A Metric Taxonomy for Supervisory Control of Unmanned Vehicles

M.L. Cummings, Patricia Pina, Jacob W. Crandall

MIT Humans and Automation Lab
Phone: 617-253-0993, Fax: 617-253-4196
Email: missyc@mit.edu
URL: http://halab.mit.edu

Abstract

Unmanned vehicle systems (UVSs), whether in the air, on the ground, on or under water, are inherently complex systems that rely on remote operator supervision to accomplish often dangerous and time-critical missions. Measuring the performance of a UVS is not a trivial task since the performance of the actual vehicle does not necessarily imply a performance level of the operator and vice-versa. With the increasing presence of UVSs in both military and commercial settings, it is critical that key performance metrics be identified to indicate not only operator and vehicle performance, but integrated human-system performance as well. To this end, this paper will describe a supervisory control metric taxonomy that classifies the different types of metrics across a supervisory control UVS, how they relate, and how this taxonomy can be used to identify a robust set of metrics.

I. Introduction

The DoD's strategic roadmap for the future envisions a battlefield where "various classes of unmanned systems operate together in a cooperative and collaborative manner to meet the joint warfighters' needs [1]." This vision means that there will continue to be rapid growth in all aspects of unmanned vehicle research, development, and operational fielding. One problem with such rapid growth is the ability to judge, either in a test or operational environment, whether the unmanned vehicle system (UVS) is adding value above and beyond a baseline, manned system.

In addition, in terms of system acquisition, it is often difficult to compare competing systems because of a lack of standardization in metrics, either for the system or for operator performance.

Indeed, for many UVS evaluation programs, large sets of metrics are gathered, which often include traditional human factors measures such as reaction time, error rates, and the overly-taxing NASA TLX (a subjective workload rating scale). In addition, often vague and context-dependent mission performance measures are gathered to prove or disprove system effectiveness, e.g., situation awareness and time to mission completion. While these measures are no doubt of importance, it is not clear how they could equitably be compared across systems with different human-computer interfaces. More importantly, the "measure everything and hunt for significant relationships in the post-hoc data analyses" approach provides little diagnostic information that indicates how and where interventions are needed to improve a UVS. In addition, this shotgun approach is also very expensive both in terms of time and money.

Because little guidance exists in the literature on how to select a set of meaningful supervisory control metrics for UVSs, we have developed a supervisory control metric taxonomy that classifies the metrics that could be gathered in a UVS, which is the focus of this paper. A second phase of this project, to develop a methodology to select the most parsimonious set of metrics from these metric classes needed for effective UVS evaluation, is currently underway.

In the context of this research, a metric class is defined as the set of metrics that quantify a certain aspect or component of a system. The idea of defining metric classes is based on the assumption that metrics are mission-specific, but that metric classes are generalizable across different missions. The idea of metric classes per se is not new, and previous work has been done to determine robot effectiveness metrics (e.g., [2]), human-robot interaction metrics [3], as well as the development of single human-multiple robot metric classes [4]. The research reported here

builds on this previous work by extending and adding to these models to build a comprehensive UVS supervisory control metric taxonomy that illustrates and ties together the many different aspects of the human operator, one or more unmanned vehicles, and system performance.

II. The Single Operator – Single Unmanned Vehicle Model

While there are many possible configurations of humans and unmanned vehicles (UVs), we first will describe our taxonomy for the single operator-single UV, and then build from this model. We propose that there are four conceptual groupings that form four metric classes for the single operator-single unmanned vehicle configuration which include 1) UV behavior, 2) human behavior, 3) human behavior precursors, and 4) collaboration (Figure 1). The respective

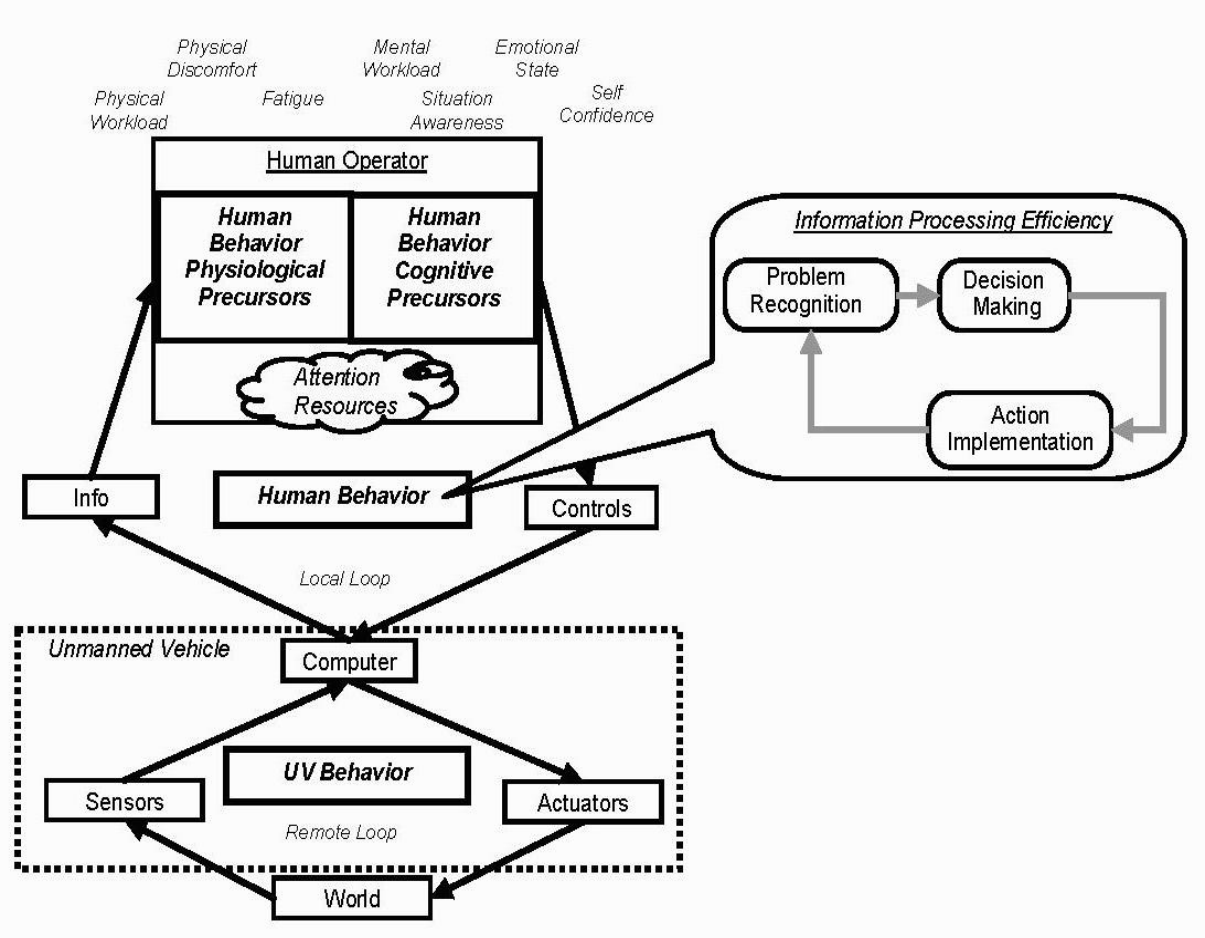

Figure 1: Supervisory Control of a Single Unmanned Vehicle.

behaviors of the UV and the human operator are represented by the two control loops shown in Figure 1. Characteristic of supervisory control systems, the operator receives feedback about the UV and mission performance, and adjusts the UV behavior through controls if required. The UV then interacts with the real world through actuators and collects feedback relating to mission performance through sensors. The evaluation of the performance of the operator-UV team requires an understanding of both human and UV control loops, so these two loops represent the two fundamental metric classes of a UVS, and are the focus of the next section.

UV and Human Behavior Metric Classes

In terms of actual metrics that populate these two fundamental classes, we propose that for the UV behavior metric class, subclass metrics include usability, adequacy, autonomy, and self-awareness. Usability refers to several related attributes typically associated with learnability, efficiency, memorability, errors, and user satisfaction [5]. Adequacy is the ability to satisfactorily and sufficiently support the mission, and this metric subclass contains measures of automation accuracy and reliability. Autonomy is the ability of the UV to function independently, and self-awareness corresponds to the UV's awareness of itself [6].

The human behavior metric class in Fig. 1 refers to the decisions made and actions taken by the human to complete the mission. We propose that the two primary metric subclasses for human behavior efficiency are attention allocation efficiency and information processing efficiency. Divided attention is inherently a human supervisory control attribute, thus the attentional resource allocation subclass assesses the operator strategies and priorities in managing multiple tasks and sharing attention among them. For the single operator-single UV case, even though only one UV is controlled, the operator still performs multiple tasks such as monitoring UV health and status as well as the environment, identifying emergent events, and

commanding the vehicle. How humans sequence and prioritize these multiple tasks provides valuable insights into the system design.

Information processing metrics measure how well the individual tasks and activities that compose the overall mission are conducted. Attention allocation efficiency metrics examine an operator's ability to manage across tasks but information processing metrics provide insight within a task. These subclasses are related in that attention allocation will drive information processing for a specific task; however, information processing efficiency provides additional information about the system. Instead of focusing on task management in attention allocation, this subclass focuses on an operator's problem recognition, decision making, and action implementation [7]. We recognize that differentiating between information processing states is often difficult, so in some cases, the use of generic task efficiency metrics such as the time required to prosecute a target can capture overall information processing efficiency.

Table 1 summarizes the metric subclasses for the human behavior efficiency metric class and provides examples for illustrative purposes. The next section will discuss in more depth the two remaining metric classes.

Table 1: Overview of Metrics Subclasses for the Human Behavior Metric Class.

Metric Subclasses			**Measure Examples**
Attention Allocation Efficiency			% of time operator is focused on the highest priority task
Information Processing Efficiency	Task Efficiency	Recognition Efficiency	Error detection rate Error detection time
		Decision Making Efficiency	Correct decision rate Quality of decisions
		Action Implementation Efficiency	Control input activity Frequency of functionality usage

Human Behavior Precursors and Collaborative Metric Classes

While the two fundamental classes of human and UV behavior are necessary to understand system behavior, they are also insufficient because they do not address the underlying cognitive processes leading to specific operator behavior. These factors are represented by the metric class of human behavior precursors, which includes both cognitive and physiological precursors. Human behavior precursors are cognitive constructs or processes that exist or occur before a certain behavioral action is observed. Human behavior is driven by high-level cognitive constructs and processes such as situation awareness (SA), mental workload, self-confidence, and emotional state (Fig. 1). In addition to cognitive precursors, physiological precursors such as fatigue or physical discomfort can also affect human performance.

Finally, the operator and the UV constitute a team that works together to conduct a mission. Therefore, evaluating how well the UV and the human collaborate motivates the fourth metric class of collaboration. The metric subclasses which examine human-UV collaboration for the single operator-single vehicle model are UV-human awareness, human mental models, and human trust.

UV-human awareness is the degree to which automation is aware of the human role, including humans' commands and constraints that may require a modified course of action or command noncompliance. Depending on the application, onboard automation may need to have knowledge of humans' expectations, constraints, and intents, thus it is critical to quantify a UV's model of the human. While not typically found on operational UVs today, with increasing use of artificial intelligence onboard UVs, the vehicles could modify their behavior based on human actions and predicted states. It will be critical that such models are accurate, so how well these models match actual human intentions and actions should be evaluated.

In terms of the mental model subclass, a human mental model is an organized set of knowledge with depth and stability over time that reflects the individual's perception of reality. Mental models allow people to describe and understand phenomena, draw inferences, make predictions, and decide which actions to take, thus automation design should be consistent with people's natural mental models [8]. Evaluation of mental models can inform display design requirements and also training material development.

Lastly, human trust in automation is another important collaborative metric. In the context of complex human-automation systems, Madsen and Gregor have defined trust as "the extent to which a user is confident in, and willing to act on the basis of the recommendations, actions, and the decisions of a computer-based tool or decision aid [9]. Operators' lack of trust in automation and the resulting automation disuse thwarts the potential that a new technology offers. However operators' inappropriate excessive trust and the resulting automation misuse could lead to complacency and the failure to intervene when the technology either fails or degrades [10]. Thus, objectively measuring trust, arguably a difficult task, is important when system reliability and the domain culture could create trust barriers.

The Fifth Metric Class: Mission Effectiveness

While not represented explicitly in Figure 1, there is a fifth metric class that measures aggregate system performance, that of mission effectiveness. Key performance parameters and effects-based outcomes represent meaningful system performance measures, but they are often system and mission dependent. However, while not always generalizable, having an overall mission effectiveness metric is critical in determining the severity of the impact of the other metric classes. For example, given a particular system, if mental workload is reported high, attention allocation seems inefficient, and SA measures low, but the overall mission

effectiveness metric is high, either the system is very robust or more likely, there is a problem with one or more of the subclass measures or some aspect of the system was not adequately measured. Thus mission effectiveness metrics are critical for determining whether a system actually meets its stated objectives, but it can also provide insight into the validity of other system metrics.

III. The Single Operator - Multiple Unmanned Vehicles Model

In a supervisory capacity, operators intermittently interact with UVs, so it is possible that an operator could control multiple vehicles, particularly as onboard automation increases. The DoD recognizes this possibility and is moving towards this future operational architecture [1]. Thus, we have adapted our single operator-single UV model above to demonstrate how these same metric classes would be characterized in a multiple UV scenario. However, single operator

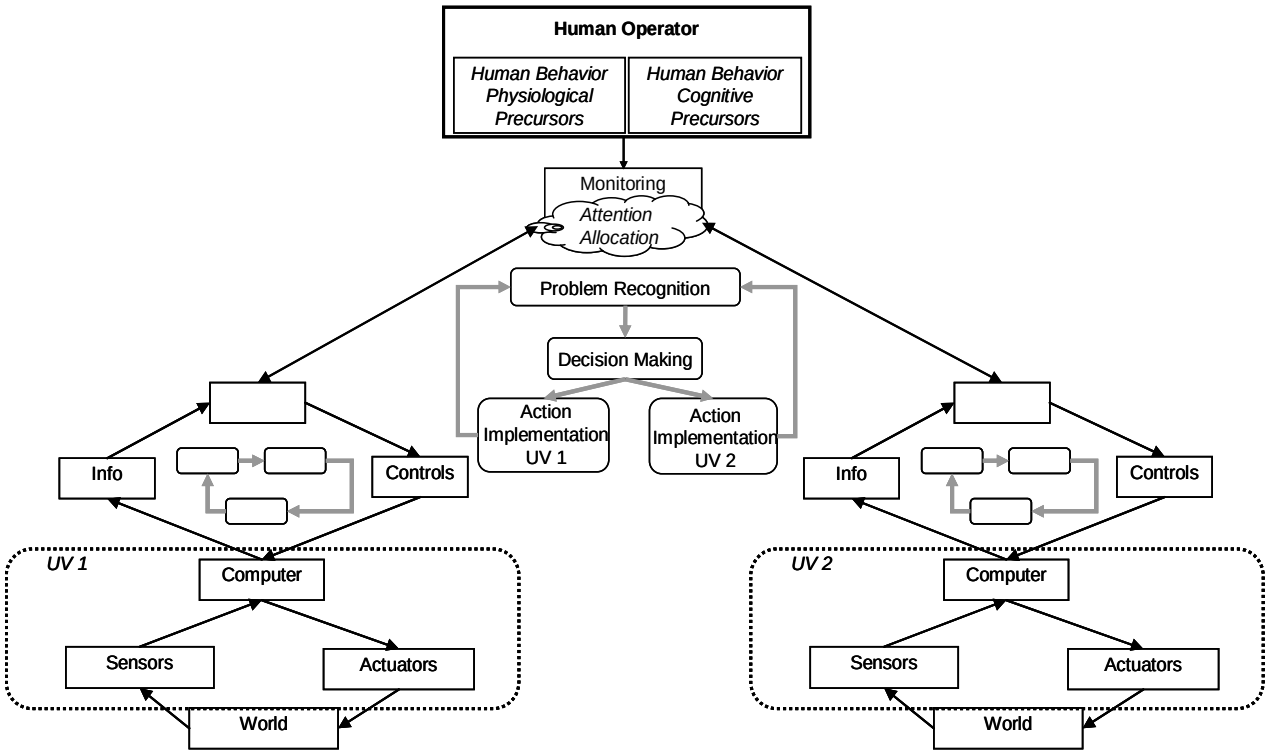

Figure 2: Supervisory Control of Independent UVs.

control of multiple UVs can be manifested in two ways: multiple UVs performing independent tasks (Figure 2), and multiple UVs performing collaborative tasks (Figure 3).

In the simpler case of an operator controlling two independent UVs, the operator monitors the environment and the UVs' status, decides on which one to focus attention, interacts with that UV and when finished, returns to group monitoring or decides to service another UV. In the independent multiple vehicle control case, no additional metric classes are needed, but there are other considerations for various subclasses. In terms of the human behavior metric class, additional attention allocation metrics should be considered such as measuring task/vehicle switching frequency, UV prioritization strategies, and length and quality of vehicle interaction.

In contrast with the independent multiple UV scenario, one operator can supervise multiple UVs that coordinate with one another, as well as with the operator (Figure 3). Because

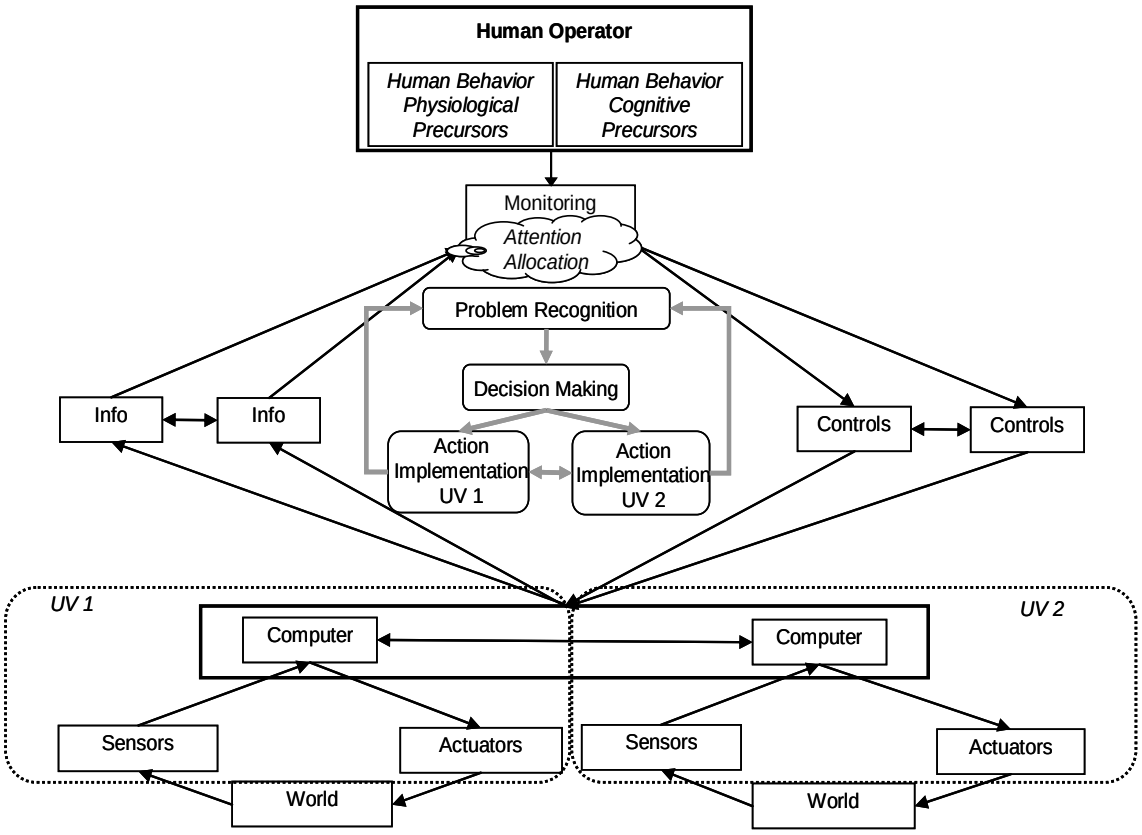

Figure 3: Supervisory Control of Collaborative UVs.

the control loops for the UVs are no longer independent, servicing the vehicles is inherently dependent. For example, making a decision for UV1 in Fig. 3 can involve acquiring and analyzing information related to UV2, and implementing an action for UV2 can require synchronizing it with UV1. Moreover, good human factors display design principles dictate that to the largest extent possible, information should be integrated [11], so the dependencies exist not just as the vehicle level, but also at the ground control station level.

Just as in the independent case, the five metric classes are still sufficient, but several subclasses are impacted by these collaborative dependencies. In addition to those subclasses discussed above, the information processing efficiency subclass in the human behavior metric class is distinctly affected in the multiple vehicle control model. While in the case of independent UVs, problem recognition, decision making, and action implementation can be evaluated separately, for the collaborative UV case, these will likely have to be analyzed in the aggregate, due to the inability to decouple the effects of the different UVs on these states.

To account for the inter-vehicle collaboration, a new subclass is needed in the collaboration metric class, which is UV-UV collaboration. In the single operator-single vehicle and single operator-multiple independent vehicles models, all collaboration took place just between the operator and a vehicle. With collaborative UVs, both the quality and efficiency of the collaboration between vehicles can be measured (e.g., information sharing such as path obstacles and the presence of unexpected threats), but also how the human collaborates with the UVs either individually or in a group. This multiple collaborative UV model reflects the swarming concept of operations, so it is critical to understand how and what hierarchical level an operator should interact with the vehicles to promote both efficient system performance and sufficient situation awareness.

IV. The Multiple Operator-Multiple Unmanned Vehicle Model

Given the inherent team nature of command and control operations, the single operator-multiple UV architecture is somewhat artificial and in most cases, will probably be a multiple operator-multiple UV scenario. Thus, we extend our model to address this configuration (Figure 4). For the collaborative metrics class, the previously discussed subclasses (human-UV and UV-UV collaboration) also apply for the multiple operator, multiple UV system. However, because of the introduction of additional operators, we add the human-human collaboration subclass.

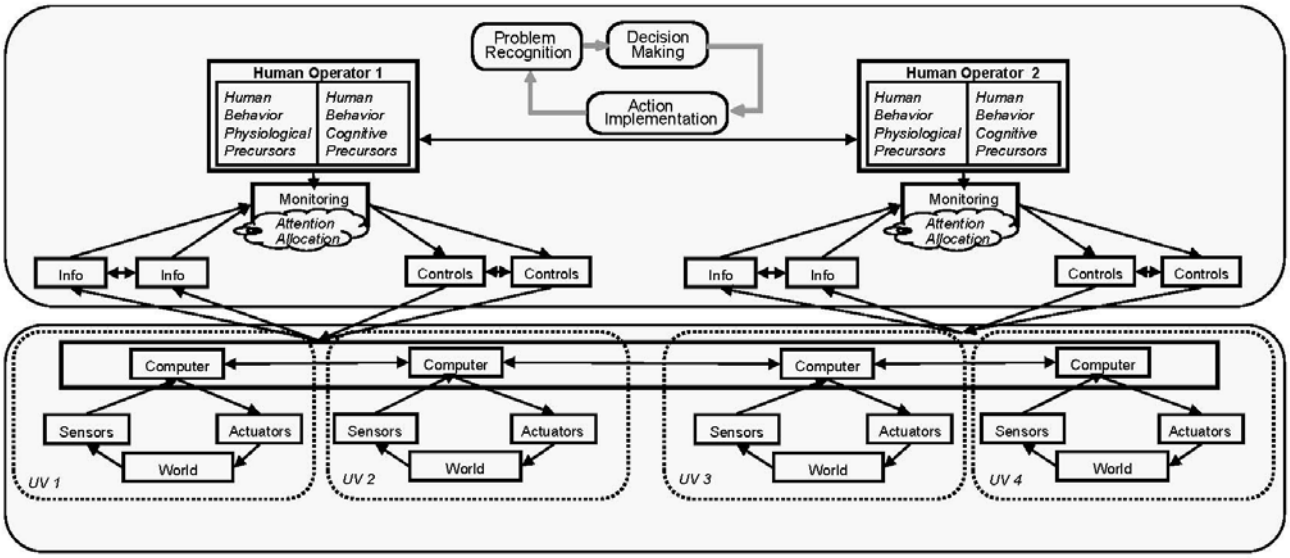

Figure 4: Human/Human and UV/UV Collaborative Metrics

In command and control settings, a human 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]. Since team members must consistently exchange information, reconcile inconsistencies, and coordinate their actions, one way to measure holistic team performance is through team coordination, which includes written, oral, and gestural interactions. Team coordination is generally assessed through communication analysis, which can include quantitative physical measures such as how long team members spend communicating, as well as more qualitative measures that focus on the communication

content. In addition, the focus of the measures can be static (at a single point in time), or more dynamic such as measuring patterns of communication. Measures of behavioral patterns such as communications, social networks, etc. are traditional metrics in team research.

In addition to measuring team coordination for the human-human metric subclass, measuring team cognition, which refers to the thoughts and knowledge of the team, can be valuable in diagnosing team performance and identifying effective training and design interventions [13]. Just as for the individual operator, the team has an aggregate mental model as well as shared SA. Since efficient team performance has been shown to be related to the degree that team members agree on, or are aware of task, role, and problem characteristics [13], team mental models and team SA should be considered when evaluating the multiple operator, multiple UV architecture.

V. Metric Class Summary

Based on the operator-UV models presented in this report, five generalizable metric classes were identified through a principled approach for human-automation team evaluation (Table 2). Examples of subclasses are included, and for a more exhaustive discussion of the specific metrics that populate these classes and subclasses, the reader is referred to Pina et al. [14]. We have shown that these metric classes apply to any system of humans and unmanned vehicle, regardless of the vehicle type, and the combination and degree of collaboration between humans and/or UVs. It is important to note that these classes are not independent, thus in many cases metrics will likely be correlated, which is discussed further in the next section.

Which specific metrics should be used to evaluate a system will depend on many factors, but as a rule-of-thumb, we propose that at a minimum, one metric from each class should be used to provide a multi-dimensional assessment of a UVS. Some metrics may be more valuable than

others, and as is discussed in the next section, determining the optimal set of metrics *a priori* is an area of ongoing research. However, failing to follow either this or any other principled system evaluation metric approach means that some aspect of the system will not be measured, and thus some latent condition could later be manifested because of the failure to comprehensively evaluate the system.

To determine what the impact on our research has been by not following such a principled approach, we evaluated several recent large-scale supervisory control experiments conducted in the MIT Humans and Automation Laboratory. The results show that prior to adapting this metric classification approach, we were fairly consistent in measuring mission effectiveness and human behavior through such metrics as reaction times and decision accuracies. However, despite our supervisory control focus, we were remiss in gathering attention allocation metrics and collaboration metrics, and we often gathered too many correlated metrics that were redundant and wasteful [14]. This meta-analysis of our experimental

Table 2: Unmanned Vehicle Human Supervisory Control Metric Classes and Subclasses

1) Mission Effectiveness (e.g., key mission performance parameters)
2) UV Behavior Efficiency (e.g., usability, adequacy, autonomy, reliability)
3) Human Behavior Efficiency
   - Attention allocation efficiency (e.g., task switching times, prioritization)
   - Information processing efficiency (e.g., decision making accuracy, reaction times)
4) Human Behavior Precursors
   - Cognitive Precursors (e.g., SA, mental workload ,self-confidence, emotional state)
   - Physiological Precursors (e.g., physical comfort, fatigue)
5) Collaborative Metrics
   - Human/UV Collaboration (e.g., trust, mental models)
   - Human/Human Collaboration (e.g., coordination metrics, team mental model, team SA)
   - UV/UV Collaboration (e.g., vehicle reaction times to situational events that require autonomous collaboration)

shortcomings reflect those in the general research population in that we tended to gravitate to popular metrics that are relatively easy to gather, without a clear understanding of exactly what aspect of the systems we were measuring and how the various metrics informed an overall research question. We are now using the metric classification framework proposed here to inform our design of experiments across a number of different projects.

VI. Towards a model for metric selection

Given that we have comprehensively defined the UV supervisory control metric classes, as well as the different subclasses for different human operator-UV configurations, the next question is, "Which specific metric(s) should I use to evaluate my system?" This is the focus of ongoing research, and we are examining criteria such as experimental constraints (e.g., time to run an experiment, access to real operators), construct validity, (e.g., is my EEG monitor really measuring workload or measuring stress?), metric value-added (e.g., does each metric contribute to my research question in distinct manner or is there significant overlap?), statistical validity (is this metric highly correlated with another, possibly inflating experimental error and wasting resources?), and the measuring technique (e.g., does interrupting users to ask situation awareness questions interfere with either performance or the gathering of other, more critical system data such as interaction times?)

From this initial taxonomy work, we hope to develop a cost-benefit methodology that can provide clear and tangible guidelines to researchers and practitioners to aid them in metric selection. While no such approach will ever be able to provide a metric checklist for every system and every research question of interest, we hope to provide theoretical grounding for why some measures could be better than others in some contexts, and how some areas of focus, such

as resource allocation problems in a UV system, can lead to a set of generalizable metrics that can be used across different systems.

Acknowledgements

We would like to thank Bianca Farrell and Dr. Birsen Donmez for assistance with this project. This research was funded in part by the US Army Aberdeen Test Center and the Northrop Grumman Integrated Systems Innovation IRAD program.

References

1. Office of the Secretary of Defense, *Unmanned Systems Roadmap 2007 (Draft v1.1)*. 2007, DoD: Washington DC.
2. Olsen, D.R. and M.A. Goodrich. *Metrics for Evaluating Human-Robot Interactions*. in *Performance Metrics for Intelligent Systems*. 2003. Gaithersburg, MD.
3. Steinfeld, A., et al. *Common Metrics for Human-Robot Interaction*. in *HRI*. 2006. Salt Lake City, Utah.
4. Crandall, J. and M.L. Cummings. *Developing Performance Metrics for the Supervisory Control of Multiple Robots*. in *Human Robotics Interaction 2007*. 2007. Washington DC.
5. Nielsen, J., *Usability engineering*. 1993, Cambridge, MA: Academic Press.
6. Drury, J.L., J. Scholtz, and H.A. Yanco. *Applying CSCW and HCI techniques to human-robot interaction*. in *CHI 2004 Workshop on Shaping Human-Robot Interaction*. 2004. Vienna.
7. Parasuraman, R., 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: Systems and Humans, 2000. **30**(3): p. 286-297.
8. Norman, D.A., *The Design of Everyday Things*. 1988, New York: Doubleday. 257.
9. Madsen, M. and S. Gregor. *Measuring human-computer trust*. in *Eleventh Australasian Conference on Information Systems*. 2000. Brisbane.
10. Parasuraman, R. and V. Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse.* Human Factors, 1997. **39**(2): p. 230-253.
11. Wickens, C.D. and J.G. Hollands, *Engineering Psychology and Human Performance*. 3rd ed. 2000, Upper Saddle River, N.J.: Prentice Hall.
12. Cooke, N.J., et al., *Advances in measuring team cognition*, in *Team Cognition: Process and Performance at the Inter- and Intra-Individual Level*, E. Salas, S.M. Fiore, and J.A. Cannon-Bowers, Editors. 2004, American Psychological Association: Washington, DC.
13. Fiore. S.M. and Schooler J.W., *Process Mapping and Shared Cognition: Teamwork and the Development of Shared Problem Models*, in *Team Cognition: Understanding the Factors that Drive Process and Performance*, E. Salas, S.M. Fiore, and J.A. Cannon-Bowers, Editors. 2004, American Psychological Association: Washington, DC.
14. Pina, P.E., B. Donmez, and M.L. Cummings, *Selecting Metrics to Evaluate Human Supervisory Control Applications*. 2008, MIT Humans and Automation Laboratory: Cambridge, MA.