41st International Conference on Environmental Systems
17 - 21 July 2011, Portland, Oregon

AIAA 2011-5248

A Survey of Quantitative Team Performance Metrics for Human-Robot Collaboration

Sharon M. Singer* and David L. Akin †

Space Systems Laboratory, University of Maryland, College Park, 20742, USA

Humans and robots have been increasingly used not only in the same workspace, but as team members that interact to accomplish overall mission goals. With a multitude of options developing for how humans and robots can simultaneously participate on a team, it has become necessary to quantitatively analyze the performance of the heterogeneous teams to enable comparison between different team configurations. This paper contains a survey of the field of collaborative human and robot team performance metric models, and examines existing overall team quantitative performance models to determine which are more applicable to future human and robotic space exploration missions.

I. Introduction

As technology has been advancing and designers have been looking to future applications, it has become increasingly evident that robotic technology can be used to supplement, augment, and improve human performance of tasks. With more options developing for how robotic technology can be used in an integrated human and robot team (HRT), new methods to assess individual agent performance and overall team performance have been developed. There has been significant work to resolve issues related to automating system functions, creating robotic agents that can resolve task-implementation issues independently, improving the communication platform, understanding of dialogue and spatial/world perspective, and pseudo-optimizing the different facets of team task performance.

There are numerous performance metrics in the literature to assess different aspects of task performance, e.g. completion time, resource usage, workload, reliability, etc. There has also been great diversity in the computational methods used to aggregate the results of the different individual task performance metrics into an overall ranking or score representing how well a task was completed. After the results for each task are tabulated, the next step is to rate the performance of the team as a whole in completing the overall scenario's goals. The objective of the team task performance analysis is to determine the quality and efficiency of the team's overall performance.

Developing application-generic quantitative team performance metrics has been a significant challenge. Which criteria to select and how to use them to evaluate performance has been much debated in the literature. The general approach has been to assess the performance ability of each agent, characterized by the important performance parameters (completion time, reliability of completion, mean time between failures, resource utilization and cost, mental workload, etc.). Some of the methods for evaluating overall team performance can accommodate ordering constraints, while others assume functional independence between the tasks.

Task performance analysis can be a useful tool during several different periods of mission operations. Experimental testing before and during the mission design phase provides mission planners with data on how well each agent could perform a wide variety of tasks in a future mission. The data evaluates each task separately rather than as part of a full mission. The data can be used during the design phase of a mission to determine which agent should perform each task (task allocation), and can facilitate arranging of each agent's schedule to ensure that all tasks are completed during the mission and that the workload is distributed between all of the agents. When performing *a priori* mission-level performance analysis, it is usually assumed that all of the goals and tasks will be completed (nominal task performance by each agent,

*PhD Candidate, Department of Aerospace Engineering, University of Maryland, AIAA Student Member.

†Associate Professor and Space Systems Laboratory Director, Department of Aerospace Engineering, University of Maryland, AIAA Senior Member.

and contingency operations are usually excluded). A priori analysis can also be performed to assess a team's ability to resolve anomalies during task performance. It is at this point in the design process that many different options can be considered, and comparisons can be drawn between different combinations of agents to determine the best performing configuration. This in turn will objectively select the final team members (agent selection), task allocation, mission objectives and task ordering.

Additionally, task performance analysis can be evaluated in real-time during a mission to assess how well the tasks are being completed. This type of performance analysis is based on real-time task performance data and can enable replanning and implementation if the team agents are not completing a task at a required level. In this case, task performance anlaysis can aid in efficient anomaly recovery.

This paper contains an extensive survey of the field of collaborative HRT performance metrics. It assesses the different priorities, assumptions, and methodologies incorporated into the leading quantitative models, with specific emphasis on determining which are most applicable in spaceflight applications. It discusses the challenges of implementing the methods, the shortcomings of the methods, and possible adaptations to increase the generality of the methods. This is a necessary first step to enable analysis of the effect of the different evaluation metrics on the design of team participation. In a unique contribution to the field, this survey sought to identify the conceptual pieces of existing HRI methods that would best synthesize into a universal, objective quantitative team task performance evaluation model, applicable to a wide range of applications.

This paper focuses on monitoring and measuring the overall team performance of a human and robot system. For more information on research about measuring and improving human workload in human and robot teams, the authors suggest reference 53[53] which contains a review of human workload models based on cognitive resource utilization.

In section 2, the component pieces of HRT architectures are discussed, in addition to individual task performance metrics used in space exploration applications to assess each. Section 3 contains methods developed in the literature for categorizing individual performance metrics. In section 4, quantitative models developed in the literature are described and categorized. Section 5 discusses the challenges to practical implementation of quantitative models, and section 6 contains the conclusions from this research effort.

II. Identifying Task-Relevant Performance Metrics

A variety of performance metrics have been identified to help define the phenomena that occur between a human and robot working cooperatively on a task. Some of the metrics are application specific, and some are more general. These individual metrics canvas the range of activities done by each agent. Common metrics include task completion time, reliability of task completion, mean time between failures, resource utilization and cost, mental workload, and a transition or switching cost for transfering mental attention from one task to another.

Burke[7] described application specific HRI metrics for urban search and rescue that could be applied to more generic scenarios (search, rescue (extrication), structural evaluation, medical assessment and treatment, information transfer, command and control, and logistics). It used existing models and software systems (Robot-Assisted Search and Rescue Coding System (RASCAR-CS) and the FAA's Controller-to-Controller Communication and Coordination Taxonomy (C4T), which capture what is communicated and how it is communicated for a team) to examine the robot's effects on human task performance within the context of human work. In essence, it logs how the robot affects and aids human performance.

Keller[36] presented a straightforward example of a human driving a car while talking on the phone to break down a multi-part task list to develop metrics that characterize the activities. The human's performance in the example was evaluated by considering a task as several simultaneous actions (visual, auditory, cognitive, psychomotor, all measured in relative rating scales). Activity in all of these pieces could lead to excessive workload demands, leading to performance errors (note, without any other agents taking part). This paper measured the scaled values of the number of human resource components needed for each task to simulate quantitative predictions of the human's workload over the entire task list.

Most models that attempt to measure the performance of a team including humans have a common metric for measuring human mental workload. The model that is used is the NASA-TLX[27] , which provides workload values for various human mental tasks. This results in subjective data from a post-experiment questionnaire to gage test subjects' estimated workload during different parts of an experiment.

An integral step in producing the most efficient task allocation and team schedules depends on the

selection of an objective set of task performance metrics to distinguish between each agent's performance, and to measure the effect of changing task allocation or task ordering on the completion of the overall mission objectives. Frequently, the metrics used to assess performance are built into the software planning and scheduling packages that automatically develop feasible mission plans. To reduce the computational load, most software packages create schedules that observe all constraints, but to reduce the computational load, often the software is written to only consider a single mission objective (e.g. reduce overall mission duration, minimize human time). A much smaller set of task performance metrics is required to analyze each agent's contribution to meeting the mission objective (e.g. task completion time). While this analysis will produce good results, the other performance parameters that are neglected (workload, resource depletion) in the analysis might have significant affects that are not brought to light.

A variety of software planning packages have been developed to plan a HRT's operations. In general, each research group working on a mission scheduling problem has used their own software packages, and there is no community consensus on which is preferred. While each package is unique, most can be applied to the same types of scenarios. For example, an overall lunar mission planning software package (HURON (Human-Robot Task Network Optimization)[17]) has been developed at JPL to facilitate task allocation, planning, and scheduling for combined human and robotic activities. It provides the architecture to develop optimal task allocation and scheduling for a scripted multi-agent lunar mission. The input problem description in this software package is decoupled from the planning and agent assignment. This indicates that this software, like many of the other packages developed, can be applied more broadly than it was originally designed.

Arnold's work[1] created an analytical framework to compare the advantages and disadvantages of different human-robot systems. Given a set of metrics, the goal of the framework was to lead towards optimal task performance. Although tasks and schedules were primarily scripted, the framework provided a guide for how to incorporate unplanned interventions into the schedule.

III. Architecture of a Human-Robot Team

The primary configuration differences between HRTs can be grouped into several broad categories. Whether the team's planning and performance analysis occurs offline or in real-time greatly influences the performance metrics that capture the important team performance characteristics. The scenario perspective of the mission planners can determine whether the humans or robots are favored to reduce workload. For real-time mission performance, the decision and control authority structure determines where critical decisions and work allocation is made. The type of communication network and the information it is capable of passing also characterizes a HRT. The autonomy level of each robotic agent determines the amount of human intervention in robotic tasks needed. How each agent (human and robot) receives information about their surrounding environment can vary greatly between different configurations. Each of these categories are described in more detail in the following subsections. A sampling of the related performance metrics is discussed for each architecture component.

A. Offline versus Real Time Performance Analysis

HRTs are either designed for use in real-time operations with real-time performance analysis feedback, or the modeling is done offline. Whether the mission scenario is in real-time or decided *a priori* has a significant effect on the task allocation, planning, and scheduling that occurs, and the performance metrics that best characterize them. Done *a priori*, the processes can be cycled through software to optimize the scenario over the entire mission length. Computational efficiency may still be an issue, but it is removed from being mission critical. As seen in Figure 1, HRT performance can be assessed at many different stages during *a priori* analysis, and can provide feedback during real-time task performance. The metrics used to assess both cases will be different, and geared towards the important components of the tasks.

Rather than relying on task completion times to measure how well an agent performs a task (as done in HURON), Schreckenghost[59] developed the concept of a work efficiency index (WEI). The WEI metric represents the ratio of productive time to overhead time that occurs in an agents' task schedule. Schreckenghost notes that WEI is more valuable in *a priori* analysis because it utilizes total productive time and total overhead time for a mission. Using these metrics, however, is not practical for real-time task performance analysis.

Real-time analysis requires a significantly different hardware architecture to support the quantity of

A priori	Real-time
Select agents for team compare skills and task performance capabilities for overall utility	**Feedback to supervisor** monitor activity, identify decreases in productivity, rearrange schedule or reallocate tasks
Task allocation distribute tasks according to which agent performs each best	**Feedback to operator** monitor activity, identify performance anomaly, fix and recover task performance
Compare each team against different schedules determine most efficient arrangement of tasks within constraints	**Feedback to other team members** monitor activity, increase knowledge of teammate's situation, alert to drop in teammate's productivity, alert to anomalous performance
Compare different team configurations select the best overall performing team for the mission	

Figure 1. Applications for Performance Analysis in A Priori and Real-time

computing and processing of sensor information that must keep up with the real-time operations. The advantage of real-time analysis is that details of team performance (including task allocation and anomaly resolution) can be addressed and altered as needed to improve the team performance for the rest of the mission.

For real-time operations, it is often advantageous to distribute more of the thinking capability to the active agents involved. A human supervisor can make these decisions in real time, but there are limits to human information retention, which depends on the quantity to be sifted through, and the complexity and heterogeneity of the task. These lead to a limit to the number of robots that an operator can control simultaneously ("fan out" (FO)[44]) without degradation of team efficiency or under-utilization of team resources. An overloaded supervisor will develop a backlog of actions and decisions. If a robot has to wait for an operator's instructions, an anomaly resolution plan, or confirmation of any kind, this reduces the efficiency of the robotic agent.

If a task is taking longer than expected, other subsequent tasks can be reassigned to balance out the schedule to maintain efficiency. With distributed levels of intelligence, there may not be an initial team plan to deviate from. Decisions could be made as each task-need arises. In operations with supervisor decisions made *a priori*, none of these performance qualities can be altered mid-task, and must wait through the end of the task or scenario.

B. Scenario perspective

Whether a scenario is designed from a human-centered or robot-centered perspective influences which variables and resources are the most mission-critical and can change the assumptions behind the scenario development. For example, a human-centered model might assume that human extra-vehicular activity (EVA) time is the most critical commodity because of risk exposure to the human outside of the spacecraft. Therefore, the scenario will be planned such that the human agent is only used when absolutely needed. Alternatively, in a robot-centered model, the robot could be designed to be fully capable and independent of the human. Any evaluation of task or team performance would be geared towards evaluating efficiency and effectiveness with respect to the robot.

One perspective to assess a team's efficiency, effectiveness, and productivity is by analyzing its resiliency

to failure. Kannan[33] defined a metric for calculating how useful fault tolerance is for a multi-robot team. The paper provided a practical method to calculate the redundancy of a system. Shah[62] examined the productivity of a HRT through the mean time between interventions (MTBI), mean time completing an intervention (MTCI), and the probability that an intervention would be needed. It described the effect of unplanned interventions on a team's productivity, and demonstrated that the team's productivity was much more sensitive to MTBI than to MTCI.

Within each scenario perspective, the roles of the humans and robots can vary. The roles have significant effect on the way tasks are completed and the workload distributed. Scholtz[57] discussed five different roles that a human can have when working together with a robot on a task: supervisor, operator, mechanic, peer, and bystander. As a supervisor, the human gives direct instructions to the robot. The detail of the instructions (high level commands versus primitive level scripted tasks) depends on the robot's abilities to interpret and implement instructions. In addition, as a supervisor the human is responsible for forming and creating a plan to pursue the overall goals. A human could also be an operator who directly influences the robot's actions. As a mechanic, the human is collocated with the robot and participates physically with the robot's hardware and actions. Humans as peers to the robot are also collocated. In this role the human gives commands or information aid to the robot, but does not participate in robot upkeep. The final role of a human interacting with a robot is as bystander. In this role the human does not participate in any tasks with the robot. The human is only an obstacle in the robot's environment.

In Singer[63] , the effect of redefining a robot's role on a HRT was analyzed. The roles were differentiated by the safety considerations that determined the ability of a robot to work within the human crew's proximity. In each role the robot exercised different physical capabilities that determined the portion of mission tasks that it could perform. With the overall objective of minimizing human EVA time, the role definitions were traced through the task allocation and scheduling process to determine how the different perspectives changed the overall team's performance and efficiency.

C. Autonomy level

Robots have been developed with increasing levels of autonomy, able to not only carry out scripted tasks by themselves, but identify anomalies, and come up with their own resolution plans. Technology has developed not only to create distinct autonomy levels for a given robotic system, but also to allow adjustable, or sliding autonomy (allows the autonomy level to change as needed within a scenario). Miller[41] has done research with varying levels of robot autonomy to determine their effect on human counterparts. Heger[28] demonstrated that the concept of sliding autonomy could reduce the probability of an irrecoverable failure and would increase overall team efficiency.

Due to safety and reliability concerns, the majority of robots used to date in space mission applications (as peers to human agents) have been controlled by a supervisor (e.g. the Space Shuttle's remote manipulator system). For this reason, much of the literature presumes supervisory control of robots in a scenario that actively involves a human presence. In the scenarios where the human supervisor is off-scene, robots are designed with a higher level of autonomy (e.g. Mars rovers), allowing them to better analyze their own situation and pursue goals and waypoints independently.

In Goodrich[25] , design principles were developed that would guide the development of human-robot autonomy architectures to make interactions more efficient. More recently, in Goodrich[26] the success of two operator management styles were described for a team of robots with adjustable autonomy. Usually supervised robots would each be given orders sequentially - the operator's attention would switch to a new robot only when finished with the current one. Goodrich proposed an alternative style which would direct the robots in a method similar to a sports playbook. After a play was announced, each robot would determine the course of its own actions to facilitate achievement of the team goal.

Howard[31] examined four different human-team leadership styles to identify their defining characteristics and adapt them to HRTs to create more effective configurations. Her analysis suggested more effective teams would be formed if humans used a directive style or a transactional style of leadership with their robotic peers. In a directive style, the human is the supervisor, giving commands and establishing the overall goals. In a transactional style, the human plays the role of the operator, monitoring the robot and its task completion, and giving commands at a lower level.

1. Metrics for Autonomy

Glas[24] discussed two new metrics to represent task difficulty for a human monitoring and controlling a multi-robot team. Situation coverage (SC) is the percentage (of total tasks) in which a robot understands the directive. In multi-robot situations, SC measures the upper bound on the ability of the system to operate autonomously. Critical time ratio (CTR) is a ratio of the time that a robot is performing mission-critical tasks to the total amount of time that a robot is actively engaged in tasks. When two robots have high CTR, an operator will have a higher workload. A human can improve performance by reducing the number of conflicts that occur between robot attention demands.

A new way of thinking about HRTs was developed by Olsen and Crandall to research the most effective ways for humans and robots to work together on tasks, specifically addressing autonomy levels. They developed a series of metrics that addressed the specific unique human and robot interactions at a generic level. In Olsen[44–46] and Goodrich[25] , they introduce several new metrics. Olsen and Goodrich concretized neglect time (NT) as a metric to measure robot autonomy[44] to reduce interaction effort (IE) for both a human operator and the robot without reducing the effectiveness of the team. This metric, however, considers human attention and does not consider the physical abilities, limitations, or usage of the human. NT is the amount of time a robot can function independently of the human. This interval is demarcated by a user-defined drop in the effectiveness of the robot's task performance to below a threshold, which can only then be raised by human intervention. NT and a quantity representing a human's interaction effort (IE) combine to calculate a quantity for robot attention demand (RAD), which is the fraction of the total task time that the robot requires attention. Fan out (FO, defined as the inverse of RAD), represents the number of robots a human can monitor and control before the decrease in overall team performance drops past a threshold. Crandall[11] used these metrics to predict the performance of a multi-robot team and validate the method with experiments taxing the operator's attention. In addition, this method found the performance thresholds that maximize team performance, given a team size[10] .

Crandall[8] solidified these concepts into a methodology that has significantly affected the field. Wang[69] extended Crandall's NT metric to include the coordination demands (CD) and the effects of robot heterogeneity. Occupied time (OT) differentiates between wait times resulting from an operator having low situational awareness and wait times resulting from a queue of needy robots. Wang[68] expanded the estimation of an operator's FO limit to allow analysis for N-robots teams.

Elara[16] extended Crandall's model to include the possibility that a robot did not correctly interpret or respond to a user's command. In a false positive, a robot rejects a "correct" interaction. In a false negative, a robot fails to reject an "incorrect" interaction. False alarm time (FAT) represents the time that is spent identifying a false alarm and recovering from the delay.

In the two papers by Schreckenghost[59,60] the authors described metrics and a quantitative model to assess real-time adjustable autonomy. They suggested measuring the degree of robot independence (to decrease human intervention time) based on the time spent on unplanned interventions. The computed performance metrics results were used in real time by controllers. It should be noted, however, that results were based on percentage of mission time but that there was no indication given to relate robot to human time. If a robot performed its share of the tasks at a slower rate than in the previous scenario run, then all else being the same, it would register as spending more time in an autonomous mode. This could skew the results because performing tasks slower does not require a different level of autonomy, and should not be perceived as such.

D. Hierarchical and Distributed Decision Making

Decision making for a HRT can either be hierarchical or distributed. Most HRTs to date have used hierarchical decision making, where a supervisor oversees the team and there can be several levels of authority. In distributed decision making, a team is made up of individuals who can each make task decisions for themselves, relying on their sensory information of the world around them, and relevant information passed to them by a neighboring agent.

In Fong[20,21] , the Peer-to-Peer Human-Robot Interaction Project was described. In the model, a task executive allocated tasks to agents which were both capable of performing the work and were not actively working on another task. A feature of this tool was that the task executive had the hierarchical decision authority to interrupt an agent's task and send it to perform another (preemption). Once a task was assigned, however, decision making was passed to the individual agent to plan and schedule its own task.

Ponda[52] described the development of a real-time decision making framework that allocated tasks for a

heterogeneous HRT. The predictive model developed schedules for all of the agents based on agent availability, workload, and any coordination requirements between the humans and robots needed to complete the given task. As the number and diversity of agents in the combinatorial problem increased, centralized planning became computationally prohibitive. Using a decentralized approach (such as decentralized auction algorithms) to facilitate the task allocation reduced this challenge, and facilitated the generation of a more efficient and effective architecture.

Billman[4] presented an extensive table of performance metrics, their relevant parameters, and the human factors concerns associated with each. These were used in several experiments to evaluate a mixed-initiative (both robot and human can initiate communication and tasks) human-autonomous unmanned vehicle teams for the Navy's Intelligent Autonomy program.

Saleh[56] expanded Crandall's model[8,10] to include two factors that represent the level of trust for both the human and the robot during cooperative interaction intervals. Trust in a HRT is representative of the human's belief that a robot understood its instructions, that it will correctly assess its situation, and that it will perform tasks correctly without requiring assistance. In Saleh's[56] work, a human's trust level was proportional to the human's FO, and indirectly proportional to the robot attention demand (RAD). RAD was defined to be a function of both direct interaction time and indirect interaction time. Indirect interaction time was explained to be the interval when the robot is working autonomously but the human, due to reduced trust, monitors the robot's progress rather than applying the human's full attention to another task. Another factor gaged the level of human trust in the autonomous system making correct choices and following through with those choices. Decreased trust increased the time duration of the interaction.

Freedy[22] presented a methodology to assess trust levels for a HRT during operations. Expected value statistics were used to decide whether to allocate control to a robot. This work used a tool developed by Visser[14] , the Mixed Initiative Team Performance Assessment System (MITPAS), which calculated a compound "goodness" score as a relative expected loss score, derived from observed human task allocation decision behavior, risk and observed robot performance. This was used to maximize the performance of the overall team. The MITPAS system was developed and validated for operator training on unmanned vehicles.

Marble[40] described the level of trust that the human had both for a robot making a correct decision, and in the robot completing a decided path effectively. This work assessed trust limitations in mixed-initiative systems. The operator's situational awareness and trust within an experiment in which operators directed mobile robots to perform tasks was measured for a given scenario under five different autonomy modes.

E. Situational Awareness

Situational awareness, as discussed in the literature, can refer to three different perspectives: that of a human supervisor referring to knowledge of overall mission operations, that of a robot control operator referring to knowledge of the robot's immediate environment and obstacles, and of the robot's awareness of its own environment. Quantifying situational awareness has been a challenge because the idea is very ethereal and has many definitions and contributing factors. Bruemmer[6] proposed using human workload, error, and overall performance as quantities that gage the effectiveness of robots in a mixed-initiative environment. The amount of human error could be seen as a reflection of operator fatigue, but more dominantly a lack of human situational awareness of the task environment. The model created a text-based dialogue between the human and robot, and produced a 3-D representation for shared understanding about the task and the environment. While this research was intended to be a proof-of-concept argument that increased autonomy and better operator interfaces could improve robot navigation, it also demonstrated collaborative control where robots were effectively used as trusted peers.

Scholtz[58] described the implementation of a tool that evaluates the situational awareness provided to an operator through a user interface used for supervisory control of autonomous vehicles. Each interface was evaluated to find how well each facilitates a user's situational awareness.

Lampe[37] described a metric representing environmental complexity and a robot's information of it as a measure of robot autonomy. Nehmzow[43] presented a novel quantitative model of a robot's interaction with its environment based on the robot's trajectory over time.

Nehme[42] created a model to evaluate the performance of supervisory control of multi-unmanned vehicles. Included in the model were both operator variables and variables that pertain to the entire team. A unique contribution of this model was that it modeled operator limits by accounting for wait time due to loss of situational awareness. The integrated model facilitated comparing design development and was capable of searching the large design space to obtain the overall goal of maximizing team efficiency.

Hwang[32] developed a model that used each agent's knowledge of the other agent's state and of the environment's state to measure the changing interactions between the agents. This model can be used to measure the situational awareness of each agent, and the interaction effort required at each state of the cooperation.

F. Communication Architecture

For a heterogeneous HRT, creating a communication architecture that facilitates passing relevant information when needed, and avoiding information overload (passing all sensory data) is imperative for an effective team. There are many different architectures that can be utilized in the communication framework, depending on the agents involved in a mission. In most supervisory control scenarios, human operators or peers communicate to the robots, but the robots are not capable of passing confirmation of task completion messages to the humans, and are not capable of querying their supervisors with questions or to address task difficulties.

Kaupp[35] presented a robot-centric model to facilitate bidirectional communication between a human peer and a mobile robot. Within this paradigm, robots passed images of the environment to the human to increase the humans' situational awareness and to process details from the images. Humans were viewed as a resource which could be queried for information and observations about the environment, but as with all resource usage, it came at a cost. The cost of querying operators was traded off within the architecture against the expected benefit of the new information. This model could be used to determine what and when to communicate between a human and robot engaged in collaborative tasks.

Approaching the difficulties of team communication from a different direction, Trafton[66] presented a cognitive architecture to facilitate perspective-taking for collaborative human and robot interaction. The goal of the model was to produce intelligent robots that were capable of reasoning from a human-perspective by modeling how humans integrate multiple information and environmental representations into a world model. This type of reasoning allowed a robot to spatially interpret relative commands from a human, e.x. "give me the wrench on the right". To select the specified object correctly, the robot needed to be able to simulate the perspective of the human, which required implicit rotations of the world environment.

The Human Robot Interaction Operating System (HRI/OS) software[20,21] was created with several distinct objectives. In addition to facilitating perspective-taking, the overall goal was to both maximize the amount of work done by the entire team and to reduce the number and duration of EVAs needed to complete the task list. To further reduce the human's workload in the scenarios, human and robot communication was designed to have as natural (for humans) interaction mechanisms as possible. This includes having a text-to-speech agent that verbalized a robot's responses for a human to hear, and a speech recognition agent that translated a human's verbal response into text for a robot to parse.

Each metric that quantifies a portion of a HRT interaction provides useful information to a mission designer or supervisor. Although there are numerous metrics to describe components of a HRT architecture and assess an agent's performance on individual tasks, it is clear that one single metric will not sufficiently explain the team's task performance. Quite the contrary. Each of the metrics mentioned in this discussion cover a different aspect of a HRT's performance, but successful performance in one architectural category does not necessarily entail good performance in the other categories. Several metrics would need to be selected to comprehensively determine a HRT's performance by analyzing the different facets of the team's configuration. The problem that remains, however, is how to integrate the results from multiple metrics into a meaningful, overall objective team performance score. It is only with this type of rating system that two different HRTs could be compared to determine their relative performance on an overall mission, rather than on single tasks or other criteria.

IV. Measuring Human-Robot Team Performance

Most team configurations and hardware used in past experiments have been subjectively selected for their specific mission scenarios. When planning task scenarios, designers frequently selected a subjective set of mission-critical constraints, and used only a small set of performance metrics with subjective weightings to measure the overall performance of a team. In human-centered operations, space mission operations sought to minimize human EVA time as the time-critical mission driver. Alternatively, because robots are a valuable resource, future missions could be designed to minimize the amount of time a robot is in standby, waiting

for instruction from an operator. Although these methods and the models they produce are valuable in evaluating mission-specific scenarios, it should be possible to generalize this procedure to select a complete set of mission-generic performance metrics to sample all of the important phenomena that occur during the human-robot interaction, and evaluate a team's overall performance based on all of these factors.

The diversity of robotic technology available creates a multitude of new opportunities in task performance. Utilizing the new capabilities for hardware, software, sensors and system integration, and communication architecture could lead to a greater level of mission diversity, and facilitate scenarios that had previously been impossible. These may help improve the attainment of the overall objective.

A primary obstacle to integrating new robotic technology into familiar mission scenarios is the inability to quantitatively compare overall team performance between very different team configurations. Is it possible to objectively and conclusively demonstrate whether a standard two-human Space Shuttle EVA team, a combined human and semi-autonomous robot team, or an autonomous robot performs an entire mission scenario better than the other configurations without limiting the analysis to a few important metrics? Can the benefit derived from using one team configuration over another be quantified? As a designer, is it important to retain knowledge of the tasks completed by each of the agents in a given scenario separately? Or is the overall interest primarily in the overall team's performance of a scenario? These are the questions that drive the development of an overall team performance metric - to enable future comparison between diverse team configurations and determine the optimal for a scenario.

Fong[19] described a proof-of-concept experiment to show that robotic reconnaissance missions could supplement human exploration. The concept was to send a mobile robot to scout out the terrain to aid in selecting a traversal path for a two human EVA team on the moon. The scout would collect environmental data that could be used to improve each day's mission plan and improve human crew performance by reducing the operational risk and increasing its situational awareness. Individual task metrics for the scout were reported in real-time to the operator, were compiled into overall metrics and the data was used by operators to adjust the current operation.

This type of overall mission analysis to differentiate team configurations could be an inordinately valuable tool for mission designers. The challenge then becomes creating a quantitative, overall model to measure a team's performance for a generic mission, to enable broad use of the tool (and remove the need for mission-specific models).

A. Team Performance Metrics in Related Fields

There are many applications that use humans and robotic technology cooperatively to complete a task. Bechar[2,3] and Oren[47] developed a methodology for a performance analysis of human-robot collaboration in visual target recognition tasks. It was based on a quantitative model[2] with four levels of human-robot collaboration, ranging from manual to fully autonomous. The metrics used were from signal detection theory: hit, false alarm, miss and correct rejection. These quantify the influence of the robot, human, environment, and task to determine the optimal operational level based on input parameters.

The development of new computer user interfaces (UI) encounter similar challenges to those of HRTs. Stanton[64] demonstrated three experiments to record the usability of the UI and how it affected the user's performance in the application of human supervisors directing urban search and rescue mobile robots. This paper collected the data and left selection of an evaluation method for future research. Yanco[71] demonstrated a proof-of-concept method for evaluating user control interfaces from rovers at an artificial intelligence search and rescue competition. Yanco categorized the evaluation of UIs into six categories of methods: effectiveness, efficiency, user satisfaction, inspection methods, empirical methods, and formal methods. Analysis of the competition runs and scoring results in several concrete UI suggestions for the type of information that, if present to the user, would have improved task performance. Bruemmer[6] measured the usability of an interface based on how the human workload and human error affected the performance of the autonomous robot.

Glas[24] developed a UI to increase the performance of multiple social robots by designing the UI to provide information to the operator in an intuitive way, increasing the human's ability to monitor several robots while specifically controlling a different one. Task difficulty, as used in this paper, referred to a task being more difficult if a robot required a user's attention to help with the task. The UI can be used to reduce the amount of attention each robot needs caused by task difficulty (to be differentiated from attention needed to relay a directive).

B. Categorization of Human-Robot Team Performance Metrics

The large variety of options available for selecting and organizing team members in HRTs raises the question of how to evaluate their performance in methods that transcend the individual details of a team: in essence, how to compare apples and oranges. One approach is to select an individual metric and apply it to different teams. This is not sufficient, however, because the different priorities of the teams cannot be reflected in a single metric. Additionally, if the robot of team A excels in a mission, but it is the human of team B that excels in the same mission, how do you objectively weigh the two options?

There have been several attempts recently to create categorizations for individual performance metrics. Categorizing the metrics would aid the search of commonality between them. It could then be anticipated that a selection of metrics could be drawn that span each of the categories and facilitates a wider understanding of a team's overall performance. One approach has been to create taxonomies for human-robot interaction to specify categories for relevant details. These taxonomies have emphasized creating a rubric for comparing and contrasting the design decisions of different HRT.

Gerkey[23] presented a taxonomy of task allocation for multi-robot systems based on the relative utility of one robot performing a task. This paper provided formal categorization of the different application problems that can occur (single-task robots versus multi-task robots, single-robot tasks versus multi-robot tasks, and instantaneous assignment of task allocation versus time-extened assignment, which assumes some predictive knowledge of the tasks that will need to be completed in the future). Several different algorithms were proposed to efficiently calculate the task allocation for each of these types of problems, and the algorithms were compared for computation requirements, communication requirements, and solution quality. While this paper focused on purely robotic teams, its methodology can easily be extended from the heterogeneous robot teams considered to a HRT.

Yanco[70] created a taxonomy that emphasized the physical characteristics of a team's dynamics. Team composition (defined as the ratio of humans to robots and including details about the different types of robots) was a defining characteristic - a team of one human and one robot will vary significantly in its overall task performance from a team of one human and a two-robot team. The degree to which a robotic agent is dependent on a human decision maker (autonomy level) will also significantly influence performance. Does each robotic agent require human intervention at some point? Can it resolve problems by itself? Can a human operate more than one robot in a given scenario? Does a robot need to resolve conflicting instructions received from different human agents? How much required interaction is necessary between team agents to complete tasks?

Yanco also described the information type, variety, and level provided to operators to facilitate situational awareness in decision making (including sensor information available, sensor fusion, and data pre-processing). Yanco cites Ellis[18] for the time and space part of the taxonomy, differentiating into four categories: humans and robots functioning at the same time (synchronous) or at different times (asynchronous), and physically located in the same place (collocated) or at a separated distance (non-collocated). This can clearly be seen to differentiate between a robot teleoperated from a control room and a robot peer working along-side its human contemporary.

Yanco's taxonomy applies to a broad variety of applications. It also includes a subjective category to indicate the relative importance of a task's performance, termed its criticality. This allows differentiation between mission critical performance applications (urban search and rescue) distinguished from applications with less severe consequences for failing a task (robot soccer team).

Steinfeld[65] provided a generalized categorization for common metrics that apply to human-robot interaction. Metrics were split into three categories: human, robot, and overall system. Each of these categories contains metrics in five additional categories (navigation, perception, management (including task allocation, resource allocation, and coordination), manipulation (interaction with the environment), and social).

These three taxonomies highlight different aspects of a HRT. Gerkey's taxonomy facilitates task assignment depending on team configuration, but it does not include categories for the planning and scheduling portion of the design problem, including workload, environmental knowledge, completion times, etc.. Yanco's taxonomy emphasizes the interaction mechanisms between the humans and robots much more highly, but provides little guidance for task allocation metrics or the effect of a HRT with distributed decision making. In Steinfeld's taxonomy, the details of each agent's task performance are clarified, and the interaction mechanisms between them are accounted for, but does not offer suggestions on how to combine a selection of metrics (including overall team performance metrics and agent performance metrics) into an overall picture. This includes incorporating multiple overall team performance metrics into a single ranking. Each of

these taxonomies has been successfully used independently, and can apply to a broad range of heterogeneous teams composed of humans and robots. A future analysis might find value in integrating them for a more comprehensive analysis of what occurs during HRT interactions.

A unique approach to categorize team metrics has been developed recently by researchers at the Massachusetts Institute of Technology. Building on their previous work, the researchers[8,12,15,50,51] developed the concept of metric classes under the application of human supervisory control. They addressed the problem of which metrics should be selected to completely assess team performance. The more metrics were included in an analysis, the more computationally complex the analysis became and, as pointed out in Donmez,[15] using metrics that correlate to the same data can result in finding false significant effects in the data. Categories of different types of metrics were created with the goal of facilitating selection of metrics to be applied to a HRT. The resulting guiding principle is that to efficiently and comprehensively measure the overall team task performance, at least one metric from each metric class category should be selected. Crandall[9] applied the formal metric classes framework to the application of UAVs to create a methodology for developing a predictive model of the interaction between humans and unmanned vehicles.

Although there are several different methods for categorizing performance metrics, the same goal is sought by each: a logical progression between groups of metrics would allow a mission designer to select the most relevant from a group, and would facilitate a mission designer's overall scenario consideration by encouraging the designer to select metrics from every category.

V. Quantitative Models of Human-Robot Team Performance

A. Building Performance Metrics into Quantitative Models

Individual performance metrics for components of a HRT provide insight into the different qualities of team configurations. What is lacking, however, is a method or framework to incorporate these metrics into quantitative models designed to determine the quality of and efficiency of a team's overall task performance. This type of system-level analysis could determine which team will perform better for a given mission scenario without needing to run additional simulations or experiments, and will result in a better, more comprehensive picture of what is actually occurring. Each model will have built into it assumptions, priorities, and differing methodologies. Differentiating which model would be preferred in different scenarios can be a difficult problem. The following discussion seeks to categorize existing models and describe their characteristics.

Without a formal framework, a designer seeking to compare different team configurations would need to compare numeric results from a multitude of performance metrics to discover which configuration is best across the board. Schreckenghost[59,60] described a real-time sliding autonomy robot assessment tool that provided several streams of individual performance metric data to an operator, covering the important components of robot operations. This type of analysis of several metric data steams becomes much more complicated with the increase in metric set size. The calculation of composite task scores are used in other models to break down the complicated analysis such that comparison between team configurations can be done by assessing a single value for each.

Rodriguez[54] provided a justification for using a composite task score by comparing the analysis to the scoring of an athletic competition. A composite numerical score for the overall competition was obtained from the scores on each of the individual events. If different sets of events were selected, the results were different. If all events were weighted equally, the individual scores were summed. If some events were deemed more valuable than others, a weighting factor was used to represent the importance. The final result from the competition was then a single score for each participant.

B. Parasuraman's Model Categorization

There are many implicit choices made in deciding which model to use, including whether the level of automation is assessed (including the coordination demands placed on other agents), and which components of the mission schedule are more important to the analyst. In a frequently cited paper, Parasuraman[49] presented two evaluation criteria to be used for this decision. The primary criterion was the effect of a given design selection on human performance, e.g. how was the human's performance influenced by this selection? The secondary evaluative criterion included several additional important pieces of a model, including reliability and cost. These criteria were for a human-centered model of HRTs, and were intended as a framework to guide an objective selection of a HRT for a given application. Parasuraman[49] identified four categories that

represent existing quantitative HRT performance models: task load models, expected value statistics models, cognitive systems models, and state transition network models. Each of these types of models combine task performance data for each agent and results in an overall scenario-level assessment of team performance. It is with these types of quantitative models that it becomes possible to compare, in essence, apples and oranges - it facilitates comparison of different team configurations to select the best overall team for a mission.

1. Task Load Models

Task load models evaluate the effort required for each agent to complete a task. The goal of task load models is to examine the effect of the tasks themselves on system performance, operator demands, and how task performance responds to a range of autonomy levels. This can be in terms of time as a resource (task completion time), resources used (either an agent's or cummulative for the team, e.g. power), or overall, repetitive, and fatiguing workload levels for a human. Selecting this model, the designer's goal is to distribute the total workload across the agents such that the total task list could be completed in the minimum amount of time, constrained by the finite amount of resources and agents available. This type of model can be used to compare which agent should perform portions of a task list, and to compare which agent contributes most to the entire mission. It therefore can also be used as a task allocation schema, and a method to objectively select amongst a set of possible agents for the final team configuration.

A common method to combine metrics for a task load model involves a pairwise comparison of effort and execution time for each agent. This approach has been used by many researchers, particularly to determine the final task allocation amongst a HRT. Howard[29] validated using a composite task score (termed the sequence execution parameter) as a task allocation scheme instead of individual task performance metrics in an experiment to guide a robot to a target location. The overall composite task score was run through a genetic algorithm in which the weightings in the fitness function changed with user feedback. In another model, Howard[30,31] used a composite task score for each agent to balance performance score and mental workload for teleoperated and autonomous robot control of an autonomous rendezvous and docking scenario. All of these works used the composite task score for the final task allocation decisions.

Task load models can be the most straight-forward of the model types to implement because the data required for standard-type missions, primarily execution time and workload quantities, will be readily available from previous missions or testing. No extra experimentation will be needed to obtain the required inputs to the model. Each agent of the HRT will have known capabilities, and the quantized effort required to complete tasks could be estimated based on similar tasks. These types of models are preferred in space applications that are planned from a human-centered perspective. As an *a priori* analysis, task load models can provide verification of the task allocation schema used, and the mission objectives can be easily interpreted into the model to produce desired results. It can be much more challenging, however, to use task load models if a wealth of performance data for an agent is not available, e.g. for a new robotic system. Task completion time is one of the primary criterion in the model. New technology or tasks that agents have not attempted before will require either significant experimental testing before it can be used in the model, or a method to estimate performance used instead. While pausing in the analysis to perform experiments to obtain necessary input data is not the most expedient solution to the challenge, it is the most rigorous answer. Estimating models are often limited by their level of accuracy, which can feed a sizeable amount of uncertainty into the task load model, and could propagate through in unknown ways.

2. Expected Value Models

The second type of model relies on the expected value to be gained from an agent performing a given task. This is a common type of quantitative model and can utilize a nearly limitless number of individual metrics. Each task performed by each agent is assessed to determine the benefit gained from that task-agent combination, and the cost of that combination. The tabulated difference between the benefit and the cost is the expected value. Evaluation is often represented in the form of statistics, where the probability of benefit and cost is used, with reference to probability of component or task failure. (Parasuraman[48] contains a more detailed set of different analysis schemes for expected value models).

Rodriguez[54] was the first to use this model to compute a composite task score as a relative value with respect to a reference (where 'reference' could be an agent, or a team configuration - the model can be used to compare performance between agents or between team configurations) to evaluate overall team performance. In this model, tasks and agents were assessed for relative task completion time and relative resource cost.

Dissimilar metrics could be used because the model resulted in a matrix of dimensionless parameters. Each agent's performance ratios and resource cost ratios were summed for the entire scenario considered. The performance ratios were summed for each agent such that an agent would have a single score that represented the difference between the relative benefit and the relative difference in cost of using the resources.

Rodriguez's model was structured such that each task primitive emphasized a different aspect of human performance (cognitive, motor, and sensory skills). This model calculated the value added by using a given team instead of the reference team. The selection of the reference does not affect the results - the analysis is relative, and the same relations would be achieved if a different reference was selected. In this paper, task difficulty included aptitude of a given system with respect to a reference, and a relative amount of power, mass, or other resource needed to implement the candidate system. The composite task score represented the ratio of value added by using a specified system or team instead of the reference.

Mann[39] suggested using MacKenzie's modification[38] to Fitts' law in information theory to use a less simplified version of Shannon's law of communication theory in Rodriguez's performance ratio equation. This equation change allowed performance, resource, and utility ratios that were nearly equal to and less than one to be used in Rodriguez's model. These values could occur in Rodriguez's model when a candidate agent or team configuration performs as well as (or better than) the reference system. Rodriguez recommended his equation be used with binary bit information theory, which required the $\log_2$ of each ratio to be used in calculations. The $\log_2$ of a value nearly equal to one or less than one will either result in a zero or a negative value. Translating that mathematical result into practical applications would result in the finding that an agent performing a task has zero (or negative) utility or performance advantage, and that the task performance required zero resources. This equation alteration desensitized Rodriguez's equation so that the relative value of an agent could not mathematically result in zero.

Tunstel[67] applied Rodriguez's methodology for composite task scores to monitor the navigation performance of the Mars Exploration Rovers, Spirit and Opportunity. Kaupp[34,35] used a composite task score's value of information to determine an appropriate level of autonomy for navigating a maze. Using the HURON software[20,21] to plan a lunar mission, Elfes[17] sought to maximize the value per cost of a mission by analyzing the required input effort to the expected output benefit of each task.

Expected value models are structured to facilitate comparison between different candidate systems. They require approximately the same input data as task load models, but the data is used as relative values rather than directly. This provides more robustness to uncertainty in the data itself. These models can also be run with only a rough estimation of how systems or agents perform in relation to each other (if task and resource data does not presently exist), removing the need to run more experiments before beginning analysis. The application of the models are more flexible than task load models: rather than necessitating a quantized overall score (in which case the range of scores would need to be analyzed to determine if a gap between two values is significant or negligible), the scores are immediately referenced for relative comparison. Expected value models are best applied when the designer seeks to select a configuration from a set of possibilities. On the other hand, if any other summary data is sought (comparison of the types of workload, distribution of the workload, or duration of larger workload quantities), futher analysis capabilities would need to be built into the models, or a different model would need to be used.

3. Cognitive Systems

The goal of the third type of model is to evaluate the effect of tasks on human mental processes. The methods and systems of information processing fall into this domain. This model applies to HRTs because it is well adapted to cover not only human information processing about individual tasks, but also the coordination demands of working on a cooperative task (either with a human or a robot).

These models are best used for detailed analysis of the effect of different autonomy levels, situation awareness, and communication architectures on a human's mental workload. Due to their emphasis (if not exclusively) on cognitive processes only, these models do not assess physical performance. System level HRT overall performance evaluations generally do not benefit from this type of model, but these models can be invaluable in the development and verification that a misions scenario requires a feasible workload level.

4. State Transition Networks

This type of model attempts to frame a sequence of tasks into the agent states required to perform them (e.g. observer, active physical participant, standby, etc.). Agents only change from one state to another when

a task assigned to them requires a different form of involvement. Built into the model is the assumption that minimizing the number of transitions will reduce the workload required to perform a task list in its given order and improve the efficiency of overall mission performance. This type of analysis is valuable for tracing what causes the changes to an agents' state throughout a scenario, and the direct effect between the actions. Heger[28] used a state transition matrix to map the probabilities of each agent's success or failure at a given task, with a composite task score accumulating for each task attempted. The transitions in this model refered to the operator yielding control to the robot, adjusting the autonomy of a HRT to have the greatest probability of success, and a timing metric was computed to obtain the expected duration of the task list.

State transition models facilitate analysis of human attention and mental requirements for task performance. Rather than assessing from a workload perspective (as done in cognitive systems models), these models can easily incorporate physical requirements and coordination requirements into the performance analysis. Composite scores can be obtained to compare how often and what type of mental and physical mode transfer must occur during a mission, but the emphasis on the number of transitions does not necessarily correlate to better overall HRT performance. The scores instead reveal the influence of switching states on each different agent during task performance. A different type of model would be needed to represent a team's overall task performance.

Shah[61] described the initial development of a different kind of framework to walk a designer through the process of selecting a HRT and the task allocation process. Rather than using a composite score, the evaluation of task performance relied on specific methods to link together common metrics for a space exploration task scenario. Each section of the paper provided a literature review of commonly used methods. The framework proposed unifying the process of selecting a team design and assessing a common set of task-based metrics to allow comparison of disparate team performance. This is a generalized framework that is an option to be used either instead of the quantitative models described in this paper, or in addition to.

VI. Implementation Challenges and Discussion

Generating quantitative overall performance models for a HRT has been a significant challenge. Designers have offered frameworks to direct analysis to a set of common metrics to facilitate comparison between disparate team configurations. The majority of these frameworks lead to developing a quantitative composite score model to evaluate a HRT's overall performance, and their generalized procedure is summarized in Figure 2.

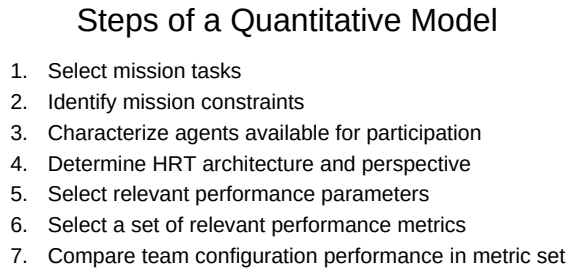

Figure 2. General Procedure For a Quantitative HRT Performance Model

Of the four types of models described by Parasuraman[48,49] , deciding which best reflects the desired analysis perspective for a given application can be difficult. The different types of quantitative models are not mutually exclusive, however. In fact, it can be productive to combine features from several of them to fully characterize a HRT's interactions. It may be advantageous to analyze the expected value of a team configuration based on a cognitive model, or to structure a task-load model into a state transition matrix. Kaupp[34] presented a method to select the autonomy level prior to deployment of a HRT that produced the highest team effectiveness for task-oriented information exchange on a HRT. This analysis included actual robot performance data, and resource costs. A composite task score was developed that included execution time, pairwise comparison of effort, value added, and a weighting for the final composite. Selecting the type of model or combination of models will have a significant affect on the results, and can either highlight or

obscure significant interactions in an application.

A. Input Data Requirements

Once the type of model has been decided, the kind of input data and level of accuracy of the data required should be addressed. The level of accuracy of the input data could have a significant effect on the predictive performance of the model. Low-fidelity estimations feed uncertainty into the model which, depending on depedencies and correlations between parameters, could propagate through the model in unknown ways.

To determine the relevant input data to the model, it is necessary to select the set of metrics that will comprehensively reflect the overall performance of the HRT. A designer must select a large enough set of metrics to cover relevant aspects of the agent interaction, but must also avoid selecting too many metrics, which runs the risk of generating false correlations by analyzing the same effect from multiple angles. It is at this point in the model-creation process that the designer must input details about the application, including the relevant subtask performance parameters, and the architecture details that facilitate the HRT's task completion. As discussed by Parasuraman (section 3.7 in[48]), the price of creating quantitative models of the parameters and properties of a HRT is a loss in generality of the resulting model.

B. Including Correlation Effects

The next challenge to implementing a quantitative model is explicitly framing the performance metrics in mathematical expressions that do not over-simplify the situation. For example, it would greatly simplify the modelling process to exclude environmental parameters from an analysis of how well one agent reliably follows the cooperative team plan. It could be assumed that an agent's comprehension of the task plan, and ability to direct its own efforts to achieve the desired goal would have a greater influence on the agent's successful performance than whether the terrain is grass, cement, or had vertical displacement. This simplified analysis would be fairly accurate in the majority of modeled cases, but it would diverge from the true behavior in cases where the environment has a significant effect on the agent performance. It is that very divergence that would be invaluable to have reflected in the model. It is necessary to ensure that the model includes all relevant dependent correlations and parameters to accurately reflect the true system performance.

C. Importance of Each Performance Metric

After selecting the metrics to be included in the model, a designer must then link them to facilitate ease of comparison between results from different team configurations. Most of the composite task scores were computed by a linear summation of the individual performance scores (or ratios, or expected values) for subtasks. In some of the implementations, weightings were used in the summation to include a measure of relative importance between the performance metrics. These weightings will make a significant difference on the optimal designs returned from the system analysis. In essence, these calculations use the individual performance metrics as objective functions, transforming the evaluation of the HRT's performance into a constrained multi-objective optimization problem.

By combining the metrics with weightings into a single summation, however, the designer applies a common optimization technique to transform a multi-objective optimization problem into a single-objective optimization problem, which is much simpler to solve. To do this, knowledge of designer preferences between metrics or an estimate of relative weightings between the metrics is needed. These have either been subjectively selected or all metrics assumed to be of equivalent value in the past. There has been significant effort in an attempt to make weightings selection more objective, but this is still a very new field. Rohrmuller[55] presented a quantitative method to map the probabilistic interdependencies between individual performance metrics, and use them and their provided data to determine relationships between the metrics, and to compile them into a composite score that could be used to predict system performance and optimize performance parameters.

For further guidance on creating this objective method, the authors suggest the extensive literature search by Bobko[5] which provided a cross discipline analysis that considered the validity of using weightings to aggregate sub-scores to represent a data group. Several methods were described and the reader was directed towards other references for more detail. If an objective method of selecting these weightings could be found, it would add rigor to the methodology of using composite scores to optimize a team configuration.

VII. Conclusion

Choosing optimum configurations of humans and robots for future missions, allocating tasks among disparate heterogeneous agents, measuring task performance, or even agreeing on what parameters are the most critical, are all topics that have been examined in detail, with no resolution as to generally-recognized agreements on approaches.

The focus of this paper is not to attempt to solve these complex problems, but to present a synoptic examination of the state of this field to date. It is the intention of the authors to use this detailed literature review in the future to inform their attempt to integrate the disparate elements of the current field in search of a supportable rationale for designing future teams of humans and multiple heterogeneous robots for space missions, based on a selection of metrics which can be justified as representing the critical parameters for mission success.

Future work will continue to prepare the groundwork for the synthesis of existing methods to compare HRTs and measure the differences in their interactions. A universal quantitative team performance model with a wide range of applications still eludes the research community, but work continues to progress towards this goal.

Acronyms

CD – coordination demand: a measure of the resources required to maintain synchronization with a robot (task performance metric)

CTR – critical time ratio: ratio of the time that a robot is performing mission-critical tasks to the total amount of time that a robot is actively engaged in tasks (task performance metric)

EVA - extra-vehicular activity

FAA – Federal Aviation Administration

FAT – false alarm time: the time that is spent identifying a false alarm and recovering from the delay (task performance metric)

FO – fan out: the number of robots that an operator can control simultaneously without performance degradation (task performance metric)

HRI – human robot interaction

HRI/OS – Human Robot Interaction Operating System (JPL software package)

HRT – human and robot team that cooperatively completes a mission

HURON – Human-Robot Task Network Optimization (JPL software package)

IE – interaction effort: a measure of the human's effort during an interaction (task performance metric)

MITPAS – Mixed Initiative Team Performance Assessment System (Perceptronics, Inc. software package, US Army)

MTBI – mean time between interventions: the average time a robot can function without requiring human assistance (task performance metric)

MTCI – mean time completing an intervention: the average time required for a human to complete an intervention and return robot to productivity (task performance metric)

NT – neglect time: the amount of time that a robot can function on its own without requiring human assistance (task performance metric)

OT – occupied time: differentiates between CD and demands of interacting with a queue of several independent robots (task performance metric)

RAD – robot attention demand: the fraction of the total task time that the robot requires attention (task performance metric)

SC – situation coverage: percentage of total tasks in which a robot understands the directive (task performance metric)

UI – user interface

WEI – work efficiency index: represents the ratio of productive time to overhead time that occurs in an agents' task schedule (task performance metric)

References

[1]J.A. Arnold. Towards a framework for architecting heterogeneous teams of humans and robots for space exploration. Master's thesis, Massachusetts Institute of Technology, 2006.

[2]A. Bechar, Y. Edan, and J. Meyer. Optimal collaboration in human-robot target recognition systems. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 5, pages 4243–4248. IEEE, 2007.

[3]A. Bechar, J. Meyer, and Y. Edan. An objective funcion to evaluate performance of human-robot collaboration in target recognition tasks. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(6):611–620, 2009.

[4]L. Billman and M. Steinberg. Human system performance metrics for evaluation of mixed-initiative heterogneous autonomous systems. In *Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems*, pages 120–126. ACM, August 2007.

[5]P. Bobko, P. Roth, and M. Buster. the usefulness of unit weights in creating composite scores: a literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, 10(4):689–709, 2007.

[6]D.J. Bruemmer, D.A. Few, R.L. Boring, J.L. Marble, M.C. Walton, and C.W. Nielsen. Shared understanding for collaborative control. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Sytems and Humans*, 35(4):494–504, July 2005.

[7]J.L. Burke, R.R. Murphy, D.R. Riddle, and T. Fincannon. Task performance metrics in human-robot interaction: Taking a systems approach. *Performance Metrics for Intelligent Systems*, 2004.

[8]J.W. Crandall and M.L. Cummings. Developing performance metrics for the supervisory control of multiple robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 33–40. ACM, 2007.

[9]J.W. Crandall and ML Cummings. A predictive model for human-unmanned vehicles systems. Technical Report HAL2008-05, Massachusetts Institute of Technology, 2008.

[10]J.W. Crandall, M.A. Goodrich, D.R. Olsen, and C.W. Nielsen. Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Sytems and Humans*, 35(4):438–449, July 2005.

[11]J.W. Crandall, C.W. Nielsen, and M.A. Goodrich. Towards predicting robot team performance. In *Systems, Man and Cybernetics, 2003 IEEE International Conference on*, volume 1, pages 906–911. IEEE, 2003.

[12]ML Cummings, P. Pina, and J.W. Crandall. A metric taxonomy for supervisory control of unmanned vehicles. In *AUVSI*, 2008.

[13]K. Dautenhahn and I. Werry. A quantitative technique for analysing robot-human interactions. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002.

[14]E. De Visser, R. Parasuraman, A. Freedy, E. Freedy, and G. Weltman. A comprehensive methodology for assessing human-robot team performance for use in training and simulation. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, volume 50, pages 2639–2643. Human Factors and Ergonomics Society, 2006.

[15]B. Donmez, P. Pina, and M.L. Cummings. Evaluation criteria for human-automation performance metrics. In *Proceedings of the Performance Metrics for Intelligent Systems Workshop*, 2008.

[16]M.R. Elara, C.A.A. Calderon, C. Zhou, and W.S. Wijesoma. False alarm demand: A new metric for measuring robot performance in human robot teams. In *Autonomous Robots and Agents, 2009. ICARA 2009. 4th International Conference on*, pages 436–441. IEEE, 2009.

[17]A. Elfes, C.R. Weisbin, H. Hua, J.H. Smith, J. Mrozinski, and K. Shelton. The huron task allocation and scheduling system: Planning human and robot activities for lunar missions. In *Automation Congress, 2008. WAC 2008. World*, pages 1–8. IEEE, 2008.

[18]C.A. Ellis, S.J. Gibbs, and G. Rein. Groupware: some issues and experiences. *Communications of the ACM*, 34(1):39–58, 1991.

[19]T. Fong, A. Abercromby, M.G. Bualat, M.C. Deans, K.V. Hodges, J.M. Hurtado, R. Landis, P. Lee, and D. Schreckenghost. Assessment of robotic recon for human exploration on the moon. *Acta Astronautica*, 67(9-10):1176–1188, November–December 2010.

[20]T. Fong, I. Nourbakhsh, C. Kunz, L. Fluckiger, J. Schreiner, R. Ambrose, R. Burridge, R. Simmons, L.M. Hiatt, A. Schultz, et al. The peer-to-peer human-robot interaction project. *Space*, 6750(AIAA 2005-6750), September 2005.

[21]T. Fong, J. Scholtz, J.A. Shah, L. Fluckiger, C. Kunz, D. Lees, J. Schreiner, M. Siegel, L.M. Hiatt, and I. Nourbakhsh. A preliminary study of peer-to-peer human-robot interaction. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 4, pages 3198–3203. IEEE, 2007.

[22]A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman. Measurement of trust in human-robot collaboration. In *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*, pages 106–114. IEEE, 2008.

[23]B.P. Gerkey and M.J. Mataric. A formal analysis and taxonomy of task allocation in multi-robot systems. *International Journal of Robotics Research*, 23(9):939–954, 2004.

[24]D.F. Glas, T. Kanda, H. Ishiguro, and N. Hagita. Simultaneous teleoperation of multiple social robots. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 311–318. ACM, 2008.

[25]M. Goodrich and D. Olsen. Seven principles of efficient human robot interaction. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pages 3943–3948, 2003.

[26]M.A. Goodrich, T.W. McLain, J.D. Anderson, J. Sun, and J.W. Crandall. Managing autonomy in robot teams: observations from four experiments. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 25–32. ACM, March 2007.

[27]S.G. Hart and L.E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human Mental Workload*, 1:139–183, 1988.

[28]F. Heger and S. Singh. Sliding autonomy for complex coordinated multi-robot tasks: analysis and experiments. In *Proceedings of Robotics: Science and Systems*, August 2006.

[29]A. Howard. A synergistic approach for maximizing human-automation system performance. Draper fiscal year 2006 final report, Georgia Institute of Technology, July 2006.

[30]A. Howard. A systematic approach to predict performance of human-automation systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 37(4):594–601, July 2007.

[31]A. Howard and G. Cruz. Adapting human leadership approaches for role allocation in human-robot navigation scenarios. In *Automation Congress, 2006. WAC'06. World*, pages 1–8. IEEE, 2007.

[32]J.H. Hwang, K.W. Lee, and D.S. Kwon. A formal method of measuring interactivity in hri. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 738–743. IEEE, 2007.

[33]B. Kannan and L.E. Parker. Fault-tolerance based metrics for evaluating system performance in multi-robot teams. In *Proceedings of Performance Metrics for Intelligent Systems Workshop*, 2006.

[34]T. Kaupp and A. Makarenko. Measuring human-robot team effectiveness to determine an appropriate autonomy level. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 2146–2151. IEEE, 2008.

[35]T. Kaupp, A. Makarenko, and H. Durrant-Whyte. Human-robot communication for collaborative decision making - a probabilistic approach. *Robotics and Autonomous Systems*, 58:444–456, 2010.

[36]J. Keller. Human performance modeling for discrete-event simulation: human performance modeling for discrete-event simulation: workload. In *Proceedings of the 34th Conference on Winter Simulation: Exploring New Frontiers*, pages 157–162. Winter Simulation Conference, 2002.

[37]A. Lampe and R. Chatila. *Performance measure for the evaluation of mobile robot autonomy*. Proceedings of the 2006 IEEE International Conference on Robotics and Automation, May 2006.

[38]I.S. MacKenzie. A note on the information-theoretic basis for fitts' law. *Journal of Motor Behavior*, 1989.

[39]G.A. Mann. Quantitative evaluation of human-robot options for maintenance tasks during analogue surface operations. In C. Pain, editor, *Proceedings of the 8th Australian Mars Exploration Conference*, pages 26–34, 2008.

[40]J.L. Marble, D.J. Bruemmer, D.A. Few, and D.D. Dudenhoeffer. Evaluation of supervisory vs. peer-peer interaction with human-robot teams. In *Proceedings of the 37th Annual Hawaii International Conference on Systems Sciences*, pages 4–8, 2004.

[41]C.A. Miller and R. Parasuraman. Who's in charge?: Intermediate levels of control for robots we can live with. *Systems, Man, and Cybernetics, 2003. IEEE International Conference on*, 1:462–467, 2003.

[42]C. E. Nehme. *Modeling Human Supervisory Control in Heterogeneous Unmanned Vehicle Systems*. PhD thesis, Massachusetts Institute of Technology, February 2009.

[43]U. Nehmzow. Quantitative analysis of robot-environment interaction–towards scientific mobile robotics. *Robotics and Autonomous Systems*, 44:55–68, 2003.

[44]D.R. Olsen and M.A. Goodrich. Metrics for evaluating human-robot interactions. In *Proceedings of PERMIS*, 2003.

[45]D.R. Olsen, B. Wood, and J. Turner. Metrics for human driving of multiple robots. In *International Conference on Robotics and Automation*. IEEE, 2004.

[46]D.R. Olsen and S.B. Wood. Fan-out: measuring human control of multiple robots. In *Proceedings of the Special Interest Group on Computer Human Interaction Conference on Human Factors in Computing Systems*, pages 231–238. ACM, 2004.

[47]Y. Oren. Performance analysis of human-robot cooperation in target recognition tasks. Master's thesis, Ben-Gurion University of the Negev, Aug. 2008.

[48]R. Parasuraman. Designing automation for human use: empirical studies and quantitative models. *Ergonomics*, 43(7):931–951, July 2000.

[49]R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Sytems and Humans*, 30(3):286–297, May 2000.

[50]P. Pina, ML Cummings, J.W. Crandall, and M.D. Penna. Identifying generalizable metric classes to evaluate human-robot teams. In *3rd Annual Conference on Human-Robot Interaction: Metrics for Human-Robot Interaction Workshop*, 2008.

[51]P. Pina, B. Donmez, and M.L. Cummings. Selecting metrics to evaluate human supervisory control applications. HAL Report HAL2008-04, Humans and Automation Laboratory, Massachusetts Institute of Technology, 2008.

[52]S.S. Ponda, H.L. Choi, and J.P. How. Predictive planning for heterogeneous human-robot teams. *AIAA Infotech@ Aerospace*, 2010.

[53]M. Prewett, R. Johnson, K. Saboe, L. Elliott, and M. Coovert. Managing workload in human-robot interaction: a review of empirical studies. *Computers in Human Behavior*, 26(5):840–856, September 2010.

[54]G. Rodriguez and C.R. Weisbin. A new method to evaluate human-robot system performance. *Autonomous Robots*, 14(2-3):165–178, 2003.

[55]F. Rohrmuller, O. Kourakos, M. Rambow, D. Brscic, D. Wollherr, S. Hirche, and M. Buss. Interconnected performance optimization in complex robotic systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4113–4118. IEEE, 2010.

[56]J.A. Saleh and F. Karray. Towards generalized performance metrics for human-robot interaction. In *Autonomous and Intelligent Systems, 2010 International Conference on*, pages 1–6. IEEE, 2010.

[57]J. Scholtz. Theory and evaluation of human robot interactions. In *Proceedings of the 36th Hawaii International Conference on System Sciences*, 2002.

[58]J. Scholtz, B. Antonishek, and J.D. Young. Implementation of a situation awareness assessment tool for evaluation of human-robot interfaces. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Sytems and Humans*, 35(4):450–459, July 2005.

[59]D. Schreckenghost, T. Fong, H. Utz, and T. Milam. Measuring robot performance in real-time for nasa robotic reconnaissance operations. In *Proceedings of NIST PERMIS Workshop*, 2009.

[60]D. Schreckenghost, T. Milam, and T. Fong. Ai space odyssey: Measuring performance in real time during remote human-robot operations with adjustable autonomy. *IEEE Intelligent Systems*, 25(5):36–44, 2010.

[61]J.A. Shah, J.H. Saleh, and J.A. Hoffman. Review and synthesis of considerations in architecting heterogeneous teams of humans and robots for optimal space exploration. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 37(5):779–, September 2007.

[62]J.A. Shah, J.H. Saleh, and J.A. Hoffman. Analytical basis for evaluating the effect of unplanned interventions on the effectiveness of a human-robot system. *Reliability Engineering and System Safety*, 93(8):1280–1286, 2008.

[63]S.M. Singer and D. Akin. *Role definition and task allocation for a cooperative EVA and robotic team*. SAE International, 2009.

[64]B. Stanton, B. Antonishek, and J. Scholtz. Development of an evaluation method for acceptable usability. In *Proceedings of PERMIS*, 2006.

[65]A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 33–40. Association for Computing Machinery, 2006.

[66]J.G. Trafton, N.L. Cassimastis, M.D. Bugajska, D.P. Brock, F.E. Mintz, and A. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Sytems and Humans*, 35(4):460–470, July 2005.

[67]E. Tunstel. Operational performance metrics for mars exploration rovers. *Journal of Field Robotics*, 24(8-9):651–670, 2007.

[68]H. Wang, M. Lewis, P. Velagapudi, P. Scerri, and K. Sycara. How search and its subtasks scale in n robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 2009.

[69]J. Wang and M. Lewis. Assessing cooperation in human control of heterogeneous robots. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, March 2008.

[70]H.A. Yanco and J.L. Drury. A taxonomy for human-robot interaction. In *Proceedings of the AAAI Fall Symposium on Human-Robot Interaction*, 2002.

[71]H.A. Yanco, J.L. Drury, and J. Scholtz. Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition. *Human-Computer Interaction*, 19(1):117–149, 2009.