Identifying Generalizable Metric Classes to Evaluate Human-Robot Teams

P. Pina
Massachusetts
Institute of Technology
Cambridge, MA

M. L. Cummings
Massachusetts
Institute of Technology
Cambridge, MA

J. W.Crandall
Massachusetts
Institute of Technology
Cambridge, MA

M. Della Penna
Delft University of
Technology
The Netherlands

ABSTRACT

In this paper, we describe an effort to identify generalizable metric classes to evaluate human-robot teams. We describe conceptual models for supervisory control of a single and multiple robots. Based on these models, we identify and discuss the main metric classes that must be taken into consideration to understand team performance. Finally, we discuss a case study of a search and rescue mission to illustrate the use of these metric classes to understand the different contributions of team performance

Categories and Subject Descriptors

J.7 [**Computers in Other Systems**]: Command and Control;
H.5.2 [**User Interfaces and Presentation**]: Evaluation/ methodology

General Terms

Measurement, Performance, Experimentation, Human Factors

Keywords

Metrics, Human-Robot Teams, Performance, Supervisory Control

1. INTRODUCTION

Mission effectiveness is the most popular metric to evaluate the performance of human-robot teams. However, frequently this metric is not sufficient to understand team performance issues and to identify design improvements, and additional metrics are required.

Despite the importance of selecting the right metrics, few general guidelines that apply to a wide range of human-robot applications are available in the literature. In many cases, researchers rely on their own experience, selecting metrics they have used previously. Alternatively, other experiments measure every system parameter to ensure that every aspect of system performance is covered. These approaches lead to ineffective metrics and excessive experimental and analysis costs. Moreover, existing metrics for evaluating human-robot teams are usually application-specific, which makes comparison across applications difficult.

The goal of this research is to provide general guidelines for metric selection that are applicable to any human-robot team operating under a supervisory control paradigm. We believe that identifying generic metric classes that organize the different types of metrics available will help researchers select a robust set of metrics that provide the most value for their experiments and allow comparison with others. Metrics may still be mission-specific, however metric classes are generalizable across different missions. In the context of this paper, a metric class is defined as the set of metrics that quantify a certain aspect or component of a system.

The idea of developing a toolkit of metrics and identifying classes to facilitate comparison of research results has already been discussed by other authors. For example, Olsen and Goodrich proposed four metric classes to measure the effectiveness of robots: task efficiency, neglect tolerance, robot attention demand, and interaction effort [1]. This set of metrics measures the individual performance of a robot, however, a particular robot performance does not necessarily imply a level of human performance. Since human cognitive limitations often constitute a primary bottleneck for human-robot team performance, a metric framework that can be generalized should also include cognitive metrics to understand what drives human behavior and cognition.

In line with this idea of integrating human and robot performance metrics, Steinfeld et al. suggested identifying common metrics for human-robot interaction in terms of three aspects: human, robot, and the system [2]. Regarding human performance, they discussed three main metric categories: situation awareness, workload, and accuracy of mental models of device operations. This work constitutes an initial step towards developing a metric toolkit, however it still presents some limitations. On the one hand, this framework suffers from a lack of metrics to evaluate collaboration effectiveness among humans and among robots. On the other hand, a more comprehensive discussion on human performance is still required. For example, the authors discuss trust as a task-specific metric for social robots but it is not included as a common metric required to evaluate operator performance. We believe that operators' trust in robot behavior is often a key factor in team performance.

The research presented in this paper builds upon previous efforts conducted by Crandall and Cummings [3]. It refines, expands, and generalizes the set of metric classes already identified for human-robot teams consisting of a single human and multiple robots. The paper builds a conceptual model for human supervisory control of multiple robots. Then metric classes are identified from this model. Finally, a case study on a search and rescue mission is discussed to illustrate some of the proposed metric classes.

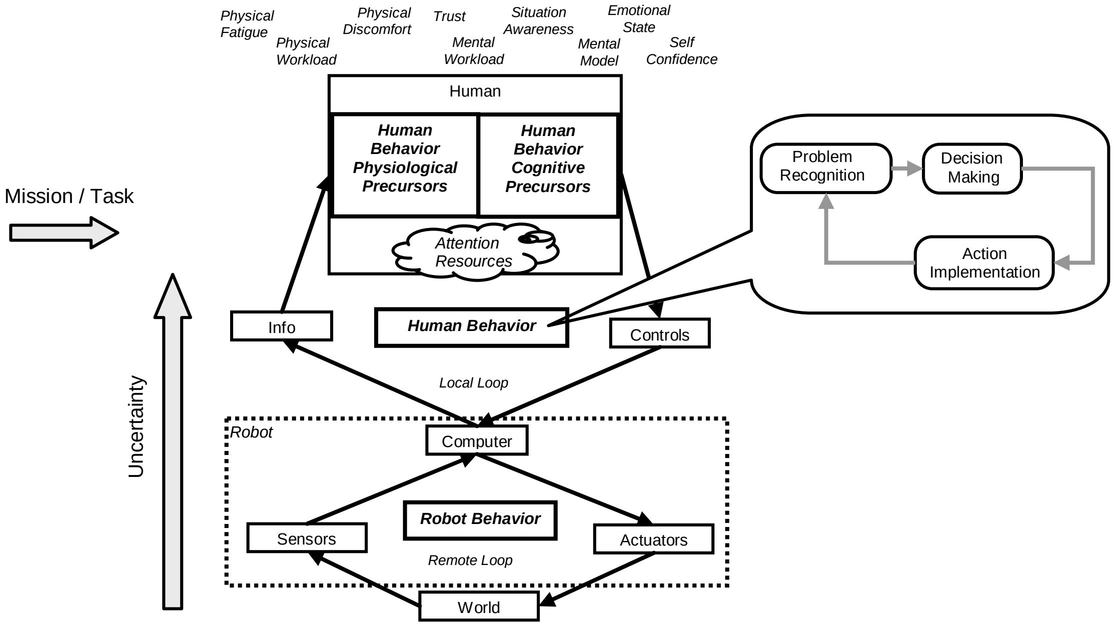

Figure 1. Conceptual Model of Human-Robot Interaction in Supervisory Control.

2. CONCEPTUAL MODEL

This section presents and discusses our conceptual models of human supervisory control of robots, including a single operator controlling a single robot, a single operator controlling multiple robots, and multiple operators controlling multiple robots.

2.1 Supervisory Control of a Single Robot

"Supervisory control means that one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment [4]." Most human-robot teams operate under a human supervisory control paradigm where robots have a certain degree of autonomy and the human guides them, monitors their performance, and intervenes when needed. Examples of this are found across several domains and applications: surveillance and target identification for military operations, health care applications such as mobility assistance and therapy, rock sampling for geology research, or other logistic applications such as personnel or material delivery.

All these examples can be conceptually represented by the model shown in Fig. 1. This model is composed of four interrelated main elements: robot behavior, human behavior, human behavior cognitive precursors, and human behavior physiological precursors. We believe that these four elements delineate the main metric classes for single operator-single robot teams. In addition to these four elements, two other concepts are represented in Fig. 1: uncertainty, and the mission or the task. Uncertainty refers to the uncertainty associated with sensors (e.g., accuracy) and actuators (e.g., lag), displays (e.g., transforming 3D information into 2D information), and the real world. This uncertainty propagates through the system reaching one or more operators who adapt their behavior to the uncertainty level by applying different cognitive strategies.

Regarding the mission or the task imposed on the operator, human behavior and system performance depend on the nature of the tasks. High structured tasks, those that can be planned in advanced and are procedurally-driven, are very different, from a human perspective, from those that have low structure levels, which are generally emergent tasks that require solving a new problem under time-pressure. Human-robot team performance can only be understood if considered in the context of the mission and the task.

The goal of this paper is to develop a general framework for the analysis of human-robot team performance. However, our focus is on those metrics of human behavior efficiency and human behavior precursors, rather than metrics of robot behavior efficiency. The fact that many human-robot teams are remote makes it essential to measure the human component. Operators who remotely operate a robot do not physically perceive the interaction of the robot with the real world. This can have a negative impact on situation awareness and human trust, which in turn can affect performance.

2.1.1 Robot & Human Behavior Efficiency

Robot and human behavior are represented by the two control loops shown in Fig.1: the human control loop and the robot control loop. The operator receives feedback on robot and mission performance, and adjusts robot behavior through controls if required. The robot interacts with the real world through actuators and collects feedback on mission performance through sensors. The evaluation of team performance requires an understanding of both control loops. The rest of this section focuses on human behavior.

Human behavior, in the context of Fig. 1, refers to the decisions made and actions taken by the human while controlling the robot. The model presented in Fig. 1 categorizes human behavior in terms of problem recognition, decision making, and action implementation. These three categories are based on the four-stage model of human information processing described by Parasuraman, Sheridan, and Wickens: 1) information acquisition, 2) information analysis, 3) decision and action selection, and 4) action implementation [5]. Our model merges the stages of information acquisition and analysis into the problem recognition category. Acquisition and analysis of information are often hard to differentiate, and the human ability to recognize problems is a more valuable metric for our purposes. Thus, understanding human performance requires evaluating each one of the three categories defined by our model.

Human-computer interactions (HCIs) are the observable outputs of human decisions, and they are commonly used to measure human behavior efficiency. Based on our model, these interactions should also be analyzed in terms of problem recognition (e.g., access to information about the environment dynamics), decision making (e.g., use of what-if functionalities to explore consequences of actions), and action implementation (e.g., entering new coordinates for a robot's destination). Such decomposition enables a more comprehensive evaluation of team performance. However, disaggregating HCIs may not always be possible.

In addition to human efficiency for problem recognition, decision making, and action implementation, human attention allocation is a key component of human behavior. The evaluation of attention resource allocation helps in the understanding of operators' strategies and priorities. Operators have limited attention resources that need to be shared between multiple tasks [6]. Although as seen in Fig. 1, one single robot is controlled, the operator still performs multiple tasks such as monitoring the dynamics of the environment, identifying emergent events, monitoring robot health, or executing manual control of the robot. How humans sequence and prioritize these multiple tasks provides valuable insights into the system.

2.1.2 Human Behavior Cognitive and Physiological Precursors

Evaluating human observable behavior can still be insufficient since all mental processes do not have immediate and observable outcomes. The evaluation of human performance requires understanding what motivates the behavior and the cognitive processes behind it. Human behavior is driven by high level cognitive constructs and processes such as mental models[1] and situation awareness[2] (SA). For our discussion, mental models refer to long-term knowledge, whereas SA reflects dynamic knowledge. Understanding human mental models is important because ideally, an interface design should be consistent with people's natural mental models about computers and the environment [10]. Poor SA or lack of understanding of a dynamic environment, when performing complex cognitive tasks, can have dramatic consequences such as the incident at Three Mile Island [11].

Mental models and SA are not the only human behavior cognitive precursors. In the context of this paper, human behavior cognitive precursors refer to cognitive constructs or processes that existed or occurred before a certain behavioral action was observed. Human trust in the robots, mental workload, and operator emotional state are other examples of cognitive constructs and processes that can also cause certain human behaviors.

Furthermore, physiological processes can reflect physical states such as fatigue, or physical discomfort which can also motivate certain human attitudes.

2.1.3 Conclusions

Our model represents the need for evaluating four main elements to understand the performance of a single operator-single robot team: robot behavior, human behavior, human behavior cognitive precursors, and human behavior physiological precursors. These four elements are all interrelated. For example, events in the real world are captured by the robot sensors and presented to the human operator through the display. Modifications on the display can affect human attention allocation and SA, which in turn will result in changes in HCI patterns, which can ultimately affect robot performance. Understanding system performance implies understanding the relations among these elements.

2.2 Supervisory Control of Multiple Independent or Collaborative Robots

The previous section discusses a model for one operator-one robot team, but operators can simultaneously control multiple robots. In order to expand the model in Fig. 1, we consider two different scenarios: a) multiple robots performing independent tasks, and b) multiple robots performing collaborative tasks. In this paper collaboration between robots means two or more robots working together to accomplish a shared goal under human supervision.

In the case of independent robots, servicing robot 1 and robot 2 are two independent tasks. The operator monitors the environment and the robots, decides on which one to focus his/her attention, interacts with that robot, and returns to monitoring or decides to service another robot. While servicing one of the robots, the operator behaves similarly as if he/she supervised only one single robot. Our model assumes that the operator does not service multiple robots in parallel. This assumption is based on the limited human cognitive resources and the high task demands imposed by supervising complex and dynamic environments under time pressure. Figure 2 illustrates this model of human supervisory control of multiple independent robots.

[1] The phrase "mental models" refers to organized sets of knowledge about the system operated and the environment that are acquired with experience [7].

[2] SA is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [8]. In the context of human-robot teams, SA encompasses awareness of where each robot and team member is located and what they are all doing at each moment, plus all the environmental factors that affect operations [9].

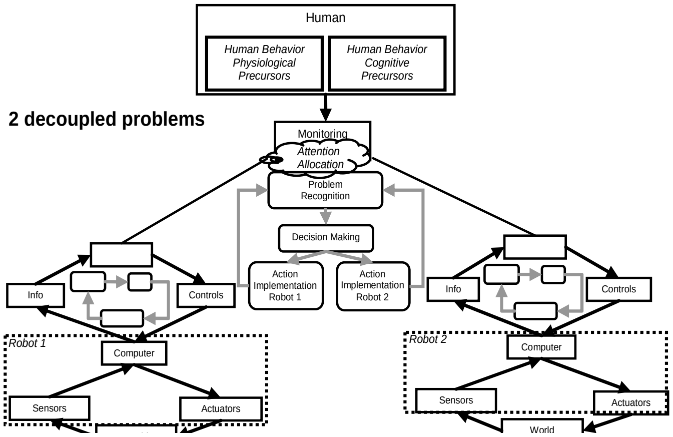

Figure 2. Supervisory Control of Independent Robots.

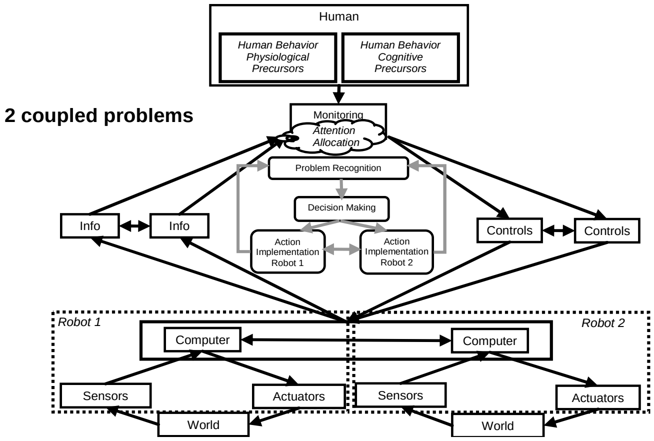

Figure 3. Supervisory Control of Collaborative Robots.

Multiple robots working together to achieve a common goal can autonomously collaborate or be manually coordinated by the operator. In the case of autonomous collaboration among robots without the possibility for human intervention, collaboration only occurs at the level of the robot behavior loop and the model in Fig. 2 is still valid. However, in the case of active human coordination, the operator executes two dependent tasks (i.e., servicing robots 1 and 2) that cannot be decoupled. Figure 3 illustrates the later model, where the control loops for robot 1 and robot 2 are not independent and separated entities. Controlling collaborative robots requires the operator to understand the consequences of an action across both control loops and to actively coordinate between them. For example, making a decision for robot 1 can involve acquiring and analyzing information related to robot 2, and implementing an action for robot 2 can require synchronizing it with another action for robot 1. Interfaces for collaborative robots should aggregate data from each control loop and display it so that the operator can easily understand the interconnections and the consequences of these dependencies.

In our previous example with independent robots, the three categories of human behavior (i.e., problem recognition, decision making, and action implementation) could be evaluated separately for robot 1 and robot 2. In the case of collaborative robots, these three categories have to be analyzed for both robots aggregately.

2.3 Human Collaboration in Supervisory Control of Multiple Robots

This section expands previous models to the case of multiple humans collaborating to control multiple robots. In these situations, system performance is directly linked to human collaboration. Our model considers two main dimensions of collaboration: team behavioral actions and team cognition. Figure 4 illustrates this model.

The evaluation of team behavioral actions consists of measuring both the efficiency of team coordination and the team efficiency in each of the three categories of human behavior (i.e., problem recognition, decision making, and action implementation). The team works together as a single entity to perform collaborative tasks so performance should be measured at the holistic level rather than aggregating team members' individual performance [12]. Team coordination comprises of written, oral, and gestural interactions among team members.

Team cognition refers to the thoughts and knowledge of the team. Measures of team cognition can be valuable in diagnosing team performance successes and failures, and identifying training and design interventions [12]. Moreover, efficient human collaboration is often shown to be related to the degree that team members agree on, or are aware of task, role, and problem characteristics [13]. Thus, team mental model and SA are two precursors of team performance.

The efficiency of the team mental model includes assessing the similarity, overlap, and consistency of the individual mental models. For team SA, both environment and team dynamics need to be understood. However, each member does not have to be aware of every change; the common picture is shared by the team, not necessarily by all its members individually. As Gorman et al. discuss, better performance does not necessarily mean all team members sharing a common picture [14]. In addition, evaluating team cognitive precursors can also include evaluating workload distribution and social patterns and roles within the team.

3. GENERALIZABLE METRIC CLASSES

Based on the models presented in this paper, we can infer six generalizable metric classes relevant for human-robot team evaluation. Examples of sub-classes are included in brackets.

- Mission Effectiveness (e.g., key mission performance parameters)
- Human Behavior Efficiency (e.g., attention allocation efficiency, problem recognition efficiency, decision making efficiency, action implementation efficiency)
- Robot Behavior Efficiency (e.g., error-proneness, robustness, autonomy, learnability, memorability)

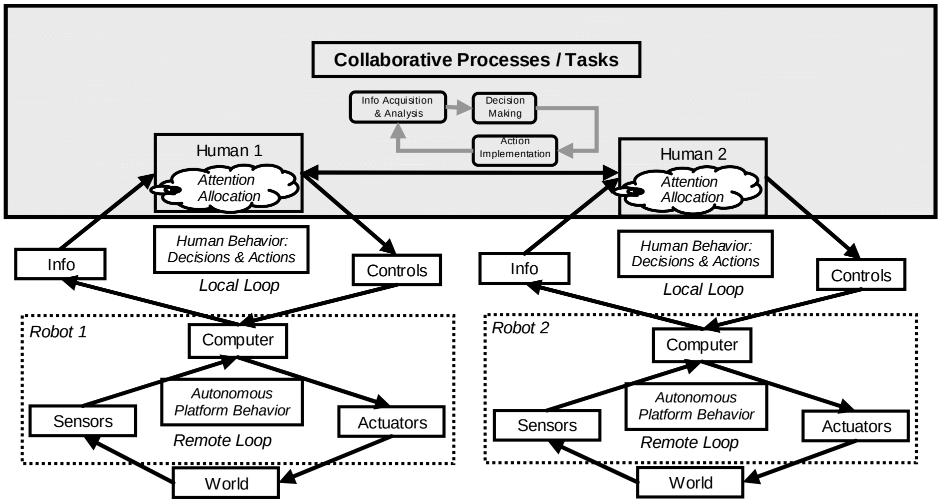

Figure 4. Human Collaboration in Supervisory Control of Robots.

- Human Behavior Cognitive Precursors (e.g., mental models, SA, mental workload, trust in automation, self-confidence, emotional state)
- Human Behavior Physiological Precursors (e.g., physical workload, physical comfort, physical fatigue)
- Collaborative Metrics

– Team Behavioral Action Efficiency (e.g., coordination efficiency, collaborative problem recognition efficiency, collaborative decision making efficiency, collaborative action implementation efficiency)

– Team Cognition Efficiency (e.g., team mental models, team SA, workload distribution, social patterns and roles)

– Robot Collaboration Efficiency

Evaluating the performance of the whole human-robot team requires applying metrics from each of these classes, but including metrics of every sub-class for every experiment can be inefficient and costly. As a rule of thumb, in addition to the more popular mission effectiveness and robot behavior efficiency metrics, incorporating at least one metric from the classes of human behavior efficiency, human behavior cognitive and physiological precursors, and collaborative metrics enables better team performance evaluation.

The next section discusses an experiment where a single human controlled multiple robots conducting a search and rescue mission. This study considered metrics for mission effectiveness, human behavior efficiency, and human behavior cognitive precursors. The value of incorporating metrics from each of these classes is discussed in the context of this experiment.

4. A CASE STUDY: SEARCH AND RESCUE MISSION

4.1 Experiment Description

In this experiment, a human participant teamed with multiple simulated robots to perform a search and rescue mission: removing objects from a maze[3] using different number of robots (2, 4, 6, or 8). The goal was a) to remove as many objects from the area as possible during an 8-minute session while b) ensuring that all robots were out of the maze when time expired. Collecting objects from the maze required the user to perform navigation and visual search tasks. First, the user assigned an object to the robot and the robot moved to that location. Second, the robot "picked up" the object, which in the experiment was simulated by the visual search of identifying a city on a map of the United States using Google Earth-style Software. Third, the user assigned one of the two maze exits to the robot and the robot carried the object out of the maze. The objects were randomly spread through the maze.

The maze was initially unknown, but the robots created and shared a map of the maze as they moved around it. Each robot could choose its path, choosing to explore an unknown path if it thought that path could possibly be shorter than the shortest known path to its user-specified destination. In addition, the robot would automatically choose an object or an exit after it had been waiting for a user-command for longer than 15 seconds. The user could at any moment redirect the robots to different locations by reassigning their destinations or rerouting them through a different path.

Sixteen people between the ages of 19 and 49 years old participated in the study. After completing a training and a comprehensive practice session, each subject participated in four 8-minute sessions, each with a different robot team size. The conditions of the study were randomized and counter-balanced. More details on the experimental setup can be found in [3].

4.2 Metrics Considered

This study measured metrics for mission effectiveness, human behavior, and human behavior cognitive precursors in an

[3] In this experiment, the routes within the maze are unknown but the locations of objects to rescue are known.

attempt to understand the final outcome of the mission, the decisions made and actions taken by the operator, and the causes driving those actions and decisions.

We believe that at least one metric from each class is necessary to understand team performance. However, we recommend for the human behavior efficiency class, both attention allocation efficiency and human efficiency in conducting mission's tasks should be measured because they represent different aspects of the system. In addition, if the mission is composed of tasks of different cognitive nature, one human behavior efficiency metric for each task is also recommended. For the human behavior cognitive precursor class, the number of metrics selected depends on the actual research question and experimental setting. For this experiment, we measured trust and mental workload because both factors can influence human use of automation (i.e., robots' autonomy) [15]. Automation mistrust, which refers to over-reliance on automation, occurs in decision making because humans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct [16]. This effect is known as automation bias.

We did not measure human behavioral actions separately for problem recognition, decision making, and action implementation because of the difficulty of distinguishing among these three categories in this particular testbed. No additional data that could support this analysis was recorded during the experiment.

This experiment did not measure metrics for human behavioral physiological precursors because with the 8-minute session time, these could not provide any meaningful insight. Collaborative metrics were also not considered since the focus was on single operator control, and robot efficiency was also not considered since they were simulated. Table 1 summarizes the metrics considered in this experiment.

Performance score, an indication of mission effectiveness, was defined as the total number of objects collected minus the number of robot lost (i.e., number of robots that did not get out of the maze when the 8-minute session expired).

HCIs were categorized in terms of robot navigation planning, robot navigation replanning, and visual search. The metrics selected were the time to complete a visual search, the time to assign a robot's destination, and the times to reroute a robot and reassign its destination.

The metric selected for attention allocation efficiency was the time required to decide which robot to service next, also known as the switching time. This metric included both the time it took for the user to decide which robot required his/her intervention, and the time required to select that robot on the display.

The frequency of overriding robot decisions was selected as an indication of operators' trust in robots. Finally, a five-point Likert scale was used to subjectively measure mental workload.

Table 1. Metrics Measured in the Case Study.

Metric Class	Selected Metric
Mission Effectiveness	• Performance score
Human Behavior Efficiency	• Average time to complete a visual search (indication of human efficiency in visual search) • Average time to complete a robot destination assignment (indication of human efficiency in planning robot navigation) • Average time to reroute a robot or reassign its destination (indication of human efficiency in replanning robot navigation) • Switching Time (indication of attention allocation efficiency)
Robot Behavior Efficiency	None
Human Behavior Cognitive Precursors	• Frequency of overriding robot decisions (indication of over-reliance on robots' autonomy) • Subjective rating of operator workload (indication of mental workload)
Human Behavior Physiological Precursors	None
Collaborative Metrics	None

4.3 Mission Effectiveness

Figure 5 shows the performance score as a function of the robot team size. A one-way ANOVA analysis showed that the robot team size significantly contributed to its variability (p-value = 0.018, R^2 = 15.31%). However, the R^2 of this model implies that it explained little of the performance variability. The Tukey test showed only difference in workload for 2 robots as compared to 8 robots.

Thus, evaluating performance in terms of robot team size does not provide much information, which confirms that additional metrics are required to really understand what happened in this experiment.

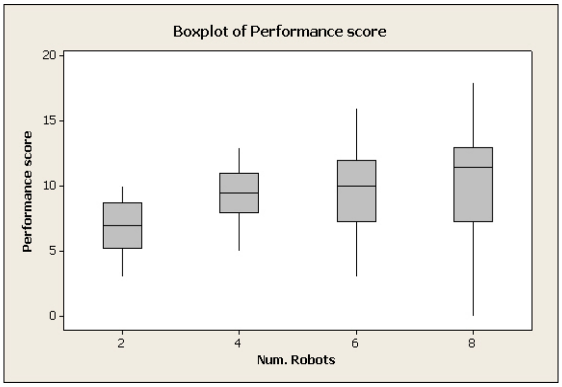

Figure 5. Performance Score vs. Size of the Robot Team.

4.4 Human Behavior Efficiency

Results suggest that the faster the subject completed a visual task, the higher the performance score (Pearson correlation = -0.594, p-value < 0.001). Results also suggest that subjects who were fast performing the visual search were also fast when selecting robot destinations (Pearson correlation = 0.479, p-value < 0.001).

Regarding navigation tasks, the average time to complete a destination assignment and that required to complete a reassignment are not correlated (Pearson correlation = 0.163, p-value =0.214). This result confirms that the task of goal assignment for initial planning and for replanning were distinct.

Regarding replanning, robot destination reassignment ratio and rerouting ratio are strongly correlated (Pearson correlation = 0.526. p-value < 0.001), suggesting that subjects performed both reassignments and rerouting with a similar frequency. Results also suggest that people who were faster in the visual search, conducted more rerouting and reassignments (Pearson correlation of reassignment frequency & time for the visual search = -0.388, p-value = 0.002; Pearson correlation of rerouting frequency & time for the visual search = -0.345. p-value = 0.005).

Using an ANOVA model with the number of robots as the main factor and the average time to complete a visual search as a covariate, we obtained statistical significance for both variables (p-values < 0.001). The R^2 of this model was 59.38%, which means that 59.38% of the performance variability is explained with these two variables. The Tukey post hoc test showed only difference in performance for 2 robots as compared to the other robot levels. This result confirmed the trend seen in Fig.5 and additionally pointed that that there was also difference in performance for 2 robots as compared to 4 and 6 robots. Including in the ANOVA model other variables such as time to replan, or time to assign robot destinations did not improve the model. Thus, the average time to complete a visual search was the main factor driving the performance score. In this analysis, it was important to use these additional metrics to confirm our initial results and ensure consistency across metrics.

Regarding attention allocation efficiency, results show a strong correlation between performance score and switching time (Pearson correlation = -0.533, p-value < 0.001). Thus, performance scores tended to be higher with low switching times. Interestingly, the switching time and the time to complete a visual search are not correlated, which indicates that these are two independent sources of performance variability (Pearson correlation = -0.098, p-value = 0.441). This result demonstrates that the two human behavior metric classes (attention allocation efficiency and human efficiency in the visual search) are measuring different aspects of the system that should be considered separately to understand team performance.

4.5 Human Behavior Cognitive Precursors

Figure 6 shows that as the robot team size increased, subjects overrode fewer robot autonomous decisions. A one-way ANOVA analysis of the overriding frequency showed that the robot team size significantly contributed to its variability (p-value < 0.001, R^2 = 50.86%). The Tukey post hoc test showed only difference in overriding frequency for 2 robots as compared to the other robot levels. As task load, which refers to the task demands imposed on an operator, increased, users decreasingly overrode robot decisions. This result suggests that workload was affecting subjects' pattern for overriding automation.

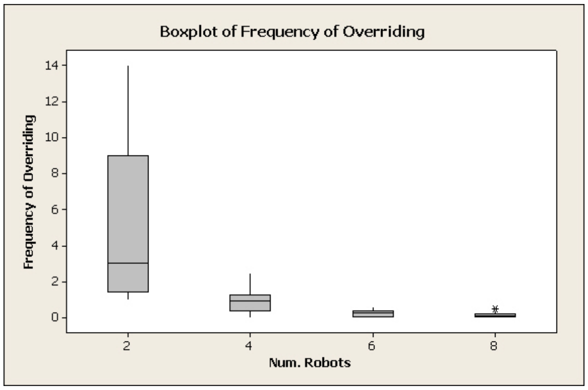

Figure 6. Overriding Robot Autonomy.

Additional investigation is needed to distinguish between subjects' cognitive saturation and subjects' over-reliance on robots. Subjective metrics for trust would allow further discussion. Since trust is a purely psychological state, subjective ratings are necessary to understand trust issues [17].

Figure 7 represents the perceived workload as reported by the subjects at the end of each scenario, 1 being nothing to do and 5 being completely overwhelmed. A one-way ANOVA analysis of workload showed that the robot team size significantly contributed to its variability (p-value = 0.005, R^2 = 18.86%). However, the R^2 of this model implies that it explained little of the workload variability. The Tukey test showed only difference in workload for 2 robots as compared to 6 and 8 robots. Subjective metrics are inexpensive and easy to administer, however they should be used to complement rather than to replace other forms of metrics.

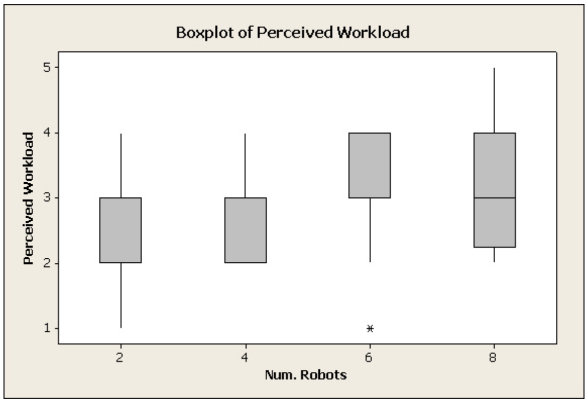

Figure 7. Perceived Workload.

4.6 Conclusions from the Case Study

This case study illustrates the need to measure multiple metrics across different metric classes to understand human-robot team performance, its underlying drivers, and effective design interventions.

In this experiment, analyzing human behavioral actions in the context of the tasks allowed us to identify that the visual search was the primary task driving the performance score. In addition, metrics of attention allocation efficiency pointed to an additional source of performance variability, switching time. Metrics of human behavior cognitive precursors allowed identifying that task load and over-reliance on robots' autonomy are interconnected. However, additional metrics for workload and trust that were not recorded during the experiment are necessary to distinguish between user cognitive overload and automation bias.

One potential drawback to the selection of metrics was that we did not explicitly measure behavioral actions in terms of problem recognition, decision making, and action implementation. Without this information, it is hard to say whether additional user support for problem recognition (e.g. which robot should I service next?) or decision making (i.e. which is the optimal route for this robot if I want to replan?) would be a better intervention to improve team performance. For interface design, measuring separately these three categories is essential because it allows exploring and understanding which parts of the mission require additional support and which design improvements can be more effective to maximize team performance. Measuring the complexity of the decisions that compose the mission and its workload as well as collecting more in-depth user feedback would also provide valuable information about future improvements.

However, problem recognition and decision making are highly interconnected and it can be difficult to measure them separately. As Klein and Klinger discuss, decision-making in complex environments under time pressure seems to be "induced by a starting point that involves recognitional matches that in turn evoke generation of the most likely action" [18]. Researchers should measure the observable outcomes of humans' decisions, and analyze and understand the decision process with other techniques such as verbal retrospective protocols.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a set of generalizable metric classes to consider for the evaluation of human-robot team performance. A case study of a single operator controlling multiple robots conducting a search and rescue mission illustrates the usefulness of measuring multiple metrics across these different classes.

Future work will populate these metric classes with the different types of metrics available and link them to actual research questions to help experimenters select the set of metrics that provide the most value for their experiments.

6. ACKNOWLEDGMENTS

This research was funded by MIT Lincoln Laboratory and U.S. Army Aberdeen Test Center.

7. REFERENCES

[1] Olsen, R., O. and Goodrich, M.A. 2003. Metrics for evaluating human-robot interations. In Proc. NIST Performance Metrics for Intelligent Systems Workshop.

[2] Steinfeld, A., et al. 2006. Common Metrics for Human-Robot Interaction. In Proceedings of the Conference on Human Robot Interaction (Salt Lake City, Utah, USA, March 2 - 3, 2006). HRI'06. ACM Press, New York, NY,

[3] Crandall, J.W. and Cummings, M.L. 2007. Identifying Predictive Metrics for Supervisory Control of Multiple Robots. IEEE Transactions on Robotics – Special Issue on Human-Robot Interaction, 23(5), 942-951.

[4] Sheridan T.B. 1992. Telerobotics, Automation, and Human Supervisory Control. The MIT Press. Cambridge, MA.

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

[6] Wickens, C.D. and Hollands, J.G. (1992). Engineering psychology and human performance. Third Edition. New York: HarperCollins.

[7] Rouse, W.B. and Morris N.M. (1986). On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulleting, 100, 349-363.

[8] Endsley, M.R. and Garland D.J. (Eds.) (2000) Situation Awareness Analysis and Measurement. Mahwah. NJ: Lawrence Erlbaum Associates.

[9] Drury, J.L., Scholtz, J., Yanco, H. (2003). Awareness in Human-Robot Interaction. In Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, October 2003.

[10] Norman, D.A. (2002). The design of everyday things. New York: Basic Books.

[11] Durso, F.T., Rawson, K.A., Girotto, S.(2007). Comprehension and Situation Awareness. Handbook of Applied Cognition. Second Edition. Edited by Francis Durso.

[12] Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (in press). Advances in measuring team cognition. In E. Salas, S. M. Fiore, & J. A. Cannon-Bowers (Eds.), Team Cognition: Process and Performance at the Inter- and Intra-Individual Level. Washington, DC: American Psychological Association. (Orlando, Florida, USA, 26-30 September 2005).

[13] Fiore. S.M., Schooler J., W. (2004). Process Mapping and Shared Cognition: Teamwork and the Development of Shared Problem Models. In E. Salas, S. M. Fiore, & J. A. Cannon-Bowers (Eds.), Team Cognition: Understanding the Factors that Drive Process and Performance. Washington, DC: American Psychological Association.

[14] Gorman, J., Cooke, N., Pederson, H., Connor, O., DeJoode, J. (2005). Awareness of Situation by Teams (CAST): Measuring Team Situation Awareness of a Communication Glitch. In Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting.

[15] Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Human Factors 39, 230-253.

[16] Cummings, M. L. (2004). "Automation Bias in Intelligent Time Critical Decision Support Systems." Paper presented at the AIAA Intelligent Systems Conference.

[17] Wickens, C. D. and Xu, X. (2002). Automation Trust, Reliability and Attention HMI 02 03, AHFD-02-14/MAAD-02-2, AHDF Technical Report.

[18] Klein, G., & Klinger, D. (2000). Naturalistic Decision Making. Human Systems IAC GATEWAY, 11 (3), 16-19.