tasks

Most recently, fluency metrics have been applied in physical shared manipulation tasks. Nikolaidis *et al.* analyzed models for human–robot mutual adaptation in a collaborative table-holding scenario [23]. The authors used ten subjective metrics, including measures of trust, WAI, and satisfaction with the robot. The findings demonstrated that their model of human adaptation was better at establishing human–robot trust.

Faria *et al.* evaluated the impact of different motion types in a collaboration scenario between a robot and multiple people [24]. Delay was used to evaluate objective performance. Subjective metrics evaluated user perception of fluency using an undisclosed questionnaire. The results found that although legible motions are more expressive than predictable, workspace configuration and the existence of other bodies are crucial in humans' understanding of the robot's objective.

VIII. Conclusion and Future Work

In this paper, we discuss metrics to evaluate human–robot collaborative fluency—the successful coordination and meshing of actions in a team. These metrics include subjective measures made up of internally valid scales and individual indicators. They also include four objective measures that could provide benchmarks for evaluating the fluency of a human–robot collaborative interaction.

Subsets of these metrics have been used in the past years to evaluate human–robot fluency in our own work and in that of other researchers. However, this is the first attempt to systematically examine fluency metrics and to empirically evaluate the validity of the objective metrics in terms of a subset of the subjective metrics used. To do so, we presented analytical forms of idealized metrics as well as insights from computer simulations of a minimal shared workspace or handover model. Our findings indicate that the temporal dynamics of the proposed metrics may be more complex than previously considered and should be taken into account by human–robot collaboration researchers.

Relating objective and subjective metrics, we find F-DEL to have the strongest correlation with subjective fluency perception, and that it mostly correlates with team fluency, robot contribution, and commitment indicators. H-IDLE also shows a significant correlation with fluency perception. Anecdotally, R-IDLE and C-ACT appear to be related only in their extreme nonfluent case.

Effect sizes were not large in our study. This could be because fluency perception by an outside observer is not as sensitive as that of a participant in the collaboration. We are currently working on a participant-centric version of the shared workspace study described above, as well as an experimental protocol examining causal relationships between high-fluency robot policies and the human's sense of fluency.

A literature review of the preferred use of fluency metrics in the past decade shows a roughly equal use of the four objective metrics presented in this paper, often used in combination with overall task time. H-IDLE slightly leads the use table of objective metrics. Researchers have found general fluency, along with trust and robot contribution to be the most useful subjective metrics. This observation is in line with our own findings here, suggesting

these metrics to be more applicable than others. That said, some of the presented subjective metrics are minimal and best fit for a fast assessment of the downstream effects of HRI. For some of these constructs, more extensive metrics have been studied, for example, Muir and Moray's work on human–machine trust [26]. Such validated measures should be considered in future research on the downstream effects of fluency.

In summary, it is worth noting that fluency in human–robot collaboration is not a well-defined construct and is inherently somewhat vague and ephemeral. That said, our work is based on the contention that fluency is a quality that can be positively assessed and recognized when compared to a nonfluent scenario. Moreover, it is the very tacit nature of fluency that necessitates tools for its evaluation toward the design of successful robotic teammates.

There are aspects of collaborative fluency that these metrics do not yet address and that should be considered for future work. These aspects include: How to take into account correct and incorrect actions of the robot and the human? How can we model uncertainty in action start and end times? Does the role relationship between human and robot (e.g., supervisor, subordinate, or peer) affect perceptions of fluency? How should one account for corrections and repetitions of identical actions? And how can one extend these measures to larger teams than just one human and one robot? It is also worth noting that the metrics presented here are highly dependent on the definition of activity start and end points. These points may be task-specific and ambiguous, and their specification may significantly affect the resulting metrics.

The metrics herein are an evolving work in progress. Over the years, we have added, refined, and removed some of these metrics from our inventory. This work is a step to systematically review and validate the metrics as they have been used, in order to develop a generally agreed-upon set of fluency metrics that can serve the human–robot collaboration community. This can enable clearer benchmarks to compare human–robot collaborative systems, advancing the goal of designing more useful robotic team members.

Acknowledgment

The author would like to thank R. Slyper, K. Vanunu, Y. Feldman, and G. Struble for their assistance in this project.

References

[1] G. Hoffman and C. Breazeal, "Collaboration in human-robot teams," in *Proc. AIAA 1st Intell. Syst. Tech. Conf.*, 2004, Paper AIAA 2004-6434.

[2] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2015, pp. 189–196.

[3] J. Shah, J. Wiken, B. Williams, and C. Breazeal, "Improved human-robot team performance using Chaski, a human-inspired plan execution system," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2011, pp. 29–36.

[4] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational human-robot interaction," *Found. Trends Robot.*, vol. 4, nos. 2/3, pp. 105–223, 2016.

[5] G. Hoffman and C. Breazeal, "Cost-based anticipatory action-selection for human-robot fluency," *IEEE Trans. Robot.*, vol. 23, no. 5, pp. 952–961, Oct. 2007.

[6] M. Cakmak, S. Srinivasa, M. Kyung Lee, S. Kiesler, and J. Forlizzi, "Using spatial and temporal contrast for fluent robot-human hand-overs," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2011, pp. 489–496.

[7] C. Chao and A. L. Thomaz, "Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 4–25, 2012.

[8] S. Nikolaidis and J. Shah, "Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2013, pp. 33–40.

[9] G. Hoffman, "Evaluating fluency in human-robot collaboration," in *Proc. Robot.: Sci. Syst. Workshop Human-Robot Collaboration*, 2013.

[10] G. Hoffman and C. Breazeal, "Effects of anticipatory perceptual simulation on practiced human-robot tasks," *Auton. Robots*, vol. 28, no. 4, pp. 403–423, May 2010.

[11] A. O. Horvath and L. S. Greenberg, "Development and validation of the Working Alliance Inventory," *J. Counseling Psychol.*, vol. 36, no. 2, pp. 223–233, 1989.

[12] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: Improving robot readability with animation principles," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2011, pp. 69–76.

[13] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How quickly should communication robots respond?" in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2008, pp. 153–160.

[14] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?" *Perspectives Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[15] V. V. Unhelkar, H. C. Siu, and J. A. Shah, "Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2014, pp. 82–89.

[16] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2015, pp. 51–58.

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

[18] M. Gombolay, C. Huang, and J. A. Shah, "Coordination of human-robot teaming with human task preferences," in *Proc. AAAI Fall Symp. Ser. AI-HRI*, vol. 11, 2015, pp. 68–73.

[19] M. Gombolay, A. Bair, C. Huang, and J. Shah, "Computational design of mixed-initiative human-robot teaming that considers human factors: Situational awareness, workload, and workflow preferences," *Int. J. Robot. Res.*, vol. 36, nos. 5–7, pp. 597–617, 2017.

[20] C. Huang, M. Cakmak, and B. Mutlu, "Adaptive coordination strategies for human-robot handovers," in *Proc. Robot.: Sci. Syst. Conf.*, 2015.

[21] J. Baraglia, M. Cakmak, Y. Nagai, R. Rao, and M. Asada, "Initiative in robot assistance during collaborative task execution," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2016, pp. 67–74.

[22] M. Maniadakis, E. Hourdakis, and P. Trahanias, "Time-informed task planning in multi-agent collaboration," *Cogn. Syst. Res.*, vol. 43, pp. 291–300, 2017.

[23] S. Nikolaidis, D. Hsu, and S. Srinivasa, "Human-robot mutual adaptation in collaborative tasks: Models and experiments," *Int. J. Robot. Res.*, vol. 36, nos. 5–7, pp. 618–634, 2017.

[24] M. Faria, R. Silva, P. Alves-Oliveira, F. S. Melo, and A. Paiva, ""Me and you together"; movement impact in multi-user collaboration tasks," in *Proc. Int. Conf. Intell. Robots Syst.*, 2017, pp. 2793–2798.

[25] S. M. Rahman, "Cyber-physical-social system between a humanoid robot and a virtual human through a shared platform for adaptive agent ecology," *IEEE/CAA J Autom. Sinica*, vol. 5, no. 1, pp. 190–203, 2018.

[26] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

Guy Hoffman (M'10) received the M.Sc. degree in computer science, as part of the Adi Lautman interdisciplinary excellence scholarship program, from Tel Aviv University, Tel Aviv, Israel, in 2000, and the Ph.D. degree in human–robot interaction from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2007.

He is currently an Assistant Professor and the Mills Family Faculty Fellow with the Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, USA. Prior to that, he was an Assistant Professor with IDC Herzliya and co-Director of the IDC Media Innovation Lab.