Metrics for Robot Proficiency Self-Assessment and Communication of Proficiency in Human-Robot Teams

ADAM NORTON, University of Massachusetts Lowell
HENNY ADMONI, Carnegie Mellon University
JACOB CRANDALL, Brigham Young University
TESCA FITZGERALD, Carnegie Mellon University
ALVIKA GAUTAM, Brigham Young University
MICHAEL GOODRICH, Brigham Young University
AMY SARETSKY, University of Massachusetts Lowell
MATTHIAS SCHEUTZ, Tufts University
REID SIMMONS, Carnegie Mellon University
AARON STEINFELD, Carnegie Mellon University
HOLLY YANCO, University of Massachusetts Lowell

As development of robots with the ability to self-assess their proficiency for accomplishing tasks continues to grow, metrics are needed to evaluate the characteristics and performance of these robot systems and their interactions with humans. This proficiency-based human-robot interaction (HRI) use case can occur before, during, or after the performance of a task. This paper presents a set of metrics for this use case, driven by a four stage cyclical interaction flow: 1) robot self-assessment of proficiency (RSA), 2) robot communication of proficiency to the human (RCP), 3) human understanding of proficiency (HUP), and 4) robot perception of the human's intentions, values, and assessments (RPH). This effort leverages work from related fields including explainability, transparency, and introspection, by repurposing metrics under the context of proficiency self-assessment. Considerations for temporal level (*a priori*, *in situ*, and *post hoc*) on the metrics are reviewed, as are the connections between metrics within or across stages in the proficiency-based interaction flow. This paper provides a common framework and language for metrics to enhance the development and measurement of HRI in the field of proficiency self-assessment.

CCS Concepts: • **Computer systems organization → Robotics**; **Robotic autonomy**; • **General and reference → Metrics**; • **Human-centered computing → Empirical studies in interaction design**.

Additional Key Words and Phrases: human-robot interaction, proficiency self-assessment, metrics, performance evaluation

Authors' addresses: Adam Norton, University of Massachusetts Lowell, 1 University Ave, Lowell, MA, 01852, adam_norton@uml.edu; Henny Admoni, Carnegie Mellon University, henny@cmu.edu; Jacob Crandall, Brigham Young University, crandall@cs.byu.edu; Tesca Fitzgerald, Carnegie Mellon University, tescaf@andrew.cmu.edu; Alvika Gautam, Brigham Young University, alvikag@byu.edu; Michael Goodrich, Brigham Young University, mike@cs.byu.edu; Amy Saretsky, University of Massachusetts Lowell, amy_saretsky@student.uml.edu; Matthias Scheutz, Tufts University, Matthias.Scheutz@tufts.edu; Reid Simmons, Carnegie Mellon University, rsimmons@andrew.cmu.edu; Aaron Steinfeld, Carnegie Mellon University, steinfeld@cmu.edu; Holly Yanco, University of Massachusetts Lowell, holly@cs.uml.edu.

Manuscript submitted to ACM

ACM Reference Format:
Adam Norton, Henny Admoni, Jacob Crandall, Tesca Fitzgerald, Alvika Gautam, Michael Goodrich, Amy Saretsky, Matthias Scheutz, Reid Simmons, Aaron Steinfeld, and Holly Yanco. 2021. Metrics for Robot Proficiency Self-Assessment and Communication of Proficiency in Human-Robot Teams. 1, 1 (February 2021), 37 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Robots that can self-assess their abilities to perform tasks can potentially improve human-robot interaction (HRI). *Proficiency self assessment* (PSA) is the ability of a robot to predict, estimate, or measure how well it can perform a task in a given context and environment. Human-robot teaming not only benefits from identifying a set of practicable metrics for PSA, but also from developing metrics that evaluate how the robot communicates its proficiency to a human, how the human understands the communication, and how the robot perceives the human. Together, these metrics enable comprehensive evaluation of human-robot teaming.

Accurate self-assessment is a feat that some human experts exhibit. They know what they can or cannot do under a variety of circumstances, they can often estimate the likelihood of success (though sometimes with biases [12, 77]), and they usually know how well they can do it. Human self-assessment is grounded in extensive experience of one's own behavior, requiring perceptive observation of environmental conditions and introspective access to one's abilities, limitations, and goals. Moreover, humans can include their knowledge of other humans and their capabilities in performance assessments of tasks that involve multiple humans. Human self-assessment serves as evidence that robots should also be able to self-assess their proficiency provided that they are equipped with the necessary metrics.

This paper presents a review of PSA metrics as well as metrics for evaluating how PSA impacts communication and other aspects of human-robot teaming. The metrics lead to the following operational definition of task-based proficiency: The extent to which a given robot, its sensors, actuators, and computational resources is *proficient* at a task is determined by four factors: 1) the probability and extent to which the robot will accomplish the task (i.e., achieve the task goal or set of task goals), 2) within a time bound or throughout a time period, 3) given a set of environmental variations, and 4) relative to contextual standards. *Proficiency assessment* is the ability to accurately make assertions about a robot's proficiency given the task, the context, the robot's observations about and behaviors within the world, relative to contextual proficiency standards. Naturally, the term *proficiency self-assessment* refers to proficiency assessment performed by the robot about itself and the teams in which it may participate.

Assessment can be performed at multiple temporal phases: *a priori* (before the task is executed), *in situ* (while the robot is performing its task or mission), or *post hoc* (after the mission terminates). *A priori* assessment enables performance predictions; *in situ* assessment enables adaptation to changes or transformation to new goals or operational envelopes [4, 9, 69]; and *post hoc* assessment leads to evaluations that can inform future behaviors and longer term learning. PSA can take on three increasingly sophisticated forms:

(1) An *estimate* of the probability that (or extent to which) a robot is proficient, perhaps accompanied by information about the uncertainty associated with the estimate,
(2) *Measurements* from a set of *metrics* that correlate, predict, or set bounds on a robot's proficiency, or
(3) An *explanation* of the causal factors that led to a particular assertion about proficiency.

Not every PSA form is possible in every problem, so PSA metrics naturally cover a range of forms.

This paper's perspective is that proficiency self-assessment is part of a larger system of humans and possibly other robots or agents. For ease of exposition, it is assumed (though this assumption is revisited in the discussion, Section 7) that the robot is part of a simple human-robot interaction dyad, where the behavior of the robot impacts the human in some meaningful way. More specifically, this paper adopts a rhetorical framing where the human is the *problem holder* [164] and the robot is assigned to perform tasks or accomplish goals in pursuit of the problem held by the human. Thus, the paper emphasizes metrics for robots to self-assess proficiency and to communicate its self-assessed proficiency to a human partner, leading to the important communicative dimension of proficiency-based assertions: the efficacy of the process of communicating proficiency between human and robot. While the rhetorical framing of a team comprised of a single human and single robot is used, the metrics in this paper apply not only to human-robot dyads, but also to teams with multiple human or artificial partners.

This paper is organized as follows: Section 2 describes the scope of the paper, including a review of related concepts, a summary of relevant roles in human-robot teaming, a temporal flow for how PSA is embedded in human-robot teaming, an illustrative case study, and limitations of the paper. Sections 3–6 describe metrics for each state in the temporal flow. Section 7 summarizes the metrics, identifies relationships between the metrics, and discusses open problems on PSA in human-robot teaming. Finally, Section 8 presents conclusions.

2 SCOPE

Proficiency self-assessment and communication of such is related to several other areas of research in autonomous systems and HRI. In this section, related work is reviewed and compared to the area of proficiency self-assessment as a means of positioning it within HRI research. Throughout the paper, metrics from these related research areas are leveraged to form the basis of metrics for robot proficiency self-assessment and communication of proficiency. Connections and overlaps between related research areas were considered; as such, some metrics are combined and/or recontextualized to match this domain. New metrics and evaluation criteria are also proposed. This section also presents four stages of proficiency-based human-robot interaction scenarios for which metrics are defined along with an example scenario that will be referenced throughout the remainder of the paper.

2.1 Concepts Related to Proficiency Self-Assessment

Proficiency assessment is closely related to *explainability* in artificial intelligence (explainable AI, or XAI). Hoffman et al. [70] identify three purposes of XAI: “How does [the AI] work?”, “What mistakes can [the AI] make?”, and “Why did [the AI] just do that?” XAI emphasizes causal factors that a human can use to calibrate trust in and reliance on decisions made by an AI algorithm. The explanation of the causal factors might include bounds on the algorithm’s confidence in its performance or reliability. The overlap between proficiency self-assessment and XAI would include intersecting sets, but neither XAI nor proficiency self-assessment is a subset of the other. For example, proficiency self-assessment might include an explanation, but it might also be a clear statement about how proficient the agent is without any explanation. The complement to this is when XAI includes a discussion of what bounds influence the competency of the algorithm without yielding a clear assertion about whether the algorithm will be useful in the present context. Furthermore, explanations of “Why did the AI just do that?" emphasize *post hoc* evaluation and may de-emphasize *a priori* and *in situ* assertions about likely success. Finally, proficiency self-assessment can be used by an agent to autonomously *initiate* a change in goals or behaviors, thus supporting mixed initiative interactions that tend to be outside the scope of XAI.

Communicating proficiency is also closely related to *transparency* in human-machine interaction. Chen et al. define transparency as “... the descriptive quality of an interface pertaining to its abilities to afford an operator’s comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process" [21]. These elements of transparency suggest that the concepts of transparency and proficiency self-assessment overlap, but, as with proficiency self-assessment and XAI, neither is subsumed by the other. Transparency emphasizes *in situ* assessments of the causal factors contributing to agent behavior including the beliefs and motivations of the agent, and can include projections of the likely success of the agent in its task. Transparency is a property of an interface, and the interface may not explicitly report real-time performance metrics [134] or other assessments of proficiency.

Endsley's three levels of situation awareness (SA) – perception, comprehension, and projection [40] – are widely used throughout HRI research and are particularly relevant to metrics definition for robot self-assessment of proficiency. The robot's proficiency measures could identify shortfalls or alignments between required capability to perform a task and robot capability (perception), explaining or reasoning as to why and the degree to which success or failure is likely (comprehension), or predicting the robot's ability to accomplish a task (projection). Communication of proficiency can also be categorized using these levels of SA, similar to Chen et al.'s situation awareness-based agent transparency (SAT) model [21]. In the SAT model, an autonomous agent's goals and actions (level 1 SAT), reasoning process (level 2 SAT), and projections/predictions (level 3 SAT) optionally with associated uncertainty measures (U) can be communicated transparently to improve team effectiveness. The SAT model has been used to develop interfaces that adhere to its principles at level 1, 1+2, 1+2+3, and 1+2+3+U [136], the latter of which is most related to communicating proficiency self-assessment due to its inclusion of uncertainty measures. Given that Endsley's three levels of SA are primarily used for categorizing human understanding, they are used as such for metrics of human understanding later in this paper.

Throughout the remainder of the paper, relevant metrics from each of these related concepts are referenced if similar measures exist. The intent of this paper is to leverage metrics from these areas and recontextualize them as appropriate as well as propose new metrics that are particular to each stage of the proficiency-based interaction flow.

2.2 Human-Robot Roles

For the purposes of this paper, the human is the problem holder, meaning they act as a partner who is teamed with the robot and who requested the robot to perform a task. The human's role is therefore most akin to that of a supervisor, operator, or teammate [133], in that they have some experience with the robot and are intended to work with or alongside it. More specifically, the human's role as the problem holder is somewhere between that of an information consumer [56] (receives and uses information from the robot to make decisions) and an abstract supervisor [71] (uses the information received from the robot to modify its objectives and goals). The experience level of the human may vary, though, ranging from novice to expert, which will impact the human's understanding of the robot's proficiency and therefore how the robot should communicate its proficiency. As part of the human-robot team, the robot can serve multiple roles including individual support, team support, or as a team member [153]. The robot may consider the implications of the human-robot roles during interactions to influence how it chooses to communicate its proficiency. These human-robot roles set the use case for the metrics reviewed in this paper; i.e., other roles such as the human as a bystander or the robot serving a social role are not considered.

2.3 Proficiency-Based Interaction Flow

To frame this paper's discussion of metrics, a proficiency-based interaction flow that consists of four stages is proposed:

- RSA: **R**obot performs **S**elf-**A**ssessment of proficiency,
- RCP: **R**obot **C**ommunicates its **P**roficiency to the human,
- HUP: **H**uman processes the communication and conveys their **U**nderstanding of the robot's **P**roficiency, and
- RPH: **R**obot **P**erceives the **H**uman's intentions, values, and assessments.

This four-stage interaction flow is shown in Figure 1. For a given proficiency-based interaction, all stages may not occur (e.g., interaction with a robot that does not possess the required perception capabilities for RPH will skip this stage), but the flow is intentionally abstract to encompass all proficiency-based interactions. The metrics at each stage can be used to evaluate human-robot interactions that include robot proficiency self-assessment and communication of proficiency, or can be used by the robot system to evaluate its own performance. The use of the metrics by robot systems may be particularly important for robots that attempt to improve their communications of proficiency to humans over time. Metrics are reviewed for each stage of the proficiency-based interaction flow in the following sections. The metrics reviewed at the RSA and RPH stages are categorized and defined similarly (both deal with evaluations of the task and teaming), as are those at the RCP and HUP stages (both deal with communications). Connections between the metrics within and across stages is detailed in each section as appropriate.

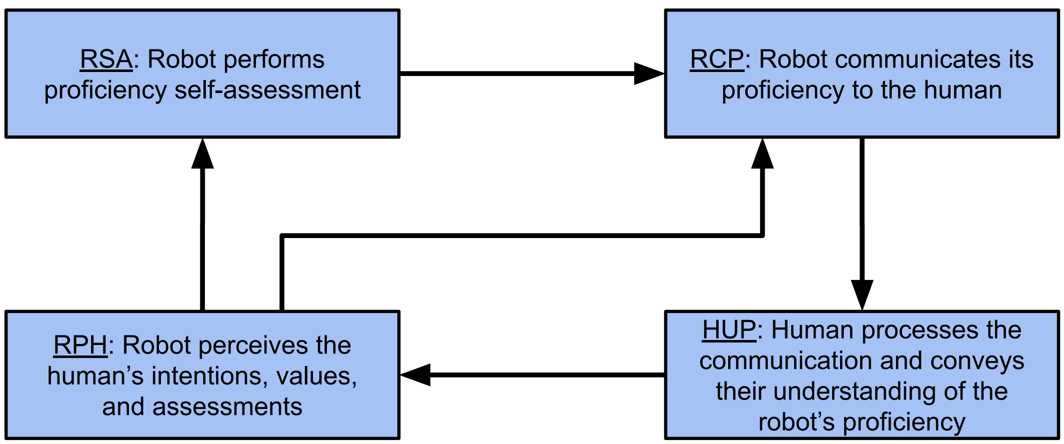

Fig. 1. Proficiency-based interaction flow.

Each stage impacts subsequent stages in the flow. The robot's self-assessed measure of proficiency at the RSA stage will be converted into a communicable form at the RCP stage. When that proficiency is communicated, the human will attempt to understand it at the HUP stage. When the human conveys that understanding back to the robot (e.g., by pointing and asking about an object the robot referenced in its communication), the robot may perceive the human's behavior in the RPH stage to infer information about the human, such as the human's intention. Based on how the robot uses this information in the RPH stage, the RSA stage may be repeated (e.g., if the human made any physical updates to the task to improve the likelihood of success) which may include updating assessments about its own proficiency (e.g., if the human suggests different strategies to assist in accomplishing the task). Alternatively, the interaction may progress from the RPH stage to the RCP stage wherein the robot communicates its proficiency again using a different method (e.g., if it perceived the human did not understand the previous communication) and/or updating the information being communicated (e.g., if the human asked a clarifying question in response to the previous communication).

A summary of the metrics presented throughout the rest of this article can be seen in Table 1, organized by the stage of the proficiency-based interaction flow they are associated with and grouped into categories.

Stage	Category	Metrics
Robot self-assessment of proficiency (RSA)	Uncertainty	Alignment of uncertainty and performance Risk-averse reward Uncertainty reduction
	Performance	Mission progress Replanning triggers Accuracy vs. rejection curves
	Time	Predicted vs. actual completion time Productive time Reliability Forecasting time
	Events	Interventions Repeated attempts Violation of performance envelopes
Robot communication of proficiency (RCP)	Attributes	Information communicated Nature of communication Communicability
	Complexity	Abstraction Clutter Comprehensiveness Simplicity Size
	Efficiency	Communication time Conversion time Communication latency Transmission time
Human understanding of proficiency (HUP)	Perception	Perception clarity Perception completeness Perception time
	Comprehension	Comprehension clarity Comprehension completeness Communication consistency Content quality Processing difficulty Comprehension time
	Projection	Expectations Projection clarity Congruity Command changes Environment changes
Robot perception of human intentions, values, and assessments (RPH)	Uncertainty and coherence	Model consistency Behavior consistency Model uncertainty Model prediction accuracy
	Performance	Mission progress Replanning triggers Violation of performance envelopes
	Time	Persistence Expected execution time Coordination time Event timing
	Events	Corrections Interventions Modifications Communications
	Human factors	Workload Stress Trust Situation awareness Violations of human performance limitations

Table 1. Summary of metrics at each stage of the proficiency-based interaction flow.

2.4 The Fetch Scenario

To ground the discussion of metrics to a common reference point, a human-robot interaction scenario is defined that will be referenced throughout the paper: a mobile manipulator robot is tasked with finding bolts, screws, and other components to place into a specific type of bin and then deliver the bin to another station for the human to use to build a gearbox (Fig. 2). The process will be performed continuously until several bins are filled by the robot and used by the human to build several gearboxes, aiming to produce a specified number of gearboxes per hour. The human problem holder in this scenario has commanded the robot to perform the task, will monitor the robot's performance, can ask questions about progress made, can intervene to augment the scenario as needed, and will inspect the final state of the bin before building the gearbox. This scenario is similar to that described in Frasca et al. [50]. Throughout the rest of this paper, this scenario is referred to as "the Fetch scenario." Robot proficiency self-assessment techniques and communication methods are not specified here, but will be as needed when the scenario is referenced. To illustrate the proficiency-based interaction flow in Figure 1, below is an example of the events at each stage when occurring *in situ* for the Fetch scenario:

- **RSA**: The robot conducts proficiency self-assessment at picking up a bin filled with screws, producing a measure of low confidence at succeeding.
- **RCP**: The robot speaks to the human, "I am unlikely to pick up the bin without dropping it."
- **HUP**: The human heard the robot, but does not fully understand why the robot has low confidence, so they follow up and ask "Why is your confidence low?"
- **RPH**: The robot perceives this communication and decides that it will communicate its proficiency again with more detail (i.e., advancing to the RCP stage next).

Fig. 2. Example HRI scenario referenced throughout the paper as "the Fetch scenario."

2.5 Limitations

It should be noted that the metrics presented in this paper are each at varying levels of development, maturity, and prominence. Some metrics have not yet been robustly validated through experimentation while others have been utilized substantially throughout research albeit in different contexts. Readers are encouraged to refer to the research articles cited for each metric (if available) for further information regarding implementation and nuances. The authors acknowledge that the newly proposed metrics in this article may be less substantially defined than others. Continued definition, development, and comparison of these metrics towards effective implementation will be required as proficiency self-assessment research progresses, so the metrics presented in this paper should be considered a starting point for the field.

3 METRICS FOR ROBOT SELF-ASSESSMENT OF PROFICIENCY (RSA)

The role of the robot within a team affects how proficiency is evaluated and, as a result, informs the metrics used to perform this evaluation. In the context of a robot that collaborates with humans to complete a task, the robot should perform self-assessment over its task knowledge, determining whether it has sufficient knowledge and capabilities to complete the task or should, instead, interact with a human teammate to request additional assistance (e.g., request additional training data). In the context of a fully autonomous robot, the robot may assess its task knowledge and capabilities to determine whether it should accept or reject the task, or otherwise inform its human collaborator about its proficiency. In either setting, it is important for the robot to assess its uncertainty when attempting to complete a task according to its current task knowledge. Uncertainty based metrics for self-assessment are discussed in Section 3.1.

In addition to assessing uncertainty, the robot may need to assess the performance of its actions and whether they are producing the expected outcome. This performance assessment may occur at multiple stages of task planning and execution. For example, failures may occur during task planning, after which (*post hoc*) the robot will need to update its planning parameters or otherwise attempt the task in another way. Alternatively, self-assessment may involve evaluating *a priori* whether a candidate plan will achieve the task goals (e.g., using a goal reasoning framework [75]). Finally, self-assessment may serve to monitor *in situ* the robot's actual performance during task execution in comparison to its expected performance. Metrics for assessing performance are discussed in Section 3.2.

Task uncertainty and performance may also affect another aspect of human-robot teams. In the context of mixed-initiative human-robot teams, the robot's level of autonomy is dynamic and ideally determined according to the robot's own proficiency at performing a task. In this setting, the robot should have the capability to deliberate over its current level of autonomy to efficiently use human assistance when it is required and available. Additionally, the robot would ideally assess the *type* of assistance that would enable it to optimally address a particular problem, which may involve identifying the interaction modality and/or constraints under which that assistance should be obtained [43]. Thus, such *events* along with *timelines* of certain operations that may impact the robot's role in the team can be simple, but useful, indicators of potential task failures, the robot's autonomous capabilities, and overall task performance. Time and event based metrics are discussed in Sections 3.3 and 3.4 respectively.

3.1 RSA: Uncertainty

This category of metrics relates to the robot's confidence in its ability to complete the task to a particular performance standard. Uncertainty is different from performance prediction; an agent can have high confidence that a particular sequence of actions will result in task success or failure, whereas a robot that is *uncertain* in its task model may produce a range of possible outputs without a means of distinguishing higher-performing outputs from lower-performing ones. While the following metrics are largely used for *in situ* self-assessment, metrics of uncertainty can also be used *a priori* and *post hoc*. For example, Johnson et al. [75] evaluated model uncertainty according to an acceptable bound of uncertainty, considering the maximum uncertainty to be that of the model before any training data has been obtained (i.e., *a priori* self-assessment). The level of "acceptable" uncertainty (i.e., the level of uncertainty at which the model's output is accepted), then, could then serve as a useful *post hoc* metric.

Alignment of uncertainty and performance. This metric is measured through correlations between the variance of the model's output and the actual task performance achieved using that output. Fleming and Daw [46] used a similar metric for evaluating uncertainty-based approaches, correlating the confidence in the model to the actual error incurred by its output; they also used the correlation between the model confidence and the strength of the input stimulus as a metric. When a robot may represent a task according to multiple models, the uncertainty of the candidate models may indicate which is best suited for a particular learning problem. For example, Fitzgerald et al. [45] evaluated the correlation between (i) the uncertainty of two models trained over a set of noisy training data and (ii) their respective performance in performing a novel task. The resulting comparison is an evaluation of whether model uncertainty serves as a proxy for expected task performance.

Risk-averse reward. The total reward over a sequence of actions by the robot, based on the cost of correct, incorrect, and undecided actions is used to calculate this metric. When used as a reward function in self-assessment problems, risk-averse metrics consist typically of three rewards: the reward associated with (1) making a correct decision, (2) making an incorrect decision, and (3) making no decision (and thus rejecting the task). An agent trained to optimize this reward function via reinforcement learning will (ideally) learn to gauge its confidence in a decision such that it rejects the task when it is uncertain in its model's output. When used as a metric for robot self-assessment (e.g., [27, 73, 171]), the cumulative reward over a series of tasks reflects the robot's ability to assess risk. This cumulative reward is increased by rejecting tasks in lieu of achieving poor performance on them, while also attempting tasks that the robot expects to complete successfully.

Uncertainty reduction. Calculating the accuracy in predicting what training data will reduce the robot's uncertainty is another means of evaluating a robot's ability to assess which training data points will improve its task model. In this context, it is assumed that the robot has access to a human teacher who may provide additional training data to the robot. If the robot's assessment of its uncertainty is correct, it will be able to correctly identify training samples that, once labeled by a teacher, will lead to reduced uncertainty in the overall task model. Rakicevic and Kormushev [119] presented an online active learning approach for a task learning and transfer application. The exploration component was decoupled from the task model and performed an informed search in the trial-parameter space to generate the subsequent most informative trials, by simultaneously exploiting information from previous trials and reducing the task model's overall uncertainty.

3.2 RSA: Performance

This category of metrics relates to the extent to which the robot was able to complete its task. While performance and proficiency are not the same thing, measures of performance can be used to provide information about the robot's proficiency. A robot that is performing a task poorly when previously believing it was proficient at the task should consider whether the failure is a result of a lack of proficiency or simply circumstances outside its control. When a robot has performed the same or similar tasks multiple times in the past, then it can use its past performance in these tasks as an *a priori* assessments of its proficiency [73, 171]; this assumes that the robot is capable of determining how past performance relates to the current task, particularly the current context.

Performance metrics primarily relate only to *in situ* and *post hoc* applications. Cumulative scores, rewards, and penalties can be assessed after a mission to determine how well a task was performed. Additionally, performance summaries [135] can be used to evaluate proficiency. Estimates of performance and proficiency can be further evaluated via subjective scoring by the human at the HUP stage (see Section 5). Additionally, when evaluating a robot's self-assessment ability, a set of experts may use appropriate behavioral coding standards to define a "ground truth" label for proficient and inproficient robot behavior.

Mission progress. In sequential tasks, mission progress refers to the number or percentage of sub-tasks that have been successfully completed. Real-time estimates of mission progress or partial summaries of mission state can include reaching mileposts, satisfying preconditions or postconditions, and deviations from scripts [134, 135]. When the robot achieves a milestone or satisfies necessary preconditions or postconditions, it can potentially increase its confidence in its proficiency in that task. On the other hand, failures to achieve milestones or satisfy preconditions or postconditions might reflect negatively on the robot's proficiency in the absence of a more detailed understanding to explain away such failures.

Replanning triggers. The metric evaluates the rate or frequency of replanning a mission. Replanning can be triggered by changes in the environment [51, 167], lack of progress [161], at regular time intervals [26], or when a plan drifts away from the problem-holder's intention [137], or when a goal-reasoning system indicates that the goal currently being pursued cannot be accomplished [4]. Changes to the environment or robot that lead to state spaces in which the robot is no longer proficient should be detected in order to enable replanning.

Accuracy vs. rejection curves. The relationship between the agent's task rejection rate and its performance on accepted tasks. An accuracy vs. rejection curve reflects (1) the range of risk assessments (rejection rates) that are plausible for a particular task, and (2) the acceptable level of risk for a particular task according to the corresponding performance expectation. In practice, a single rejection rate may be selected such that the robot rejects tasks that exceed that risk threshold. This metric has been demonstrated in the context of assessing the reliability of a robot's perception [27, 171] and to evaluate the tradeoff between the accuracy and autonomy afforded to a robot as a function of how much assistance it receives from a human teacher [44]. Overall, accuracy vs. rejection curves provide a *post hoc* evaluation of the robot's performance across a range of risk tolerances.

3.3 RSA: Time

This category of metrics relates to the robot's uncertainty and expected performance over time. The extent to which a robot is proficient at a task is highly governed by the probability that the robot will accomplish the task within a *time bound.* In a human-robot interaction context, time-based thresholds are often used as indicators to predict and evaluate failures and to aid the robot in making decisions as to when to ask for human assistance to recover from a failure.

Predicted vs. actual completion time. Accuracy in predicting the robot's task completion time according to changes in system and environment. Assessment of predicted task completion time against a time threshold is a another metric for how proficiently the robot will be able to complete a task. In the Fetch scenario, the robot can use the planner's output of "predicted time to execute a particular trajectory" to decide *a priori*, based on a predefined time threshold, whether to re-plan or to execute the recommended trajectory. Additionally, the combination of uncertainty and time (e.g., the ratio of maximum uncertainty, as discussed in Section 3.1, to expected task completion time) can be a useful metric to choose among goals relevant to the task when there are conflicting goals. This metric may also serve to measure the difference between the predicted failure time and actual time when a failure occurred [132]. A comparison of predicted completion time to actual completion time as a *post hoc* self-assessment is a simple indicator of how well a task was performed.

Productive time. The duration of time in which the robot operates autonomously is its productive time. This metric, which is primarily applicable in supervisory control tasks, measures the continuous amount of time that the agent can remain autonomous before requiring assistance. It may serve as an *in situ* metric towards measuring a robot's proficiency towards a task or a subgoal. For example, Olsen and Goodrich [113] defined productive time as ratio of time spent in autonomous operations to the total time spent across autonomous, manual, and unscheduled manual operations.

Reliability. The duration of time in which the robot meets performance standards under defined working conditions. Some of the commonly used time parameters for this metric include mean time between failures (MTBF) and mean time to failure (MTTF) [78]. In the Fetch scenario, the system time between perception failures is used as a decision-making parameter to reset the vision system.

Forecasting time. The time elapsed between the prediction of a failure and the occurrence of that failure. This metric may be measured at multiple levels of confidence in the form of a prediction time curve, measuring the time elapsed between initial fault detection and the confirmation of that fault [131, 132]. Alternatively, the volume of data required to detect a failure may be tracked as a proxy for time [132].

3.4 RSA: Events

This category of metrics relates to changes in the system status that affect the robot's performance. At the most basic level, counting events can provide insight into how well the robot is performing a task. Examples include, but are not limited to, counts of failures, human interventions, requests by the robot for assistance, corrections, and the amount of required training. These types of measurements are largely applicable to *in situ* and *post hoc* interactions. Additionally, in a human-robot team context the numerical thresholds on these measures for acceptable performance might directly impact other measures of proficiency. For example, in a time critical mission, a lower threshold on the allowed number of replans or reattempts for the robot would result in a higher number of human interventions.

Interventions. External interventions that can affect (positively or negatively) the robot's performance are counted and can be compared to another set of countable instances (e.g., successfully performed tasks without interventions). In a human-robot teaming context, interventions may be initiated by the robot (e.g., in response to expected need for data [44]) as well as by the human (e.g., in response to an impending robot failure or collision [27]). Depending on their nature, these interventions can provide useful insight into proficiency of the robot. Proportions of autonomously-completed and human-assisted steps within a task can also be an indicator of *productivity*. For example, Fitzgerald et. al [44] measures the ratio of supervised and unsupervised steps in an assembly task, where supervision consisted of indicating the object the robot should use in the next step of the task. In a safety critical setting, a higher ratio of human-initiated autonomy switches to total autonomy switches or interventions suggests an overall poorer task performance, as compared to a scenario with a lower ratio.

Repeated attempts. An action performed in response to system or environment anomalies that affects the robot's performance, resulting in repeat attempts of that action. When a robot fails in an operation during a task, it might reattempt the operation or ask a human for assistance. Monitoring the number of reattempts (or the number of failures) can characterize a robot's progress on a task. In the Fetch scenario, the robot failing to pick up a screw after a pre-set number of reattempts could initiate a request for human assistance. For scenarios where direct human assistance is not possible, a robot might use the number of failures as an indicator to abort or reinitialize the operation. Often counts of such failure events may be used in combination with a time metric. Returning to the Fetch scenario, a certain number of detection failures by the perception system, combined with time between failures, could be used as a decision making parameter to reset the camera and vision system. Monitoring the ratio of number of successful task executions vs. total task executions [123] can serve as an indicator of a robot's reliability, which can be used *a priori* to update the robot's self-confidence prior to the next task repetition.

Violation of performance envelopes. The number of times when the robot's performance is not within acceptable bounds in a mission. Performance envelopes [69] can be established based on prior experience. In a human-robot teaming context, frequent violation of performance envelopes can indicate that robot is not proficient according to the performance standard set by the human for a particular environment. Examples of ways to measure whether performance is within acceptable bounds include (1) comparing *in situ* performance estimates to predefined numerical thresholds [9], (2) testing whether perforamnce exceeds history-based aspiration levels from Simon's theory of satisficing behavior [140], (3) using barrier function-based measures to evaluate proximity to constraint violations [37], and (4). detecting the violation of verification and validation assumptions in real-time with a human-in-the-loop [129].

4 METRICS FOR ROBOT COMMUNICATION OF PROFICIENCY (RCP)

Once the robot has assessed its own proficiency, it must turn that assessment into a communication provided to its human partner. The manner in which the robot communicates its proficiency may be influenced by the human's role and/or the robot's perception of the human (i.e., outputs from the RPH stage) and tuned accordingly. These characteristics of proficiency communication are akin to those used to define the properties of explanations: *content* (what is being explained), *communication* (how the system interacts with the user), and *adaptation* (tuning explanation methods to be most effective for the intended user) [91]. In this section, considerations for communication modalities are presented (Section 4.1) followed by metrics for the RCP stage, detailed according to the contents of the communication and capability of the robot to communicate those contents (Section 4.2), the complexity of the communication's contents (Section 4.3), and the efficiency of communication (Section 4.4). All of the metrics described in this section are applicable across all three temporal levels, *a priori*, *in situ*, and *post hoc*. Many of these metrics can be used to evaluate robot communications in other domains; this section specifically contextualizes each metric for proficiency self-assessment. It also should be noted that these metrics do not pertain to a robot's ability to self-assess its proficiency at communicating, which is covered later in the Discussion (Section 7).

4.1 Modalities for Communicating Proficiency

The robot will use one or more modalities to communicate its proficiency to the human. The inherent limitations of the communication modality will affect what kind of proficiency measures can be communicated and what aspects of the communication the human is expected to understand (i.e., inputs to the HUP stage).

A visualization display consisting of graphics, charts, or images displayed on a monitor (e.g., visualizations of input-output relationships of a neural network [172] or heat maps to convey image saliency [22, 65]) can be rich in information and can be transmitted instantaneously, but may require some additional interpretation from the human. These could include text as part of a visualization (e.g., highlighting text that indicates what component of the robot system is being used to perform a task [165]), as a historical log, or sentences via natural language generation (e.g., narrating robot experiences during task execution [126], summarizing explanations based on situation criteria [172]). Natural language generated by the robot can also be conveyed visually and could appear instantaneously while still requiring interpretation from the human (i.e., reading), albeit the format may be more immediately recognizable and can afford increased user satisfaction and confidence with a system [38]. Audio can also be used for conveying natural language to form statements (e.g., communicating faults such as low battery [62]), explanations (e.g., explaining the values, tradeoffs, and competing objectives of planned behaviors [152]), or questions (e.g., the robot asking if the human understood what it communicated [41]). Non-speech audio, such as "auditory icons," or sounds that represent concepts being conveyed [104] may also be used for alerts to get a user's attention. Auditory communication is most appropriate for short, concise messages that require immediate response whereas visual displays might be preferred in cases where the message is complex, must be referred to later, or has a spatial component [29].

Lights positioned on the robot may also be used for alerts (e.g., when encountering a person or obstacle [144]) or for communicating directional information (e.g., which direction the robot is planning to turn [139]). Robot motion and augmented reality (AR) communication methods can utilize a common reference frame of the physical scene wherein parts of the environment can be explicitly referenced [18]. For example, motions performed by the robot may reference itself (e.g., arm movement to express incapability at lifting an object [90], torso movement to correlate with confidence level [154]) and/or the environment (e.g., adjusting motion trajectories to more clearly communicate robot intent for object selection [34, 98], pointing to objects for the human to manipulate [60], gazing towards points of interest [2]). AR projections into the physical environment can also reference the robot (e.g., light projection onto the floor to depict the planned path of travel [15, 138]) or parts of the environment (e.g., projecting a hologram icon onto a tool the robot is planning to use [18]). Virtual reality (VR) methods can display animations of simulated robot movement, such as showing alternatives that could have been performed as the result of a *post hoc* assessment after a failed task. These allow for adjustable perspective that may not be feasible using other modalities (e.g., an immersed operator judging planned robot movements prior to their execution [23]).

Multi-modal interfaces can maximize the richness of communications and optimize human workload. In the context of a swarm of robots, transparent communications of the collective have utilized color-coded visualizations of status combined with robot speech and vibrations for haptic feedback to alleviate workload and increase situation awareness [59]. For human-robot handovers, combinations of robot motion and haptic feedback have been used when a robot holds onto an object longer than expected to convey information to the human partner, which can influence the human's behavior [1]. In a multi-modal interface, the information being communicated may be redundant across modalities (e.g., text-captioned speech [80]) to ensure that the information is received by the human (which has been shown to be effective at resolving human uncertainty in complex tasks [3]), whereas non-redundant information in two separate modalities may contain two different pieces of information (e.g., visual display showing low battery, speech describing possibility of future task success) or a modulation of the information [114] (e.g., speech describing task failure with robot motion to explain which part of the robot failed). An advanced robot system may also choose one modality over another to communicate an aspect of proficiency based on contextual factors and the inherent limitations of single modalities described here.

4.2 RCP: Attributes

The proficiency measures being communicated by the robot will possess one or more possible attributes. For the proficiency self-assessment context, attributes from the explanation research literature are leveraged including those that may be predictive of explanation quality [169], failure types for explaining impossible robot behaviors [120], truthfulness of agents (e.g., transparency vs. deception [33]), and others to distill a common set. These attributes can be used to characterize the contents of a robot communication and will factor into complexity metrics (see next section). See Table 2 for examples of each of the attributes as applied to the Fetch scenario.

Information communicated. The following attributes characterize the type of proficiency information included in the communication from the robot:

- **Alternatives**. These are task strategies that could enable higher proficiency self-assessments to be made either by increasing the likelihood of success, optimizing performance thresholds (e.g., a robot presenting possible grasps it could execute for an operator to select [94]), or enabling otherwise impossible tasks to be performed. Each of the alternatives may also have associated probabilities for success or failure and their communication may possess some of the other attributes in this list. Alternatives presented *a priori* would be predictions of expected failure (e.g., predicting and mitigating potential trust-induced failures by proposing alternatives [160]), *in situ* alternatives could be caused by detecting a failure in progress, and evaluations of proficiency *post hoc* may lead to communicating alternatives for next time.
- **Contextual reference**. Elements in the scene that contribute to the self-assessed measures of proficiency may be referenced within the communication. Reference to elements in the scene may be made relatively (e.g., the object to the robot's left) or absolutely (e.g., the red object); this attribute is referred to as locality in [126] where a robot narrates its experiences. Contextual reference can include characteristics of the environment, objects interacted with, and reference to either agent.
- **Governance**. The robot's self-assessment may be governed via an internal set of rules or policies used to drive the resulting proficiency measures and then referenced in the communication. Explainable AI systems often have these types of mechanisms (e.g., [117]) to provide reasoning for their decision-making to a user. Inclusion of this attribute in a communication may also inform how the human tasks the robot in the future while those same rules and policies are active.
- **Impact**. This attribute notes the effect of anticipated success or failure on subsequent actions or future tasks. Inclusion of this type of information may be based on predicting performance on part of an upcoming task and describing how the robot's proficiency may propagate failure and success throughout the entire task (e.g., if the robot predicts issues with lifting a type of object and the task specification includes lifting several of that object). The robot may have a governing set of rules or policies as described previously that, if adhered to for a given scenario, may violate each other (e.g., the robot cannot follow a person while also avoiding the kitchen if the person walks into the kitchen [121]).
- **Novelty**. The robot may communicate its proficiency based on prior experience with the same task, similar tasks, or that the task is perceived to be novel; this metric attribute captures the amount of novelty. Similarity-based self-assessment techniques (e.g., [7, 58, 74]) utilize these types of comparisons as part of measuring proficiency and could therefore output them as part of the communication.
- **Probability**. Measures of probability will be included in all communications of proficiency, albeit likely at different abstraction levels. These types of measures are akin to expressions of uncertainty (e.g., [20, 163]), self-confidence (e.g., [5, 89]), or strictly as probabilities when used in explanations (e.g., [91]). Probability measures can be conveyed explicitly (e.g., displayed text or spoken words) or implicitly (e.g., visual display characteristics to correlate with uncertainty communication [81, 83, 97]). Conveying measures of uncertainty may impact the credibility of the robot and therefore the human's trust of the robot's proficiency.

Nature of communication. The nature of the proficiency information being communicated (i.e., how it is being communicated) can be characterized using the following attributes:

- **Framing**. The proficiency measures will be framed either as measures of success or failure (i.e., attribute framing as positive or negative proportions [95], or critical or complimentary recommendations [163]). Raman and Kress-Gazit [120] describe three types of failure conditions that can be communicated: unsynthesizable when the robot attempted the task but failed (applicable to *post hoc* communication only), unsatisfiable when an unchangeable condition of the scenario physically prevents the task from being attempted, or unrealizable when the task could be attempted if something was changed. These same types of failures can be inverted for types of success measures when the robot was able to attempt the task attempted the task and succeeded (synthesizable; applicable to *post hoc* communication only) or the conditions of the scenario enable the task to be attempted (satisfiable and realizable; applicable to *a priori* communication only).
- **Truthfulness**. The robot may be truthful in its communications or intentionally acting subversively or deceptively rather than transparently in order to elicit a particular response from the human at the HUP stage of the interaction flow (e.g., giving the human the perception of having control over the robot or that the robot is more capable than it actually is [150]). Deceptive robot motion may also be used to exaggerate, quickly switch, or communicate ambiguous goals [33], which could be the case if the robot is in an adversarial role.
- **Redundancy**. The communication of proficiency may be redundant across modalities when a robot utilizes multi-modal communication. The entire communication may be redundant across modalities (e.g., text-captioned speech) or only certain attributes of the communication may be redundant (e.g., contextual reference to an object in the space may be described via speech while the robot motions towards it). Redundancy in a communication may be used to ensure that the communication is received by the human, but also may lead to cognitive overload [114].
- **Scope**. Scope relates to how much of the task or to what property of the task the proficiency measure refers to. The scope could be the entire task, a subtask, or a requirement of the task/subtask.

The proficiency measures derived from the RSA stage will inform the possible attributes the communication could possess at the RCP stage (e.g., communicating alternatives if a counterfactual self-assessment method is used). These attributes can also be used to inform metrics at the HUP stage whereby evaluations can be performed to determine if the human perceives and comprehends these attributes (which may or may not be required for effective task performance).

Communicability. Some aspects of the proficiency measures derived at the RSA stage may not be communicable by the robot. A lack of communicability could be due to the inherent limitations of its available communication modalities or a lack of functionality to convert proficiency measures into a communicable form. Communicability concerns a robot's overall proficiency-based communication capabilities, comparing the number of unique attributes the robot is able to communicate to a master list of possible attributes. This characterizes the overall robot capability rather than individual instances of communication. For example, if the robot in the Fetch scenario lacks the ability to specify which object is obstructing its path, then proficiency measures that make contextual reference to the object are not communicable. A system may also be able to reason about communicability, conducting feasibility analyses of the limitations of its available modalities [13]. This is an important characterization to make in order to set expectations for measures at the HUP stage and as a guideline when designing a robot's proficiency self-assessment and communication methods.

Group	Attribute	Proficiency communication examples with each attribute	Explanation
Information communicated	Alternatives	The red bin is closer than the green bin, so I have a higher chance of successfully placing the bolts in that bin instead of the green bin.	Presenting the red bin as an alternative.
	Contextual reference	I couldn't grab the bin because the table to my left is obstructing my arm movement, so the task was not completed.	Reference to the table as an environmental factor.
	Governance	I am unable to move beyond the bounds of my designated area, so I can not pick up the dropped screw.	Justifying failure due to internal rule that drives behavior.
	Impact	There is a good chance I will drop a bolt on the floor and I have trouble picking up items off of the floor, so the likelihood of successful task completion will decrease if that happens.	Projection of possible failure onto the rest of the task.
	Novelty	*Novel:* I do not have experience with screws of this size, so there is a good chance I won't be able to grasp it. *Similar:* I dropped the bin because I have had problems with grasping curved objects in the past.	The size of the screw is novel. Considering prior experience using similar objects.
	Probability	There is a good chance I will successfully place the bolts in the bin, but there is only a 25% chance I will succeed in placing the screws in the bin.	Two types of probability measures communicated: "good chaance" and "25%".
Nature of communication	Framing	*Failure, unsatisfiable:* My gripper cannot open wide enough to grasp the bin, so I cannot attempt this task. *Success, synthesizable:* I successfully placed a bin filled with bolts and screws on the specified station.	Maximum gripper width cannot be changed. Task success is communicated post hoc.
	Truthfulness	There are no problems. [when there are in fact low-level problems unlikely to impact performance]	Deceptive statement about task performance status.
	Redundancy	I am not able to grasp this screw. [while the robot simultaneously tilts its head in the direction of the screw]	Speech and motion communication about the same object.
	Scope	*Task:* I successfully completed, filled, and moved the bin as specified. *Requirement:* I was not able to complete the task in the amount of time allotted.	The entire task is described. Exceeding maximum performance time is described.

Table 2. Examples of proficiency communications that possess each of the attributes described in Section 4.2 as applied to the Fetch scenario.

4.3 RCP: Complexity

Several metrics cover the complexity of communicated proficiency and are influenced by the attributes the communication possesses (see prior section) with some modality-specific measures. These qualitative and quantitative metrics are most useful for relative comparisons between communications of proficiency (e.g., one communication is more or less abstract than another) rather than absolute scales.

Abstraction. The level of abstraction refers to the amount of detail in the description that is given about a piece of information, which can be represented in coarse or fine resolution [147]. A measure of abstraction can be applied to each of the previously reviewed communication attributes. For example, in the Fetch scenario, contextual reference to a bolt could be communicated by the robot pointing to it (coarse) or it could simultaneously state via speech that it cannot pick up the bolt by its head (fine). For communicating probability, the robot's confidence in task success could be stated as high/low (coarse) or as a percentage (fine). While no absolute scale exists for information abstraction, some efforts specify a mapping of task characteristics into abstraction levels. For example, Rosenthal et al. [126] describe a mobile robot communicating information about its navigation performance, defining levels of increasing abstraction starting with coordinates of movements (level 1), then to traversal times and distances (level 2), then to right/left turns and straight segments (level 3). Similar scales could be developed based on different task types.

Clutter. Any extraneous information that is included as part of the communication of proficiency that may distract from or obfuscate the core of the communication is considered clutter. Objective metrics to evaluate clutter are specific to the modalities used for the communication. Visual clutter metrics include feature congestion and edge density [125]; Roundtree et al. [127] specify a set of metrics (including visual clutter) for swarm visualizations that were shown to be predictive of human perceived transparency and performance. Visual clutter metrics are also useful when considering the entire scene where activities occur (e.g., for everyday driving [86]). The use of extra words or description used in text and speech modalities can be evaluated by comparing the number of unique words used to the total number of words in the communication [130]. Readability analyses of text or speech based communications can be performed by utilizing readability indices such as the Flesch-Kincaid readability test [47] or the FOG index [57]. These analyses are devoid of context, however, so additional subjective ratings should also be conducted at the HUP stage.

Comprehensiveness. The number of attributes possessed by a communication is indicative of its comprehensiveness. Comprehensiveness can be expressed as a ratio of the number of unique attributes the communication possesses to the number the robot is capable of communicating (i.e., its communicability). At the HUP stage, assessments of comprehensiveness can be conducted to evaluate human understanding of completeness (the quantity of information received compared to the quantity expected [87, 93]).

Simplicity. For the purposes of this paper, simplicity refers to the degree to which there is a "simple" explanation for the proficiency measurement being communicated. This concept is borrowed from evaluation of explanations which are measured by generating a causal model for the explanation and counting the number of root causes in the model [169]. Such a model may be generated by identifying causal language provided in the explanation and then counting the connections between causes and effects [142]. Evaluating the number of root causes has been shown to be predictive of explanation quality [169], but determining the causal pathways for an explanation is somewhat subjective; more formal methods may be needed to validate this metric.

Size. The size of a communication refers to the number of features it contains, which are modality specific. Note that size is not the same as length, which is time-based (see Section 4.4). The size of a text or speech display can be measured by counting the number of words or sentences [126]. Similar quantitative metrics can be applied to robot motions used for communication (either physically or simulated in VR) including the distance traveled by the robot's limbs, the number of repetitions of communicative movements, and the speed of the movements [90]. In Kwon et al. [90], repetition of expressive robot motion at a fast or moderate speed was shown to increase a human's understanding of the robot's goal and made the cause of robot incapability clearer. Static visual displays may lack a discernible set of features to produce a metric for size.

4.4 RCP: Efficiency

The robot will communicate its proficiency to the human over a period of time. The amount of time used to communicate is a function of the length of the message and the speed at which it is communicated.

Communication time. The amount of time required for a message to be (1) generated, (2) transmitted, and (3) understood by the recipient [102] comprises communication time. For the purposes of this paper, these phases are adapted to fit the proficiency-based interaction flow and are delineated into three separate metrics: conversion time, transmission time, and perception/comprehension time. The latter is out of scope for RCP metrics, but in scope for HUP metrics (see Section 5.1). Communication time will be most impacted by the overall complexity of the communication.

A highly transparent robot will communicate more information which may be less efficient than others, producing a transparency-efficiency tradeoff [72]. It is intended that the complexity and efficiency measures provided in this paper could be used to conduct similar analyses for evaluating HRI within the proficiency-based interaction flow.

Conversion time. The first phase is the time required by the robot to convert the measures of proficiency it intends to convey to the human into a communicable form (see Section 3), a common step in facilitating information exchange in human-robot teams [14]. Depending on how the robot's self-assessment functions, measurement of conversion time may be difficult. Concretization is a process used to produce a communicable version of an explanation generated by a system [146]. If an explicit process like this is used by the robot, then conversion time could be measured and reported by the robot.

Communication latency. Latency is the amount of time between when a message is sent and when it is received; it is a common metric for evaluating HRI communications [151]. As an efficiency metric at the RCP stage, communication latency can either be measured as the amount of time between the RSA stage and the transmission of the communicated proficiency (see next metric) or as the amount of time between conversion and transmission. Unless the robot explicitly communicates when it is in each phase, latency measures between conversion and transmission may be difficult to measure. Latency between transmission and perception/comprehension time is covered in Section 5.2. High communication latency times may also impact HUP metrics, particularly if the human is expected to react in a timely manner to the communication, such as to intervene to improve the robot's proficiency measures.

Transmission time. The second phase refers to the length of time from the moment the robot starts communicating its proficiency until it stops. This can be measured according to the observable start and end time of communication (e.g., the length of a sentences spoken by or movements performed by the robot). If the robot produces speech, natural language measures including the number of words [126], conciseness [19, 146, 168], and talking speed will impact this metric. A concise communication can be further qualified as one that is considered complete, but minimally represented [84]. Transmission time for communication via physical or simulated robot motion in VR will be impacted by movement speed, distance traveled by the robot's limbs, and the number of repetitions of movements [90]. Kwon et al. [90] used those measures to derive a cost metric when determining effective methods for communicating incapability with robot motions. Robot movements may have to be communicated at a particular speed if it pertains to the proficiency measure. For static visual displays on a screen or via AR, transmission time may be (near) instantaneous.

5 METRICS FOR HUMAN UNDERSTANDING OF PROFICIENCY (HUP)

Communicating proficiency is only effective if the intended receiver correctly understands the message. Therefore, part of examining a robot's ability to communicate proficiency is analyzing how well a human can interpret and understand the robot's communication, thereby revealing how effective the robot was in conveying its proficiency to the human.

Unlike metrics at the RCP stage of the proficiency-based interaction flow (Section 4), which measure the communication in isolation, metrics related to human understanding necessarily involve the human. In particular, the metrics in this section focus on measuring downstream effects of the robot's communication by measuring its impact on the human. These metrics inherently account for context such as the human's prior knowledge, the complexity of the task, or any limitations such as time pressure. These measurements can only be conducted after the communication has been delivered, as they require measuring the effects of the communication on the human, so they do not function *a priori*.

Some of the metrics in this section can be measured explicitly, while others are measured implicitly. Explicit measurements directly test the concept at hand, often by querying the human to gauge their understanding. These kinds of measurements typically happen outside the flow of an interaction, so the interaction must pause while the human responds to an explicit measurement request. Explicit measures can test objective information (e.g., a knowledge test in which a human reports their understanding of the robot's communication), but it can often be useful to assess the human's subjective experience with explicit measures as well. Adding a measure of the user's confidence via a Likert or continuous scale, for example, can help to assess perceived certainty, which can be used in collaboration with knowledge questions to tune future robot communications. In contrast to explicit measurements, implicit measurements are based on what is already occurring in an interaction, so they do not require additional action outside of the typical interaction flow. While explicit measures allow for more precise measurements of specific concepts, implicit measurements are less disruptive. Which type of metric to use depends on the situation.

In this work, human understanding of a robot's proficiency is grouped into three levels, inspired by the three levels of situation awareness [40]. Each level represents an increasingly deep human understanding, so the levels are sequential, not categorical:

(1) Perception: how well the communication is received by the human
(2) Comprehension: how accurately the human extracted meaning from the communication
(3) Projection: how well the human can apply the effect of this communication for future coordination

There is an inherent downstream effect of the metrics from one level to another. However, it may not be necessary for the human to perceive or comprehend every attribute of the proficiency communication in order for the interaction between human and robot to be effective (i.e., effective projection). For example, the human may perceive the proficiency communication and not comprehend all of the attributes it possesses, but still initiate an appropriate change of action or plan. The following metrics can be used to measure human understanding of the entirety of the robot's proficiency communication or could instead be modality specific (e.g., the human may fully comprehend what the robot said, but did not accurately perceive the robot's movement). The metrics apply regardless if the proficiency-based interaction occurs *a priori*, *in situ*, or *post hoc* to a task.

5.1 HUP: Perception

The first level of human understanding involves perceiving the communication. This is influenced by factors under the robot's control, such as the volume used by the robot or the visibility of its movement to the human receiver. It is also influenced by factors outside of the robot's control, such as ambient noise or environmental occlusions. Finally, it can be influenced by cognitive factors of the receiver, such as distractedness or a lack of shared context with the robot. In the Fetch scenario, the robot might communicate its proficiency using speech, such as "I am unlikely to succeed at retrieving the large screw." Accurate perception of this statement would mean hearing all of the words correctly. If a visual communication method was used, accurate perception would require seeing the communication.

Perception clarity. The message should be perceived clearly by the human partner. Perception clarity can be measured by evaluating the accuracy of the received message; less accuracy means less clarity. For text- or speech-based communication methods, accuracy can be measured by using a read-back method in which the human repeats back the communication provided by the robot, measuring the deviation between the human's read-back and the robot's original message. This method is similar to the read-back/hear-back method employed by airline pilots and air traffic controllers [141]. For nonverbal communication such as robot motion, the human could repeat the motion instead (e.g., mimicking a hand signal). Another method for measuring perception clarity is to count the number of *follow up* queries made by the human that ask the robot to repeat the communication or clarify what attributes were included in the communication (called "explainee return questions" in Maduma et al.[99]). For example, using the Fetch scenario, the human may inquire "Did you say 'small' or 'large' screws?" This *follow up* method is also used in subsequent HUP levels to evaluate comprehension clarity (Section 5.2) and projection clarity (Section 5.3). Characterizing the type of *follow up* (whether for perception, comprehension, or projection clarity) may be difficult to objectively discern and is subject to interpretation. Follow up communications from the human to the robot can be tracked at the RPH stage as a count of the number of communication events that occur (see Section 6.4).

Perception completeness. Measures of perception completeness compare the quantity of the information that is received compared to the quantity that is expected [87, 93]). At the perception level, this is a measure of how much signal is received by the human. For example, in the Fetch scenario, the human might only receive the first five words of the robot's utterance ("I am unlikely to succeed") and therefore wouldn't have enough information to accurately perceive the communication. A human may or may not be aware that they have incompletely received a message. The amount of completeness could be assessed through explicit questions.

Perception time. This metric refers to the amount of time required for a human to receive the proficiency communication. Measures of communication complexity including clutter, comprehensiveness, and size from RCP will impact the perception time. For example, larger messages may take longer to receive due to longer transmission and reading times. Perception time can be measured using an end-of-message response, such as the start of a read-back or a simple acknowledgment of receipt. Perception time is not exclusively a feature of the communication, as it can be affected by the partner's current cognitive processing capacity.

5.2 HUP: Comprehension

After perception, the robot's proficiency communication must be processed and interpreted to extract task-relevant meaning. This may involve combining the current communication with prior ones to provide complete information. It may also involve resolving ambiguity in the communication, perhaps by using context, prior knowledge, or prior communications. For example, in the Fetch scenario, the robot can communicate its proficiency by saying "I am unlikely to succeed at retrieving the large screw." While there might be several screws on the table, the adjective "large" enables a human listener to resolve the potentially ambiguous reference, increasing the likelihood of successful comprehension. The robot might also point to the screw, multimodally communicating redundant information to increase comprehension.

Given that the levels of human understanding at the HUP stage are sequential, all metrics for comprehension will be impacted by those for perception (i.e., the human cannot comprehend that which they do not perceive). Even with perfect perception, comprehension can be affected by a variety of factors. Comprehension issues stem from a mismatch between how the robot has formulated its communication and how the human interprets it. This mismatch can arise from vocabulary-based ambiguity in the communication, such as using unfamiliar words, words with multiple meanings, or unclear referents like "that one" (i.e., in these examples, varying abstraction of the contextual reference attribute of the communication). Errors in comprehension can also arise when the robot incorrectly assumes the human has contextual information or prior knowledge that, when missing, affects the comprehension of the communication. In the example above, if the human only knew about small screws, they might incorrectly comprehend the robot's statement to be about the small screws.

Comprehension clarity. The clarity of the proficiency communication at this level of HUP refers to the accuracy of human understanding. Assessments at this level largely consist of subjective ratings from the human and can be evaluated along several parameters that will impact overall comprehension clarity. The parameters deal with several factors of the interaction including comprehension of the properties of the communication, human expectations, awareness of goals, and the internal machinations of how proficiency measures were derived. Parameters of comprehension clarity include:

- **Coherence**. External coherence refers to whether the communication overlaps or fits with the human's existing mental models and internal coherence refers to how well the parts of a communication fit together with each other [169]. Each of the attributes possessed by a proficiency communication should be internally coherent with each other to increase comprehension.
- **Comprehensibility**. A comprehensible communication is one that is presented in a human understandable fashion, meaning in terms, structure, and semantics similar to that a human expert would produce if provided the same input (adapted from [42, 105]).
- **Explicability**. Evaluations of explicability compare what the robot communicated regarding its proficiency to what the human had expected it to. This is adapted from explainable systems' research that typically refers to an agent communicating its plan [17], so is most relevant for *a priori* and *in situ* communications. The use of explanatory actions during planning have been used to increase explicability of a robot's actions [145].
- **Fluency**. The linguistic quality of a communication—sometimes referred to as **naturalness**—typically assessed of automatically generated text or speech communications [52]. Evaluating this metric (and others related to natural language generation) as a relative comparison of communications rather than absolutely has found to greatly improve consistency of the measure [112]. It should be noted that for the HUP stage of the proficiency-based interaction flow, fluency does not refer to the evaluation of collaborative activities between a human and a robot (as in [67]).
- **Informativeness**. This parameter is a measurement of how much informational content a communication possesses [52]. This can be evaluated in reference to the attributes the communication possesses, i.e., if some attributes are informative or if they are extraneous.
- **Intelligibility**. Evaluations of intelligibility measure the human's understanding of how the robot's model functions [8]. This is a similar goal of some transparent robot systems, such as those that aim to communicate the cause of incapability to the human [82]. This metric could also be evaluated objectively in comparison to how the robot actually functions.
- **Legibility**. The human's comprehension of a proficiency communication is considered legible when the goal of the communication is not known (adapted from [17, 32]). This is sometimes also called readability or transparency. In the context of this paper, the goal of a communication refers to what attributes of the robot's proficiency it is attempting to communicate (e.g., contextual reference; see Section 4.3 for other attributes). This metric should only be evaluated in scenarios where the human is not aware of the robot's communication goal, such as *a priori* to a task being performed or *post hoc* if the human did not observe any of the robot's previous activities (i.e., scenarios wherein the human is less aware of what the robot might communicate). The legibility rating scale introduced in Dragan et al. [32] can be used for robot motion communication modalities and adapted for others.
- **Predictability**. If the human knows the goal the robot is attempting to communicate, predictability can be evaluated as to how the proficiency communication aligns with the human's expectations (adapted from [17, 34]). The human may be able to predict what the robot's proficiency communication will be *post hoc* to a task if they observed the robot's previous activities. If the human has prior, repeated experience with the robot, predictability of the proficiency communication could be evaluated at any temporal level. A predictability scale is provided in

Parameter	Examples with high comprehension clarity	Explanation
Coherence	I cannot grasp the large screw, so I cannot place it in the gray bin.	Describes a single, coherent sequence of actions.
Comprehensibility	My arm cannot reach to the large screw from here.	Uses human-understandable terms.
Explicability	You requested that I move the screw, but I am not able to grasp it.	Describes ability with respect to human expectation.
Fluency	I'm sorry, I'm not able to reach the screw.	Uses natural, human-like speech.
Informativeness	I'm not able to grasp the screw because it is one meter beyond the reach of my left arm.	Contains high informational content.
Intelligibility	I cannot reach the screw because I do not want to risk tipping over.	Describes additional considerations by the robot while manipulating.
Legibility	[Robot reaches unsuccessfully toward screw that's out of range]	The robot's action clearly indicates the goal.
Predictability	[Robot picks up bin after it has been filled with the appropriate bolts and screws]	The action is unsurprising, given human expectation.

Table 3. Comprehension clarity can have several parameters, each of which reflect an element of human understanding of proficiency. Using the Fetch scenario, we provide examples of communications with high comprehension clarity for each parameter as applied to when the robot has been asked to reach for a screw that is outside its workspace.

Dragan and Srinivasa [35] that—similar to the legibility rating scale—is designed for subjective rating of robot motion, but can be adapted for other communication modalities.

Comprehension clarity can be objectively evaluated by counting the number of *follow up* queries made by the human to the robot in order to clarify the meaning of a proficiency communication. These queries can be characterized according to the proficiency attributes they refer to (see Section 4.3). For example, using the Fetch scenario, a *post hoc* inquiry relevant to comprehension would be asking "Is this the type of screw you had an issue with?" in order to clarify the contextual reference attribute of the proficiency communication. Comprehension clarity can also be identified through backchanneling actions, such as head nods, that indicate that the human understands what the robot is trying to communicate [64].

Comprehension completeness. At the comprehension level, completeness refers to the quantity of information understood by the human compared to the quantity that is expected (adapted from [87, 93]). For example, the human may query the robot in the Fetch scenario *in situ* and ask "Which object are you trying to grab?" and will expect the robot's responding communication to include contextual reference to the object. This is similar to explicability, which measures how close a robot's plan aligns with the human's expectations [17]. This definition can be adapted to refer to proficiency communications regarding future events that occur *a priori* or *in situ* to a task.

Communication consistency. Consistent communication refers to information presentation over time that has repeatable formatting and is compatible with prior information [84]. In the context of proficiency communication, consistency can be subjectively rated by the human or objectively evaluated along several axes according to the properties of the communication produced at the RCP stage (e.g., the communication modality used, transmission time, and the attributes it possesses). Consistency is similar to predictability in that evaluating it assumes the human is expecting the robot to communicate its proficiency in a certain way. In some explainability research, consistency is referred to as precision; a precise communication is one that the human believes to be sufficient and not surprising in relation to prior explanations [168].

Content quality. The quality of the content provided in the communication can be evaluated in terms of goodness or satisfaction, two terms that are used frequently throughout explainability research. Subjective evaluation scales can be used to assess these aspects of comprehension, such as the checklists provided in Hoffman et al. [70] where the human rates their level of agreement with various statements related to understanding, detail sufficiency, completeness, and trust. Several of the previously mentioned metrics have been shown to be predictive of subjective ratings of explanation quality including the possession of attributes like alternatives, completeness, and internal coherence [169]. Evaluations of this metric are impacted by the clarity metrics previously mentioned in that evaluations of poor content quality may be influenced by poor clarity.

Processing difficulty. Depending on factors like the communication's clarity, quality, and timing, the human may face challenges in processing the conveyed information while attempting to comprehend it. Broadly used measures for evaluating cognitive workload in human-robot interaction can be utilized, including subjective ratings like the NASA-TLX scale [61] or objective evaluations such as the introduction of a secondary task [24] and biometric measures like the use of functional near-infrared spectroscopy (fNIRS) to classify workload of auditory or visual processing [118]. Difficulty in processing may also be evidenced by additional *follow up* inquiries made by the human. These measures may also be used at the RPH stage as inputs to the robot's perception of the human.

Comprehension time. This is the amount of time required for a human to make sense of a communication. Perception time (Section 5.1) plus comprehension time is the total amount of time that a human needs to understand a communication from a robot. In our conception, comprehension time starts when the human has fully received the message, and ends when they fully understand the message's content. In reality, perception and comprehension time often overlap, because people make sense of a communication as it arrives, rather than waiting for a complete communication to arrive before beginning to interpret it. The use of anticipatory motions prior to robot actions has been shown to increase reaction times when participants were tasked with labelling robot actions [54]. More complex communications will require more time to interpret, because it will require more cognitive processing. Verbal messages may need to be repeated back, textual statements may be read more than once, and visual displays may be inspected repeatedly by the human before the message is fully interpreted.

5.3 HUP: Projection

When interacting with a robot communicating its proficiency, projection refers to the human's ability to use the proficiency communication to plan future actions and interactions between the robot and the human. In the Fetch scenario, if the robot has informed the human that it is unlikely to succeed at picking up the large screw, successful projection would involve adapting future actions to incorporate this new information. For example, the human might decide to pick up and deliver the screw themselves, rather than depending on the robot to do it. Alternately, the human might adapt the robot's environment to make task completion more likely, such as by reorienting the screw to a position that is easier for the robot to grasp.

Even if the prior levels of HUP (perception and comprehension) are successful, a human partner can still make errors in projection if they have an inaccurate understanding of the task or the robot's abilities. For example, the human might *think* that setting the screw upright would improve the likelihood of the robot completing the screw retrieval task, when in fact it has no bearing on the robot's proficiency because the screw is just too big to fit in its gripper. Projection depends on the human's model of *why* the robot's self-assessed proficiency is low, and if this model is wrong, then the human's projection will be ultimately incorrect. Having a robot provide its proficiency assessments at higher levels like *explanation* and *prediction*, described in Section 2.1, can help mitigate some of these errors.

As with perception and comprehension, projection can be measured explicitly and implicitly. Explicit measures might involve questionnaires for the human that examine their expectations of future robot behavior. Implicit measures include observations of what intervention a human chooses after receiving the robot's communication (e.g., moving a problematic object, instructing the robot to perform the task differently) and whether their plans appear to change based on that communication. Implicit measures assume a model of the human's reasoning (e.g., if they predict X, they will do Y). Mistakes in this human reasoning model will lead to inaccurate measurements of projection.

Expectations. This metric addresses how accurately a person can identify future outcomes. A robot's communication of proficiency can influence this metric by helping improve (or harm) the accuracy of a person's expectations. To measure expectations, people can be directly queried about their projection in the task (e.g., by asking "what do you think the robot will do next?"). For example, in a study of multi-robot operation, researchers had participants operate several virtual UAVs simultaneously [124]. At two points during the task, they froze the simulation and asked participants to report which direction the robots were going to go next. This is an example of the situation awareness global assessment (SAGAT) technique [39], initially proposed for aviation.

Projection clarity. This metric refers to how easily a human is able to project into the future given what the robot has told them about its proficiency. It does not necessarily require the robot's proficiency communication to be accurate, just that the human can use it to identify future outcomes. As such, an inaccurate robot proficiency assessment may lead to an overall inaccurate expectation (the previous metric), but high projection *clarity* if the human is able to make a confident (though ultimately incorrect) projection. In one study that measured projection clarity, researchers designed a game in which a robot took actions that conveyed its (hidden) skills, and a human then selected additional actions for the robot based on their understanding of the robot's skills [128]. Projection clarity was reported as the number of participants who had greater than "neutral" ratings of confidence in their projections of the robot's skills after the game. Projection clarity can also be measured similarly to perception and comprehension clarity by objectively evaluating the number of *follow up* queries made by the human to the robot asking to clarify information about the future state of the task. In the Fetch scenario, the human may ask the robot, "If I adjust the orientation of the screw, will that help?"

Congruity. Human-robot congruity is a measure of how aligned human and robot mental models are. It is most effectively measured with respect to particular features of the agents' mental models, such as their beliefs, goals, or expectations of rewards. For projection, congruity measures the similarity between human and robot expectations about future outcomes. For example, researchers have used techniques like cross training to ensure that humans and robots have similar mental models of tasks, and then have measured the congruity of those mental models by calculating the decreasing uncertainty the robot has about observed human actions during training [111]. Note that this does not rely on either agent being correct in their mental model; congruity measures alignment, not accuracy.

Command changes. Command changes are human actions taken to augment the robot's behavior based on the human's projection level of understanding. The goal of initiating command changes is to remediate the situation and increase the robot's future proficiency. This metric is reminiscent of the metric for human interventions defined for semi-autonomous operation in other HRI scenarios [151]. Command changes to robot behavior can initiate changes in robot action or the robot's goal. Using the Fetch scenario, a change in action would be the human adjusting the order in which the robot fills in the bin (maybe starting with screws first if the screw placement obstructs the bolts), where a change in goal would be for the robot to deliver empty bins to the workstation and the human decides to gather the screws and bolts on their own. If changes occur, they can be tracked at the RPH stage as a number of correction or intervention events (see Section 6.4).

Environment changes. Changes to the environment are human actions in response to a proficiency communication that attempt to correct the situation by modifying conditions of the environment. For example, the robot in the Fetch scenario may not be able to deliver a bin to the workstation if the workstation already has a bin on it, so the human may remove that bin to modify the scene. Similar to command changes, if the robot communicates high proficiency, then a lack of environment changes may be indicative of proper projection. These changes can be tracked at the RPH stage as a count of the number of modification events (see Section 6.4).

6 METRICS FOR ROBOT PERCEPTION OF HUMAN INTENTIONS, VALUES, AND ASSESSMENTS (RPH)

This paper's rhetorical framing assumes the robot is part of a human-agent interaction dyad. This subsection further assumes that the human problem holder has an internal *intention* as well as internal *values, states and beliefs*. Intention is usually defined as a mental state that represents a commitment to a plan or activity that an agent believes will bring about a desired objective [30, 53, 100]. Examples of human intention in human-robot teaming include commitment to a planned sequence of events to accomplish a mission, a specific desire about the way a task is performed, beliefs about the way that an interaction partner will act, and beliefs about the timing or duration of events or mission stages. Through the remainder of this section, intentions, values, states, and beliefs are rhetorically referred to simply as *intentions*.

The RPH problem is for a robot to *align* its models or algorithms with a human's intention. The robot can explicitly *model* or represent human intention, or the robot might implicitly *assume* things about intention in its algorithms and behaviors. When an explicit model is used, alignment means that the robot's model is a valid representation of the human intention, and when implicit assumptions are relevant, alignment means that the robot's activities honor the human's intention. Thus, this section identifies metrics of robot-intention alignment.

Both the RSA and RPH stages measure how well a particular goal is met or how well intent is satisfied, which is the essence of proficiency assessment. Thus, this section uses four metric categories that are nearly identical to those found in Section 3: uncertainty and coherence, performance, time, and events. RPH also includes a distinct metric category, colloquially referred to as *human factors*, because RPH metrics are inherently interactive between robot and human. Each metric category is given its own subsection. Each subsection emphasizes *in situ* assessment, with applications to *a priori* and *post hoc* assessment provided where appropriate.

Human activity is essential to many of the metrics in this section. Relevant human activities include both intentional messages sent from the human to the robot (e.g., gestures, speech acts, interface commands) as well as implicit signals (e.g., posture, eye gaze, interface activity, task-related behavior). Additionally, most of the metrics require information about expected behavior or outcomes, which must be obtained in a "calibration" phase that uses experimentation and prior experiences to establish bounds and expectations.

6.1 RPH: Uncertainty and Coherence

Metrics for uncertainty and coherence evaluate how well model outputs, behaviors, and model representations are "working" in the context of a human-robot interaction.

Model consistency. Model consistency means that the predictions or estimates do not change too quickly or too slowly relative to the expectations associated with the task. Explicit models can output not only estimates of human goals, plans, and intentions [25, 106, 109, 143], but also predictions about human activities or behaviors [63, 96].

Model consistency metrics evaluate the rates that these estimates and predictions change, and compare those rates to expectations. Explicit models often include internal states such as beliefs about processes, representations about likely objects in the world, etc. If internal states change too quickly or too slowly relative to expectations, then model misalignment might be occurring.

Behavior consistency. Behavior consistency means that the robot's actions (e.g., effector activity, speech acts) induced either by explicit models or implicit assumptions do not change too quickly or too slowly relative to expectations. Both overly consistent behaviors (never changing behaviors even when the world changes) or rapidly changing behaviors (thrashing through a set of actions) can indicate that the robot's behaviors do not align with the human's intent, because the robot is not responsive enough to human activity or not consistent enough to coordinate with intended interactions. For example, (1) reactive or rapidly changing responses by a robot in a shared workspace decrease fluency in human-robot teaming [68], (2) unpredictable or illegible drone behavior can induce fear or stress in human-drone interaction [156], and (3) compliance with social or proxemic norms has a strong impact on acceptance of robots in social interactions [155].

The human also generates action or communication behaviors. The consistency of human communication can indicate how well the robot's models align with the human's intention. For example, (1) misalignment can correlate with changes in sentiment across speech acts or chats [79], (2) disfluence [110] can indicate misalignment under natural language collaboration, (3) changing spatio-temporal references in shared work spaces can indicate misalignment [48], and (4) inconsistent deictic gestures can indicate misalignment in gesture-based collaboration.

Similarly, the consistency of human actions can also indicate the quality of alignment. For example, (1) operator-induced oscillations in shared or supervisory control can indicate failures by the robot to support intended human action [28, 115], (2) spatial distancing can indicate that a human is compensating for proxemic misalignment [107], (3) extraneous commands can indicate misalignment between human intent and perceived robot in learning by demonstration [85], and (4) violations of the neglect benevolence principle can indicate misalignment between human intent and perceived swarm behavior in human-swarm interaction [108].

Model uncertainty. Uncertainty means that there is variability of estimates, predictions, or internal representations of a robot. High levels of uncertainty can indicate misalignment between the robot's models and the human's intent either because the intent is unclear to the robot or the robot does not have internal representations that allow it to properly model the intent. Uncertainty is a well-known metric category and, therefore, a short list of relevant metrics suffices for this paper. Uncertainty can be quantified using conventional methods like variance, confidence intervals, correlation quality, and the probability of outliers. Uncertainty can also be quantified using aggregation techniques such as behavioral entropy [10, 55, 166].

Model prediction accuracy. Prediction accuracy is the straightforward evaluation of how well a robot's model corresponds to actual human activity when explicit behavioral predictions are made. Like uncertainty, accuracy is a well-known metric category so a short list of metrics suffices. Prediction accuracy can be measured using conventional techniques such as the percentage of correct predictions, precision and recall, F1 scores, or ROC curves. Evaluation can be performed and data gathered *post hoc* (either after a mission is complete or between repeated tasks) to assess the alignment of the robot's models of human intentions and human intentions. *Post hoc* analysis might identify overconfidence of *in situ* estimates or inconsistent performance across repeated tasks. When combined with data-driven approaches for refining robot behavior, indirect measures like convergence of inverse reinforcement learning models or persistence of learned reward structures can indicate alignment.

6.2 RPH: Performance

Proficiency should correlate with alignment of robot behaviors and human intent. Measures of performance can be used to provide *in situ* or *post hoc* information about the alignment of the robot's models and the human's intent when prior experience can be used to establish expectations about the values of the performance criteria. Simply put, low performance probably indicates that the robot did not contribute to solving the problem held by the human. Performance measures are not applicable *a priori* since performance measures depend on activity and outcomes.

For example, as mentioned in the RSA: Performance metrics (see Section 3.2), cumulative scores, rewards, and penalties can be assessed *post hoc* by the robot to determine how well a task was performed. Poor *post hoc* scores can indicate that the robot did not successfully accomplish or honor the human's intention. To this end, the RSA: Performance and Events metrics are applicable at the RPH stage, so are not reiterated here. Similar metrics of **mission progress**, **replanning triggers**, and **violation of performance envelopes** provide information about the alignment between the robot's models and the human's intent if the human has explicitly stated a set of expectations about progress. Additionally, subjective scoring by the problem holder at the HUP stage (see Section 5) or by a set of experts using appropriate behavioral coding standards can be used to compute performance. Techniques like automatically generated performance summaries [135] facilitate *post hoc* subjective scoring. Low subjective scoring can indicate misalignment between robot behavior and human intent.

6.3 RPH: Time

Metrics related to time deal with the duration of phases of autonomy, time elapsed between signals of interest, or the perseveration of an activity. Time-based metrics are simple in the sense that they do not need access to predictions from the robot's model nor access to the states in the robot's internal model.

Persistence. Persistence-based metrics reflect the amount of time spent in a particular activity. Robot-oriented metrics include the amount of time the robot spends in a mission phase [134], at a particular location, or in a particular sub-task. Long periods of perseveration or very brief periods in an activity can indicate misalignment if the human assumes regular progress toward a goal [134]. Human-oriented metrics can include the amount of time a human fixates on a location in space [16, 148, 149] or menu of an interface [66], or works on the same sub-task. These metrics can indicate that the human is trying to figure out what the robot is doing, which indicates misalignment.

Expected execution time. The deviation between the robot's expected task completion time and the human's is a straightforward indication of how well robot behaviors align with human intent.

Coordination time. When tasks or missions can be repeated, the time required before or after a mission is executed can indicate the quality of alignment. A long debriefing time can indicate that there is misalignment between what the robot did at various phases of the mission and what the human thought was needed. Similarly, a long prebriefing time can indicate *a priori* a mismatch between the needs of the problem holder and the likely ability of the robot to be able to satisfy those needs without significant human input.

Event timing. The time elapsed between expected or planned events can indicate how well the robot's behaviors align with human intent. Relevant events include communication acts, activities by either human or robot, subtask completion by either human or robot, or changes in performance indicators. Example metrics include (1) the frequency of interactions, (2) the pacing of communications such as *mean time between connection events* [122], (3) the time at which mileposts are reached or the order with which mileposts are reached [134], (4) the rate of change of instantaneous performance indicators [24], and (5) dwell time of eye gaze in interface-mediated or social interactions [2].

6.4 RPH: Events

Section 3 on RSA metrics notes that *events* can be defined in various ways and that counting events can be useful in understanding how well a robot is performing. In the context of a robot perceiving the human, relevant events can include actions initiated by the human to affect the robot's behavior or actions initiated by the robot to gather information from the human. The metric types and metrics in this subsection are closely related to the definition of *clarity* given in Section 5.2; they are based on the assumption that a lack of clarity to a human manifests itself to the robot in various forms, with each form providing information about the alignment of robot behavior and human intention. Proper alignment may also be evidenced by a lack of interaction from the human if the robot communicates a high proficiency measure. For *post hoc* evaluations, events may be counted when debriefing and may include counting instances of overuse or disuse [92], as well as measures of transparency [21] or awareness violations [36].

Corrections. Corrections are human activities that repair or modify results from the robot's behavior. Counting the number of corrections directly measures misalignment between robot models and human intent. When the relationship between the robot and the human follows a supervisory control mode, the number of *undo* commands also indicates misalignment. When the relationship between human and robot is more collaborative, the number of times the human undoes the robot's work or redoes it in a different way indicate misalignment.

Interventions. Interventions are human activities that seek to alter the future behaviors of the robot, especially if those future behaviors are likely to produce poor results (i.e., command changes that occur at the HUP stage; see Section 5.3). Interventions can include changing the robot's actions, altering the robot's goals, or changing autonomy modes, such as the human stopping a robot to do the task without the robot or changing to a mode where the human has more control. Interventions can also include demonstrating tasks or giving negative or positive feedback [159]. Counting interventions is a metric that can indicate misalignment.

Modifications. Modifications are human activities that shape the environment or task setup (i.e., environment changes that occur at the HUP stage; see Section 5.3). Modifications can include task shaping such as cleaning up a room before a Roomba vaccuums [49], placing objects within the reach of the robot's gripper, adding RFID tags to important objects in the world [88], or changing lighting conditions to improve object recognition or mapping. Modifications can also include explicit adaptations that a human makes to a robot's model, like labeling objects or tuning parameters to alter performance. Modifications are representative of a person's understanding of the robot's abilities. As such, when the dyad is in alignment, the human is making modifications that assist the robot – but when there is misalignment, modifications may not help the task completion by the robot. Due to this, both types of modifications should be reported.

Communications. Communication includes the number of speech acts, instant messages, interface requests, or queries initiated either by the robot or the human. This number depends on context because in some applications a lot of communication indicates that the robot understands the human's intent (e.g., the team is rapidly communicating needed information in a timely way to enable task progress) and in others a lot of communication indicates poor understanding of the human's intent (e.g., repeated requests for information or repeated instructions from a human). Communication can be as simple as clicks on an interface or gestures with the hand or mouse, or as sophisticated as natural language communication, eye gaze [2], deictic gesturing [103, 116], or social cueing [11]. Any follow up inquiries made by the human at the HUP stage (see Section 5) can be counted towards this metric.

6.5 RPH: Human Factors

The alignment between the robot's models and the human's intent will impact the problem holder. Thus, there are a number of human factors measures that can provide information about the human's experience throughout the phases of interaction. Information about the human's experience reveals information about alignment. Since human factors metrics are well documented, relevant measures are simply listed. These measures include human workload (such as NASA TLX [61]), physiological indicators of workload or stress [158], real-time estimates of trust [31], real-time estimates of situation awareness [101, 170], and violations of human performance limitations [67]. High levels of workload, mistrust, stress, or frustration can indicate misalignment of robot behavior with human intentions.

7 DISCUSSION

While many of the metrics reviewed in this paper can already be used to evaluate proficiency-based human-robot interaction (HRI), there are several factors that remain as open research areas and other evaluation considerations.

Maturity of metrics and measuring across stages. The metrics presented in this paper each vary in terms of maturity, robustness, and prominence in the field of proficiency self-assessment (PSA) and HRI. For instance, the metrics at the RSA stage (Section 3) are applicable to many types of autonomous systems as all deal with some level of decision-making and uncertainty, but the method of calculation and how the resulting measure is utilized by the robot can vary (e.g., tracking violation of performance envelopes as a safety measure vs. measuring risk-averse reward to inform a PSA measure for future reference). Metrics at the RCP and HUP stages (Section 4 and 5) can also be more broadly applicable outside of PSA and have relevance to multiple domains of HRI. RPH metrics (Section 6) are those most explicitly applicable to PSA, but are informed by broader and more common elements of HRI evaluation, including measurement of interventions and the various human factors evaluation tools for trust, situation awareness, etc.

However, the connections and correlations between metrics at each stage of the proficiency-based interaction flow (see Figure 1) is still an open research area for continued investigation. Some of those relationships are described previously throughout the paper, but metrics that are representative of the entire interaction may also be warranted. For example, the Fetch scenario takes place in a manufacturing environment where overall task efficiency may be paramount to maintaining expected throughput. Evaluations of efficiency will rely on metrics evaluated at multiple stages including productive time (RSA), communication time (RCP), and comprehension time (HUP). Performance evaluation in terms of task efficacy may also rely on metrics from the RSA stage (e.g., repeated attempts), HUP stage (e.g., command changes), and RPH stage (e.g., interventions), as each action could cause unacceptable delays to task completion. While much of this is context-dependent, the metrics framework presented in this paper should enable these investigations to occur using a common lexicon and metrics.

Empirical and/or model-based evaluations. Using the metrics reviewed in this paper to evaluate human-robot interaction—particularly those metrics at the HUP stage (Section 5)—can be done empirically, in comparison to a model, or both. Empirical measures of human understanding are drawn directly from observations of human behavior. For example, whether a human decides to give control to a robot or keep it for themselves is driven (at least in part) by that human's understanding of the robot's proficiency, so observing the human's decisions can allow us to measure their understanding of the robot's proficiency. These measurements can be made after a human takes action, such as initiating follow up queries, command changes, or environment changes. Empirical measurements can also be used to update hypothetical human models to make future model-based measurements more accurate.

Model-based measurements are made using a hypothetical model of a human, rather than observations of actual human behavior. Human mental models may be formulated *structurally* based on how they work or *functionally* based on what they were doing [76]. Measures like legibility [34] use a hypothetical human mental model that includes awareness of the robot's goal. These models are often validated against empirical measurements in lab-based studies, but model-based measurements are useful because they can be taken at any point in the interaction and do not require explicit actions from the human. The robot may actively use such a model of the human to influence how it self-assesses its proficiency, or build up such a model based on the human-robot interaction.

Subjective interpretation of metrics. Some of the metrics reviewed in this paper may be subject to interpretation due to a lack of formal evaluation methods or definition of specific criteria. For instance, at the RCP stage (Section 4), identification of whether a proficiency communication possesses one or more attributes may require communication instances to be coded by multiple raters, reaching a specified Cohen's Kappa measure for inter-rater reliability. While some automated methods are used for evaluating generated language output [52], other communication modalities lack such methods for validation.

Similar determinations must be made at the HUP stage in order to characterize any follow up queries made by the human in terms of the level of human understanding (perception, comprehension, or projection) they correspond to; i.e., such queries can be used as part of a clarity measurement at each level. Follow up queries may also be tied to particular attributes of the proficiency communication; proper characterization of such would enable more detailed analyses to be made, such as determining which aspects of proficiency the robot is having trouble communicating effectively. This same issue of subjectivity continues at the RPH stage in determining the type of events (corrections, interventions, modifications, or communications) that occur during the interaction. More formal criteria to classify the activities that correspond to each of these metrics would aid their usage and adoption in the field.

Self-assessment of communication proficiency. Self-assessment may also be performed by a robot to determine its proficiency at communicating information. For example, the robot in the Fetch scenario may track HUP metrics over time to determine which types of communications result in higher human comprehension clarity as evidenced by fewer follow up inquiries. Many of the RSA metrics could be used to this end and would be utilized by the robot when converting the derived task proficiency measure into a communication.

To date, it does not appear that robot systems are being explicitly designed with this capability, but the need for such considerations has been raised in the research literature [6]. Examples of relevant robot systems include those that make *in situ* decisions about communication modality choice based on associated cost parameters for each [162] or those that conduct feasibility analyses based on the inherent limitations of certain modalities [13]. The latter referenced paper presents a method to reduce the complexity of communication by spreading information across multiple modalities. If the scenario is expanded to include additional agents, the robot may further consider how choice of communication modality and attributes conveyed to one agent may impact other agents involved (e.g., [157]).

8 CONCLUSION

This paper reviews metrics for evaluating proficiency-based human-robot interactions, organized into four stages for these interactions: robot self-assessment of proficiency, robot communication of proficiency to the human, human understanding of the proficiency, and robot perception of the human's intentions, values, and assessments. Each stage is presented with a set of metric categories; metrics, evaluation criteria, and measurement considerations in each category are reviewed with references to similar metrics in related fields if appropriate.

The aim of this paper is to provide a starting point for common definitions of metrics that can enhance the development of robots capable of self-assessment and research around proficiency-based human-robot interaction. Continued experimentation is needed to further investigate and validate the correlations between metrics for proficiency-based human-robot interactions. The authors encourage others to iterate on the metrics and concepts presented in this paper to continue their evolution, validation, and adoption throughout the field.

ACKNOWLEDGMENTS

This work was supported by the U.S. Office of Naval Research (N00014-18-1-2503).

REFERENCES

[1] Henny Admoni, Anca Dragan, Siddhartha S Srinivasa, and Brian Scassellati. 2014. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 49–56.

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[3] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 51–58.

[4] David W Aha. 2018. Goal reasoning: Foundations, emerging applications, and prospects. *AI Magazine* 39, 2 (2018), 3–24.

[5] Matthew Aitken. 2016. *Assured human-autonomy interaction through machine self-confidence*. Ph.D. Dissertation. University of Colorado at Boulder.

[6] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[7] Paola Ardón, Eric Pairet, Yvan Petillot, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. 2020. Self-assessment of grasp affordance transfer. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9385–9392.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[9] Christophe Béné and Luc Doyen. 2018. From Resistance to Transformation: A Generic Metric of Resilience Through Viability. *Earth's Future* 6, 7 (2018), 979–996.

[10] Erwin R Boer. 2000. Behavioral entropy as an index of workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 44. SAGE Publications Sage CA: Los Angeles, CA, 125–128.

[11] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 708–713.

[12] Lyle A Brenner, Derek J Koehler, Varda Liberman, and Amos Tversky. 1996. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes* 65, 3 (1996), 212–219.

[13] Guilhem Buisan, Guillaume Sarthou, and Rachid Alami. 2020. Human Aware Task Planning Using Verbal Communication Feasibility and Costs. In *The 12th International Conference on Social Robotics (ICSR 2020)*.

[14] Janis A Cannon-Bowers, Clint A Bowers, and Alicia Sanchez. 2008. Using synthetic learning environments to train teams. *Work group learning: Understanding, improving and assessing how groups learn in organizations* (2008), 315–346.

[15] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim J Lilienthal. 2015. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *2015 European Conference on Mobile Robots (ECMR)*. IEEE, 1–6.

[16] Ravi Teja Chadalavada, Henrik Andreasson, Maike Schindler, Rainer Palm, and Achim J Lilienthal. 2020. Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction. *Robotics and Computer-Integrated Manufacturing* 61 (2020).

[17] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. 2019. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 29. 86–96.

[18] Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. 2018. Projection-aware task planning and execution for human-in-the-loop operation of robots in a mixed-reality workspace. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4476–4482.

[19] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 156–163.

[20] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[21] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. *Situation awareness-based agent transparency.* Technical Report. Army research lab Aberdeen Proving Ground Human Research and Engineering Directorate.

[22] Deepak Roy Chittajallu, Bo Dong, Paul Tunison, Roddy Collins, Katerina Wells, James Fleshman, Ganesh Sankaranarayanan, Steven Schwaitzberg, Lora Cavuoto, and Andinet Enquobahrie. 2019. XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019).* IEEE, 66–69.

[23] David C Conner, Stefan Kohlbrecher, Philipp Schillinger, Alberto Romay, Alexander Stumpf, Spyros Maniatopoulos, Hadas Kress-Gazit, and Oskar von Stryk. 2018. Collaborative autonomy between high-level behaviors and human operators for control of complex tasks with different humanoid robots. In *The DARPA Robotics Challenge Finals: Humanoid Robots To The Rescue.* Springer, 429–494.

[24] Jacob W Crandall, Michael A Goodrich, Dan R Olsen, and Curtis W Nielsen. 2005. Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35, 4 (2005), 438–449.

[25] Dana K Croft. 2003. Estimating intent for human-robot interaction. In *IEEE international conference on advanced robotics.* Citeseer, 810–815.

[26] Mary L Cummings, Andrew Clare, and Christin Hart. 2010. The role of human-automation consensus in multiple unmanned vehicle scheduling. *Human Factors* 52, 1 (2010), 17–27.

[27] Shreyansh Daftry, Sam Zeng, J Andrew Bagnell, and Martial Hebert. 2016. Introspective perception: Learning to predict failures in vision systems. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE, 1743–1750.

[28] Joost CF de Winter and Dimitra Dodou. 2011. Preparing drivers for dangerous situations: A critical reflection on continuous shared control. In *2011 IEEE International conference on systems, man, and cybernetics.* IEEE, 1050–1056.

[29] Bruce H Deatherage. 1972. Auditory and other sensory forms of information presentation. *Human engineering guide to equipment design* (1972), 123–160.

[30] Daniel Clement Dennett. 1989. *The intentional stance.* MIT Press.

[31] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly A Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 251–258.

[32] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 51–58.

[33] Anca D Dragan, Rachel M Holladay, and Siddhartha S Srinivasa. 2014. An Analysis of Deceptive Robot Motion.. In *Robotics: science and systems.* Citeseer, 10.

[34] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 301–308.

[35] Anca D Dragan and Siddhartha Srinivasa. 2014. Familiarization to robot motion. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.* 366–373.

[36] Jill L Drury, Jean Scholtz, and Holly A Yanco. 2003. Awareness in human-robot interactions. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, Vol. 1. IEEE, 912–918.

[37] Magnus Egerstedt, Jonathan N Pauli, Gennaro Notomista, and Seth Hutchinson. 2018. Robot ecology: Constraint-based control design for long duration autonomy. *Annual Reviews in Control* 46 (2018), 1–7.

[38] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. *CoRR* (2019).

[39] Mica R Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference.* IEEE, 789–795.

[40] Mica R Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human factors* 37, 1 (1995), 32–64.

[41] Jürgen Falb, Hermann Kaindl, Helmut Horacek, Cristian Bogdan, Roman Popp, and Edin Arnautovic. 2006. A discourse model for interaction design based on theories of human communication. In *CHI'06 extended abstracts on Human factors in computing systems.* 754–759.

[42] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. 2019. Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? *IEEE Computational Intelligence Magazine* 14, 1 (2019), 69–81.

[43] Tesca Fitzgerald, Ashok K Goel, and Andrea Thomaz. 2017. Human-Robot Co-Creativity: Task Transfer on a Spectrum of Similarity.. In *Proceedings of the 8th International Conference on Computational Creativity (ICCC).* 104–111.

[44] Tesca Fitzgerald, Ashok K Goel, and Andrea Thomaz. 2018. Human-guided object mapping for task transfer. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 2 (2018), 1–24.

[45] Tesca Fitzgerald, Elaine Short, Ashok K Goel, and Andrea Thomaz. 2019. Human-guided trajectory adaptation for tool transfer. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems.* 1350–1358.

[46] Stephen M Fleming and Nathaniel D Daw. 2017. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review* 124, 1 (2017), 91.

[47] Rudolf Flesch. 2007. Flesch-Kincaid readability test. 26, 2007 (2007), 3. https://rockstar-english.com/lessons/advanced/12-Flesch_Kincaid_Readability_Test.pdf

[48] Terrence Fong, Illah Nourbakhsh, Clayton Kunz, Lorenzo Fluckiger, John Schreiner, Robert Ambrose, Robert Burridge, Reid Simmons, Laura Hiatt, Alan Schultz, et al. 2005. The peer-to-peer human-robot interaction project. In *Space 2005*. 6750.

[49] Jodi Forlizzi and Carl DiSalvo. 2006. Service robots in the domestic environment: a study of the roomba vacuum in the home. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 258–265.

[50] Tyler Frasca, Zhao Han, Jordan Allspaw, Holly Yanco, and Matthias Scheutz. 2020. Going Cognitive: A Demonstration of the Utility of Task-General Cognitive Architectures for Adaptive Robotic Task Performance. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8110–8116.

[51] Nuwan Ganganath, Chi-Tsun Cheng, and K Tse Chi. 2015. Rapid replanning of energy-efficient paths for navigation on uneven terrains. In *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*. IEEE, 408–413.

[52] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.

[53] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1998. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages*. Springer, 1–10.

[54] Michael J Gielniak and Andrea L Thomaz. 2011. Generating anticipation in robot motion. In *2011 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 449–454.

[55] Michael A Goodrich, Erwin R Boer, Jacob W Crandall, Robert W Ricks, and Morgan L Quigley. 2004. Behavioral entropy in human-robot interaction. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems (PerMIS)*.

[56] Michael A Goodrich and Alan C Schultz. 2007. Human–Robot Interaction: A Survey. *Human–Computer Interaction* 1, 3 (2007), 203–275.

[57] Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication* 6, 2 (1969), 3–13.

[58] Corina Gurău, Chi Hay Tong, and Ingmar Posner. 2016. Fit for purpose? Predicting perception performance based on past experience. In *International Symposium on Experimental Robotics*. Springer, 454–464.

[59] Ellen Haas, MaryAnne Fields, Susan Hill, and Christopher Stachowiak. 2009. *Extreme scalability: Designing interfaces and algorithms for soldier-robotic swarm interaction*. Technical Report. Army Research Lab Aberdeen Proving Ground, Maryland.

[60] Jared Hamilton, Nhan Tran, and Tom Williams. 2020. Tradeoffs Between Effectiveness and Social Perception When Using Mixed Reality to Supplement Gesturally Limited Robots. In *International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction*, Vol. 3.

[61] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[62] Helen Hastie, Francisco J. Chiyah Garcia, David A. Robb, Atanas Laskov, and Pedro Patron. 2018. MIRIAM: A Multimodal Interface for Explaining the Reasoning Behind Actions of Remote Autonomous Systems. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 557–558.

[63] Bradley Hayes and Julie A Shah. 2017. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6586–6593.

[64] Cory J Hayes, Maryam Moosaei, and Laurel D Riek. 2016. Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 246–252.

[65] A. Heimerl, T. Baur, F. Lingenfelser, J. Wagner, and E. André. 2019. NOVA - A tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 109–115.

[66] Jeffrey J Hendrickson. 1989. Performance, preference, and visual scan patterns on a menu-based system: implications for interface design. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 217–222.

[67] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[68] Guy Hoffman and Cynthia Breazeal. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 1–8.

[69] Robert R Hoffman and Peter A Hancock. 2017. Measuring resilience. *Human factors* 59, 4 (2017), 564–581.

[70] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018).

[71] Curtis M Humphrey and Julie A Adams. 2015. Human roles for robot augmented first response. In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 1–6.

[72] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.

[73] Brett Israelsen, Nisar Ahmed, Eric Frew, Dale Lawrence, and Brian Argrow. 2019. Machine Self-confidence in Autonomous Systems via Meta-analysis of Decision Processes. In *International Conference on Applied Human Factors and Ergonomics*. Springer, 213–223.

[74] Adrien Jauffret, Nicolas Cuperlier, Philippe Gaussier, and Philippe Tarroux. 2013. From self-assessment to frustration, a small step toward autonomy in robotic navigation. *Frontiers in Neurorobotics* 7 (2013), 16.

[75] Benjamin Johnson, Mark Roberts, Thomas Apker, and David W Aha. 2016. Goal reasoning with informative expectations. In *Planning and Robotics: Papers from the ICAPS Workshop. London, UK: The Association for the Advancement of Artificial Intelligence*.

[76] Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science* 4, 1 (1980), 71–115.

[77] Daniel Kahneman and Amos Tversky. 1977. Intuitive prediction: Biases and corrective procedures. *Management Science* 12 (1977), 313–327.

[78] Adrian Kampa. 2018. The review of reliability factors related to industrial robots. *Robotics and Automation Engineering Journal* 3, 5 (2018), 1–5.

[79] Monisha Kanakaraj and Ram Mohana Reddy Guddeti. 2015. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, 169–170.

[80] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 676–682.

[81] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5092–5103.

[82] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 80–85.

[83] Christoph Kinkeldey, Alan M MacEachren, and Jochen Schiewe. 2014. How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal* 51, 4 (2014), 372–386.

[84] Shirlee-ann Knight and Janice Burn. 2005. Developing a framework for assessing information quality on the World Wide Web. *Informing Science* 8 (2005).

[85] Nathan Koenig, Leila Takayama, and Maja Matarić. 2010. Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks* 23, 8-9 (2010), 1104–1112.

[86] Vasiliki Kondyli, Mehul Bhatt, and Jakob Suchan. 2020. Towards a Human-Centred Cognitive Model of Visuospatial Complexity in Everyday Driving. In *Proceedings of the 9th European Starting AI Researchers' Symposium 2020*.

[87] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

[88] Vladimir Kulyukin, Chaitanya Gharpure, John Nicholson, and Sachin Pavithran. 2004. RFID in robot-assisted indoor navigation for the visually impaired. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 2. IEEE, 1979–1984.

[89] Ugur Kuter and Chris Miller. 2015. Computational mechanisms to support reporting of self confidence of automated/autonomous systems. In *2015 AAAI Fall Symposium Series*.

[90] Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 87–95.

[91] Carmen Lacave and Francisco J Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17, 2 (2002), 107–127.

[92] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[93] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. 2002. AIMQ: a methodology for information quality assessment. *Information & management* 40, 2 (2002), 133–146.

[94] Adam Eric Leeper, Kaijen Hsiao, Matei Ciocarlie, Leila Takayama, and David Gossow. 2012. Strategies for human-in-the-loop robotic grasping. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 1–8.

[95] Irwin P Levin, Sandra L Schneider, and Gary J Gaeth. 1998. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes* 76, 2 (1998), 149–188.

[96] Beibin Li, Laura Boccanfuso, Quan Wang, Erin Barney, Yeojin Amy Ahn, Claire Foster, Katarzyna Chawarska, Brian Scassellati, and Frederick Shic. 2016. Human robot activity classification based on accelerometer and gyroscope. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. 423–424.

[97] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. 2012. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2496–2505.

[98] Aleck M MacNally, Nir Lipovetzky, Miquel Ramirez, and Adrian R Pearce. 2018. Action selection for transparent planning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1327–1335.

[99] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1033–1041.

[100] Bertram F Malle, Louis J Moses, and Dare A Baldwin. 2001. *Intentions and intentionality: Foundations of social cognition*. MIT press.

[101] Nikolas Martelaro, David Sirkin, and Wendy Ju. 2015. DAZE: A real-time situation awareness measurement tool for driving. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 158–163.

[102] Jeremy A Marvel, Shelly Bagchi, Megan Zimmerman, and Brian Antonishek. 2020. Towards Effective Interface Designs for Collaborative HRI in Manufacturing: Metrics and Measures. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 4 (2020), 1–55.

[103] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2556–2563.

[104] Denis McKeown. 2005. Candidates for within-vehicle auditory displays. In *Proceedings of ICAD 05- Eleventh Meeting of the International Conference on Auditory Display*. 182–189.

[105] Ryszard S Michalski. 1983. A theory and methodology of inductive learning. In *Machine learning*. Springer, 83–134.

[106] Reuth Mirsky, Kobi Gal, Roni Stern, and Meir Kalech. 2019. Goal and plan recognition design for plan libraries. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–23.

[107] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*. 331–338.

[108] Sasanka Nagavalli, Shih-Yi Chien, Michael Lewis, Nilanjan Chakraborty, and Katia Sycara. 2015. Bounds of neglect benevolence in input timing for human interaction with robotic swarms. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 197–204.

[109] Chrystopher L Nehaniv, Kerstin Dautenhahn, Jens Kubacki, Martin Haegele, Christopher Parlitz, and Rachid Alami. 2005. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN) 2005*. IEEE, 371–377.

[110] Hannele Nicholson, Kathleen Eberhard, and Matthias Scheutz. 2010. "Um... I don't see any": The Function of Filled Pauses and Repairs. In *Proceedings of 5th Workshop on Disfluency in Spontaneous Speech 2010*. 89–92.

[111] Stefanos Nikolaidis and Julie Shah. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 33–40.

[112] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 72–78.

[113] Dan R Olsen, Jr and Michael A Goodrich. 2003. Metrics for evaluating human-robot interactions. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems (PerMIS)*.

[114] Sarah R Partan and Peter Marler. 2005. Issues in the classification of multimodal communication signals. *The American Naturalist* 166, 2 (2005), 231–245.

[115] Marcelo R Petry, Antonio Paulo Moreira, Rodrigo AM Braga, and Luis Paulo Reis. 2010. Shared control for obstacle avoidance in intelligent wheelchairs. In *2010 IEEE Conference on Robotics, Automation and Mechatronics*. IEEE, 182–187.

[116] Polly K Pook and Dana H Ballard. 1996. Deictic human/robot interaction. *Robotics and Autonomous Systems* 18, 1-2 (1996), 259–269.

[117] Nicoletta Prentzas, Andrew Nicolaides, Efthyvoulos Kyriacou, Antonis Kakas, and Constantinos Pattichis. 2019. Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 817–821.

[118] Felix Putze, Sebastian Hesslinger, Chun-Yu Tse, YunYing Huang, Christian Herff, Cuntai Guan, and Tanja Schultz. 2014. Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Frontiers in Neuroscience* 8 (2014), 373.

[119] Nemanja Rakicevic and Petar Kormushev. 2019. Active learning via informed search in movement parameter space for efficient robot task learning and transfer. *Autonomous Robots* 43, 8 (2019), 1917–1935.

[120] Vasumathi Raman and Hadas Kress-Gazit. 2012. Explaining impossible high-level robot behaviors. *IEEE Transactions on Robotics* 29, 1 (2012), 94–104.

[121] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton CT Lee, Mitchell P Marcus, and Hadas Kress-Gazit. 2013. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language.. In *Robotics: Science and Systems*, Vol. 2. Citeseer, 2–1.

[122] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. 2010. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 375–382.

[123] Thomas M Roehr and Yuping Shi. 2010. Using a self-confidence measure for a system-initiated switch between autonomy modes. In *Proceedings of the 10th international symposium on artificial intelligence, robotics and automation in space, Sapporo, Japan*. 507–514.

[124] Juan Jesús Roldán, Elena Peña-Tapia, Andrés Martín-Barrio, Miguel A Olivares-Méndez, Jaime Del Cerro, and Antonio Barrientos. 2017. Multi-robot interfaces and operator situational awareness: Study of the impact of immersion and prediction. *Sensors* 17, 8 (2017), 1720.

[125] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of vision* 7, 2 (2007), 17–17.

[126] Stephanie Rosenthal, Sai P Selvaraj, and Manuela Veloso. 2016. Verbalization: narration of autonomous robot experience. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 862–868.

[127] Karina A Roundtree, Jason R Cody, Jennifer Leaf, H Onan Demirel, and Julie A Adams. 2021. Human-collective visualization transparency. *Swarm Intelligence* (2021), 1–50.

[128] Matthew Rueben, Maja J Matarić, Eitan Rothberg, and Matthew Tang. 2020. Estimating and Influencing User Mental Models of a Robot's Perceptual Capabilities: Initial Development and Pilot Study. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 418–420.

[129] Dorsa Sadigh, S Shankar Sastry, and Sanjit A Seshia. 2019. Verifying robustness of human-aware autonomous cars. *IFAC-PapersOnLine* 51, 34 (2019), 131–138.

[130] Rosario Scalise, Yonatan Bisk, Maxwell Forbes, Daqing Yi, Yejin Choi, and Siddhartha Srinivasa. 2018. Balancing shared autonomy with human-robot communication. *arXiv preprint arXiv:1805.07719* (2018).

[131] Christopher Schneider, Adam David Barker, and Simon Andrew Dobson. 2014. Autonomous fault detection in self-healing systems using Restricted Boltzmann Machines. In *Proceedings of the 11th IEEE International Conference and Workshops on Engineering of Autonomic and Autonomous Systems (EASe 2014)*.

[132] Christopher Schneider, Adam David Barker, and Simon Andrew Dobson. 2015. Evaluating unsupervised fault detection in self-healing systems using stochastic primitives. *EAI Endorsed Transactions on Self-Adaptive Systems* (2015).

[133] Jean Scholtz. 2003. Theory and evaluation of human robot interactions. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. IEEE.

[134] Debra Schreckenghost, Tod Milam, and Terrence Fong. 2010. Measuring performance in real time during remote human-robot operations with adjustable autonomy. *IEEE Intelligent Systems* 5 (2010), 36–45.

[135] Debra L Schreckenghost, Tod Milam, and Terrence Fong. 2016. Techniques and tools for summarizing performance of robots operating remotely. In *14th International Conference on Space Operations*. 2310.

[136] Anthony R Selkowitz, Cintya A Larios, Shan G Lakhmani, and Jessie YC Chen. 2017. Displaying information to support transparency for autonomous platforms. In *Advances in Human Factors in Robots and Unmanned Systems*. Springer, 161–173.

[137] Meher T Shaikh and Michael A Goodrich. 2019. Intent-based Robotic Path-Replanning: When to Adapt New Paths in Dynamic Environments. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2857–2863.

[138] Ivan Shindev, Yu Sun, Michael Coovert, Jenny Pavlova, and Tiffany Lee. 2012. Exploration of intention expression for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 247–248.

[139] Moondeep C Shrestha, Ayano Kobayashi, Tomoya Onishi, Hayato Yanagawa, Yuta Yokoyama, Erika Uno, Alexander Schmitz, Mitsuhiro Kamezaki, and Shigeki Sugano. 2016. Exploring the use of light and display indicators for communicating directional intent. In *2016 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 1651–1656.

[140] Herbert A Simon. 2019. *The sciences of the artificial*. MIT Press.

[141] SKYbrary. 2020. Read-back or Hear-back. https://www.skybrary.aero/index.php/Read-back_or_Hear-back Accessed: 10-01-2020.

[142] Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.

[143] Shirin Sohrabi, Anton V Riabov, and Octavian Udrea. 2016. Plan Recognition as Planning Revisited. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. 3258–3264.

[144] Sichao Song and Seiji Yamada. 2018. Effect of Expressive Lights on Human Perception and Interpretation of Functional Robot. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[145] Sarath Sreedharan, Tathagata Chakraborti, Christian Muise, and Subbarao Kambhampati. 2019. Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning. *arXiv preprint arXiv:1903.07269* (2019).

[146] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2018. Hierarchical expertise level modeling for user specific contrastive explanations. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 4829–4836.

[147] Mohan Sridharan and Ben Meadows. 2019. Towards a Theory of Explanations for Human–Robot Collaboration. *KI-Künstliche Intelligenz* 33, 4 (2019), 331–342.

[148] Maria Staudte and Matthew W Crocker. 2009. Visual attention in spoken human-robot interaction. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 77–84.

[149] Maria Staudte and Matthew W Crocker. 2011. Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition* 120, 2 (2011), 268–291.

[150] Aaron Steinfeld. 2011. Slightly subversive methods for promoting use of autonomy in robots. In *RSS Workshop on Human-Robot Interaction: Perspectives and Contributions to Robotics from The Human Sciences*.

[151] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 33–40.

[152] Roykrong Sukkerd, Reid Simmons, and David Garlan. 2018. Toward explainable multi-objective probabilistic planning. In *2018 IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS)*. IEEE, 19–25.

[153] Katia Sycara and Gita Sukthankar. 2006. Literature review of teamwork models. *Robotics Institute, Carnegie Mellon University* 31 (2006), 31.

[154] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*. 69–76.

[155] Leila Takayama and Caroline Pantofaru. 2009. Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5495–5502.

[156] Haodan Tan, Jangwon Lee, and Gege Gao. 2018. Human-drone interaction: Drone delivery & services for social events. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 183–187.

[157] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. 2019. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 114–122.

[158] Da Tao, Haibo Tan, Hailiang Wang, Xu Zhang, Xingda Qu, and Tingru Zhang. 2019. A systematic review of physiological measures of mental workload. *International journal of environmental research and public health* 16, 15 (2019), 2716.

[159] Andrea L Thomaz, Guy Hoffman, and Cynthia Breazeal. 2006. Reinforcement learning with human teachers: Understanding how people want to teach robots. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 352–357.

[160] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 3–12.

[161] Paul Tompkins, Anthony Stentz, and David Wettergreen. 2006. Mission-level path planning and re-planning for rover exploration. *Robotics and Autonomous Systems* 54, 2 (2006), 174–183.

[162] Vaibhav V Unhelkar, Shen Li, and Julie A Shah. 2020. Decision-Making for Bidirectional Communication in Sequential Human-Robot Collaborative Tasks. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 329–341.

[163] Ryan W Wohleber, Kimberly Stowers, Jessie YC Chen, and Michael Barnes. 2017. Effects of agent transparency and communication framing on human-agent teaming. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 3427–3432.

[164] David D Woods, James Tittle, Magnus Feil, and Axel Roesler. 2004. Envisioning human-robot coordination in future operations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (2004), 210–218.

[165] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 1424–1431.

[166] Xuning Yang, Ayush Agrawal, Koushil Sreenath, and Nathan Michael. 2019. Online adaptive teleoperation via motion primitives for mobile robots. *Autonomous Robots* 43, 6 (2019), 1357–1373.

[167] Eiichi Yoshida, Kazuhito Yokoi, and Pierre Gergondet. 2010. Online replanning for reactive robot motion: Practical aspects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5927–5933.

[168] Changhe Yuan, Heejin Lim, and Tsai-Ching Lu. 2011. Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research* 42 (2011), 309–352.

[169] Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review* 24, 5 (2017), 1488–1500.

[170] Guangtao Zhang, Katsumi Minakata, and John Paulin Hansen. 2019. Enabling Real-Time measurement of situation awareness in robot teleoperation with a head-mounted display. In *Human Factors Society Conference*.

[171] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. 2014. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3566–3573.

[172] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.