

Selecting Metrics to Evaluate Human Supervisory Control Applications

P.E. Pina
B. Donmez
M.L. Cummings

Massachusetts Institute of Technology*

HAL2008-04
May 2008



<http://halab.mit.edu>

e-mail: halab@mit.edu

*MIT Department of Aeronautics and Astronautics, Cambridge, MA 02139

Table of Contents

Abstract	3
1. Introduction	5
2. Terminology	6
3. Metric Taxonomy for Human Supervisory Control.....	7
3.1. Supervisory control of a single autonomous platform	7
3.1.1. Autonomous platform behavior	8
3.1.2. Human behavior	9
3.1.3. Human behavior cognitive and physiological precursors	10
3.1.4. Human - autonomous platform collaboration	11
3.1.5. The fifth metric class: mission effectiveness	12
3.1.6. Metric classes for the supervisory control of one platform	12
3.2. Human and platform collaboration in supervisory control	13
3.2.1. Supervisory control of multiple independent or collaborative platforms ...	14
3.2.2. Human collaboration in supervisory control of multiple platforms	17
4. Generalizable Metric Classes for Human Supervisory Control	20
5. Populating Metric Classes with Existing Metrics	21
5.1. Mission effectiveness	21
5.2. Autonomous platform behavior efficiency	21
5.2.1. Usability	21
5.2.2. Adequacy	22
5.2.3. Autonomy.....	23
5.2.4. Self-awareness	24
5.3. Human behavior efficiency	24
5.3.1. Attention allocation efficiency.....	24
5.3.2. Information processing efficiency.....	28
5.4. Human behavior cognitive precursors.....	31
5.4.1. Mental workload.....	31
5.4.2. Situation awareness	36
5.4.3. Self-confidence	44
5.4.4. Emotional state	44
5.5. Human behavior physiological precursors	45
5.5.1. Physical workload	45
5.5.2. Fatigue.....	45
5.5.3. Physical comfort	45
5.6. Collaborative metrics	46
5.6.1. Human - autonomous platform collaborative metrics	46
5.6.2. Human - human collaborative metrics.....	51
5.6.3. Autonomous platform - autonomous platform collaborative metrics	54
6. Case Studies: Past Research	55
6.1. An ecological perceptual aid for precision vertical landings	55
6.1.1. Background.....	55
6.1.2. Metrics	55
6.1.3. Results.....	57

6.1.4.	Discussion and conclusions.....	57
6.2.	Decision support for lunar and planetary exploration.....	58
6.2.1.	Background.....	58
6.2.2.	Metrics.....	59
6.2.3.	Results.....	61
6.2.4.	Discussion and conclusions.....	63
6.3.	Assessing the impact of auditory peripheral displays for UAV displays.....	64
6.3.1.	Background.....	64
6.3.2.	Metrics.....	65
6.3.3.	Results.....	66
6.3.4.	Discussion and conclusions.....	66
7.	Case Studies: On-going Research.....	68
7.1.	Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU).....	68
7.1.1.	Background.....	68
7.1.2.	Metrics.....	68
7.1.3.	Discussion and conclusions.....	69
8.	Conclusions from the Case Studies.....	70
9.	Preliminary Evaluation Criteria for Supervisory Control Metrics.....	71
Appendix A: Unidimensional Workload Self-Rating Scales.....		80
Appendix A: Unidimensional Workload Self-Rating Scales.....		80
Appendix B: Multidimensional Workload Self-Rating Scales.....		82
Appendix C: SA Subjective Metrics and Techniques.....		84
Appendix D: Multiple Rating Scales to Elicit Dimensions of Trust.....		88
Appendix E: Knowledge Elicitation Techniques.....		92
Appendix F: Mental-Model Analysis and Representation Techniques.....		94

Abstract

The goal of this research is to develop a methodology to select supervisory control metrics. This methodology is based on cost-benefit analyses and generic metric classes. 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. Generic metric classes are developed because metrics are mission-specific, but metric classes are generalizable across different missions. Cost-benefit analyses are utilized because each metric set has advantages, limitations, and costs, thus the added value of different sets for a given context can be calculated to select the set that maximizes value and minimizes costs. This report summarizes the findings of the first part of this research effort that has focused on developing a supervisory control metric taxonomy that defines generic metric classes and categorizes existing metrics. Future research will focus on applying cost benefit analysis methodologies to metric selection.

Five main metric classes have been identified that apply to supervisory control teams composed of humans and autonomous platforms: mission effectiveness, autonomous platform behavior efficiency, human behavior efficiency, human behavior precursors, and collaborative metrics. Mission effectiveness measures how well the mission goals are achieved. Autonomous platform and human behavior efficiency measure the actions and decisions made by the humans and the automation that compose the team. Human behavior precursors measure human initial state, including certain attitudes and cognitive constructs that can be the cause of and drive a given behavior. Collaborative metrics address three different aspects of collaboration: collaboration between the human and the autonomous platform he is controlling, collaboration among humans that compose the team, and autonomous collaboration among platforms. These five metric classes have been populated with metrics and measuring techniques from the existing literature.

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 the human-automation team. To determine what the impact on our research has been by not following such a principled approach, we evaluated 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. This meta-analysis of our experimental 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.

Given that we have comprehensively defined the supervisory control metric classes and subclasses, the next question is “Which specific metric(s) should I use to evaluate my system?” Based on the literature review conducted and the case studies examined, a preliminary list of evaluation criteria for supervisory-control metrics has been identified. This criteria list includes experimental constraints, construct validity, comprehensive understanding gained, statistical validity and efficiency, and appropriateness of the measuring technique. The refinement of this list and the development of a cost-benefit methodology that can provide clear and tangible guidelines to select metric is the focus of ongoing research. 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.

1. Introduction

Teams of humans and automation operating under a supervisory control paradigm are currently common across domains and applications: surveillance and target identification for military operations, health care applications such as robotics for surgery, mobility assistance and therapy, rock sampling for geology research, or other logistic applications such as personnel or material delivery. In all these examples, the operator plans and monitors the performance of an automated agent with a certain degree of autonomy, retaking manual control when needed.

The most popular metric to evaluate the performance of these teams is mission effectiveness, but, frequently, this metric is not sufficient to understand performance issues and to identify design improvements. Mission effectiveness metrics can be insufficient to extract the information necessary to design more effective systems,[1]. However, little guidance exists in the literature on how to select additional meaningful metrics. In many cases, researchers rely on their own experience, choosing those metrics they have used previously. Alternatively, other experimenters 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.

The goal of this research is to provide guidelines for metric selection to evaluate teams operating under a human supervisory control paradigm. The approach is to develop a framework based on cost benefit analyses and generic metric classes, which will enable researchers to select a robust set of metrics that provide the most value. This report summarizes the findings of the first part of this research effort that has focused on developing a supervisory control metric taxonomy that classifies the metrics that could be gathered in a human-automation team. Future research will focus on developing a cost-benefit methodology to select the most parsimonious set of metrics from these metric classes needed for effective team evaluation.

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. 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 concept 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 [2]. This set of metrics measures the individual performance of a robot; however, a particular robot performance does not necessarily explain the level of human performance. Since human cognitive limitations often constitute a primary bottleneck for human-automation teams in supervisory control applications, 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 automation performance metrics, Steinfeld et al. suggested identifying common metrics in terms of three aspects: human, robot, and the system [3].

Regarding human performance, they discussed three main metric categories: situation awareness, workload, and accuracy of mental models of device operations. This work constitutes an important effort towards developing a metric toolkit; however, this framework suffers from a lack of metrics to evaluate collaboration effectiveness among humans and among robots. In addition, a more comprehensive discussion on human performance is still required. For example, the authors do not include trust as a common metric required to evaluate operator performance. However, operators' trust in robot behavior is often a key factor in team performance because it significantly affects whether and how the robot is used [4].

This research builds upon previous efforts conducted by Crandall and Cummings [5]. It refines, expands, and generalizes the set of metric classes already identified for teams consisting of a single human and multiple robots. This report discusses conceptual models for human supervisory control applications, identifies metric classes based on these models, and populates them with existing metrics. In addition, actual experiments conducted at HAL are discussed in the context of these metric classes.

2. Terminology

“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 [6].”

This research uses the generic term “autonomous platform” to refer to the computer, the effectors, and the sensors that close an autonomous control loop. It should be noted that the term automation in the context of this research has the same meaning as autonomous platform.

Many human supervisory control applications use the more popular term autonomous vehicles¹ (AVs). However, the term vehicle has certain implications in terms of mobility that are not generalizable across all human supervisory control applications. Thus, this report will not employ the words “autonomous vehicles.”

This research also employs the term team to refer to a group of humans and automation performing a supervisory control task, in which the operators plan and monitor the performance of an autonomous platform with a certain degree of autonomy, and retake manual control when needed.

¹An autonomous vehicle is an unmanned vehicle with some level of autonomy built in, from teleoperations to fully intelligent systems [7]. UVs can be unmanned aerial vehicles (UAVs), unmanned surface vehicles (USVs), unmanned undersea vehicles (UUVs), or unmanned ground vehicles (UGVs). This definition of an AV is broad enough to include weapons systems such as torpedoes, mobile mines, and ballistic and cruise missiles.

3. Metric Taxonomy for Human Supervisory Control

3.1. Supervisory control of a single autonomous platform

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

The respective behaviors of the autonomous platform and the human are represented by the two control loops shown in Figure 1. Characteristic of supervisory control systems, the operator receives feedback on the autonomous platform and mission performance, and adjusts automation behavior through controls if required. The autonomous platform 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, so these two loops represent the two fundamental metric classes of human-automation teams.

However, evaluating the observable behavior of the human and the autonomous platform is insufficient. Optimizing team performance requires understanding the motivation and the cognitive processes leading to a specific human behavior. These factors are represented in our model by the metric class of human behavior precursors, which includes both cognitive and physiological precursors. It should be noted that human behavior is often related to the environmental conditions and the operator's state when a given event occurs. In general, the response to an event can be described in terms of three set of variables [8]: a pre-event phase that defines how the operator adapts to the environment; an event-response phase that describes the operator's behavior in accommodating the event; an outcome phase that describes the outcome of the response process. Thus, in addition to human behavior, experimenters need to measure human behavior precursors to represent the operator's state, and the autonomous platform behavior to represent the initial environmental conditions.

Finally, it should be considered that the human and the autonomous platform constitute a team that works together to conduct a mission. Therefore, evaluating how well the human and the automation collaborate motivates the fourth metric class of collaboration.

In addition to these four elements, two other concepts are represented in Figure 1: uncertainty, and the mission or the task. Uncertainty refers to the uncertainty associated with sensors (e.g., accuracy), actuators (e.g., lag), displays (e.g., transforming 3D information into 2D information), and the real world. This uncertainty propagates through the system reaching the operator who adapts his behavior to the uncertainty level by applying different cognitive strategies. The nature of the tasks/mission imposed on the operator is also represented in this figure because it affects performance. For example, high-structured tasks, which can be planned in advance, are procedurally-driven, whereas low structured tasks

are generally emergent and require solving a new problem. Team performance can be understood only if considered in the context of the mission, the task, and the existing uncertainty.

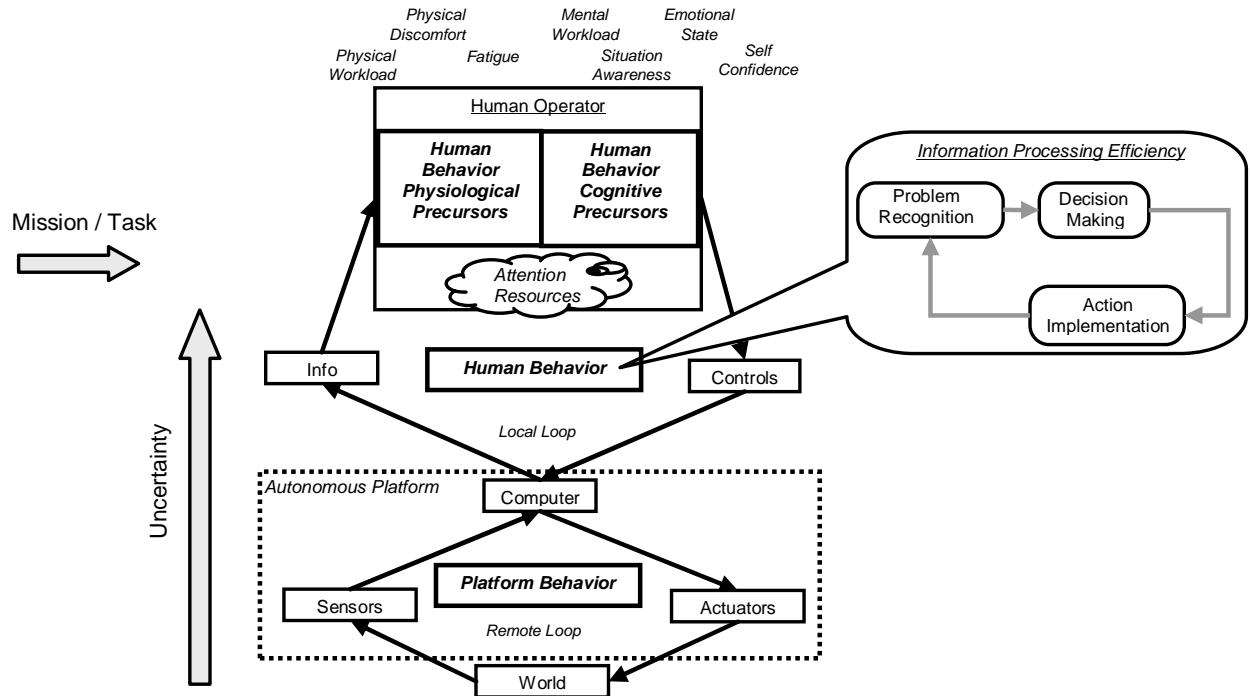


Figure 1: Conceptual Model of Human Supervisory Control Applications.

3.1.1. Autonomous platform behavior

In terms of actual metrics that populate this class, we propose that for the autonomous platform behavior metric class, subclass metrics include usability, adequacy, autonomy, and self-awareness. Usability refers to several related attributes and has been traditionally associated with learnability, efficiency, memorability, errors, and satisfaction [9]. Adequacy refers to the ability to satisfactorily and sufficiently support the operator in completing the mission, and this metric subclass contains measures of accuracy and reliability. Autonomy is the ability of the platform to function independently; and self-awareness corresponds to the autonomous platform’s awareness of itself [10].

In environments where social barriers for automation adoption are expected, it is important to evaluate automation accuracy, reliability, understandability, and ease of use as each can have a direct effect on operator’s trust on automation. Evaluating these automation characteristics that affect operator’s trust can help minimizing causes or sources of distrust, and therefore, increase the chances for a successful automation adoption.

3.1.2. Human behavior

The human behavior metric class, in the context of Figure 1, refers to the decisions made and actions taken by the human to complete the mission. Divided attention and constant information processing are inherently human supervisory attributes, thus we propose that the two primary metric subclasses for human behavior efficiency are attention allocation efficiency and information processing efficiency.

Attention allocation metric subclass assesses the operator strategies and priorities in managing multiple tasks and sharing his attention among them. Operators have limited attention resources that need to be shared between multiple tasks [11]. Although one single autonomous platform is controlled, the operator still performs multiple tasks such as monitoring the dynamics of the environment, identifying emergent events, monitoring the platform health, or executing manual control of the platform. How humans sequence and prioritize these multiple tasks provides valuable insights into the system.

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 can provide additional information about the system. Instead of focusing on task management in attention allocation, this subclass focuses on an operators' problem recognition, decision making, and action implementation. Evaluating problem recognition, decision making, and action implementation separately enables exploring and understanding which parts of the mission require additional support, and which design improvements can be more effective to maximize team performance. 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 [12]. This research merges the stages of information acquisition and analysis into the problem recognition metric subclass. Acquisition and analysis of information are often hard to differentiate, and the human ability to recognize problems is a more valuable metric for the purposes of this research.

We recognize that differentiation between information processing states is often difficult. For example, 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 [13]." In these cases, the use of generic task efficiency metrics, such as performance metrics (e.g., the number of obstacles avoided by an autonomous vehicle when navigation is a primary task of the mission) and time metrics (e.g., the time required to detect and correct a deviation from the nominal route) can capture overall information processing efficiency.

Table 1 summarizes the metric subclasses for the human behavior efficiency metric class and provides measure examples for illustrative purposes.

Table 1: Overview of Metrics Subclasses for Human Behavior Efficiency.

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

3.1.3. Human behavior cognitive and physiological precursors

While the two fundamental classes of human and autonomous platform 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. In the context of this research, human behavior cognitive precursors refer to 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 situation awareness² (SA). For our discussion, SA reflects short-term knowledge about a dynamic environment. 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 [14].

SA is not the only human behavior cognitive precursor, mental workload, and operator emotional state are other examples of cognitive constructs and processes that can also lead to certain human behaviors. Mental workload results from the demands a task imposes on the operator's limited resources; it is fundamentally determined by the relationship between resource supply and task demand [11]. Differences in operators' skill levels and strategies can lead to differences in workload for the same task demands or task load. Thus, workload is not only task-specific, but also person-specific. The measurement of mental workload enables, for example, identification of bottlenecks in the system or the mission in which performance can, but does not break down in a particular experiment. Measurement of mental workload can also enable the comparison of systems that lead to similar performance.

In addition to human behavior cognitive precursors, physiological precursors such as fatigue, or physical discomfort can also motivate certain human attitudes. Measuring physiological states can help the researchers understand the causes for observed human behavior.

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" [15]. For example, 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 [16].

3.1.4. *Human - autonomous platform collaboration*

Finally, the collaboration metric class examines how well the human and the autonomous platform collaborate. As mentioned previously, the human and automation need to work together to accomplish a common goal; the nature of their relation and how well they collaborate significantly affects the system performance. The metric subclasses that examine human-autonomous platform collaboration are autonomous platform - human awareness, human mental models, and human trust.

Autonomous platform - human awareness is the degree to which automation is aware of humans, including humans' commands and any human-originated constraints that may require a modified course of action or command noncompliance. Depending on the application, automation may need to have knowledge of humans' expectations, constraints, and intents, thus it is critical to quantify an autonomous platform's model of the human. While not typically found on operational autonomous platforms today, with increasing use of artificial intelligence onboard autonomous platforms, the automation 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 human intentions and actions should be evaluated.

In terms of the mental model subclass, a human mental model³ 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 [18]. Evaluation of mental models can inform displays design requirements and also training material development.

Lastly, human trust in automation is the third metric subclass. Trust concerns an expectancy or an attitude regarding the likelihood of favorable responses [19]. Measuring trust is important because, as Parasuraman and Riley showed, trust drives misuse and disuse⁴ of automation [20]. People tend to rely on and use the automation they trust and tend to reject the automation they do not. Operators' lack of trust in automation and the resulting automation disuse thwarts the potential that a new technology offers, and operators' inappropriately excessive trust and the resulting automation misuse lead to complacency and the failure to intervene when the technology either fails or degrades. Thus, objectively measuring trust, arguably a difficult task, is important when system reliability and the domain culture could create trust barriers.

³ The phrase "mental models" refers to organized sets of knowledge about the system operated and the environment that are acquired through experience [17].

⁴ Misuse refers to the failures that occur when people rely on automation inappropriately, whereas disuse signifies failures that occur when people reject the capabilities of automation.

3.1.5. The fifth metric class: mission effectiveness

While not represented explicitly in Figure 1, there is a fifth metric class that measures aggregates 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 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.

3.1.6. Metric classes for the supervisory control of one platform

The conceptual model of human supervisory control of Figure 1 represents the need for evaluating five metric classes to understand the performance of a team composed of a single operator controlling a single autonomous platform. These metric classes are interrelated. For example, events in the real world are captured by the platform sensors and presented to the human operator through the display. Different display designs can affect human attention allocation and SA, which in turn will result in changes in human computer interaction (HCI) patterns, which can ultimately affect platform performance. Understanding system performance implies understanding the relations among these elements.

The five generalizable metric classes that emerge from our model to evaluate human supervisory control applications are:

1. Mission Effectiveness (e.g., key mission performance parameters)
2. Autonomous Platform Behavior Efficiency (e.g., usability, adequacy, autonomy)
3. Human Behavior Efficiency
 - a. information processing efficiency (e.g., decision-making)
 - b. attention allocation efficiency (e.g., scan patterns, prioritization)
4. Human Behavior Precursors
 - a. Cognitive Precursors (e.g., SA, mental workload, self-confidence)
 - b. Physiological Precursors (e.g., physical comfort, fatigue)
5. Human - Autonomous Platform Collaborative Metrics (e.g., mental models, trust)

Evaluating the team performance 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, incorporating at least one metric from the other metric classes enables better system performance evaluation.

3.2. Human and platform collaboration in supervisory control

The previous section discussed a model for a one operator-one platform team, but multiple operators can collaborate to control multiple autonomous platforms. We have adapted our single operator-single platform model above to demonstrate how these same metric classes would be characterized in a multiple operator- multiple autonomous platform scenario. In these cases, two additional metrics subclasses for the collaborative metric class are required to evaluate not only human - autonomous platform collaboration, but also collaboration among humans, that is, human - human collaboration metric subclass, and collaboration among platforms, that is, autonomous platform - autonomous platform collaboration metric subclass.

Figures 2 and 3 illustrate a scenario with two humans collaborating while each controlling an autonomous platform. These platforms also collaborate autonomously –collaboration layers are depicted by arrows. As shown in Figure 2, the human - autonomous platform collaborative metric subclass focuses on evaluating the collaboration between the operator and a piece of automation that he controls. However, human - human and the autonomous platform - autonomous platform collaborative metric subclasses are related to the collaboration efficiency among humans, or among different pieces of automation. Figure 3 illustrates the scope of these two metric subclasses using the same conceptual model of human supervisory control that we have discussed previously. These metric subclasses are further discussed in the next sections.

Human / Autonomous Platform Collaboration Metric Class: measures collaboration between the human behavior & the automation behavior control loops

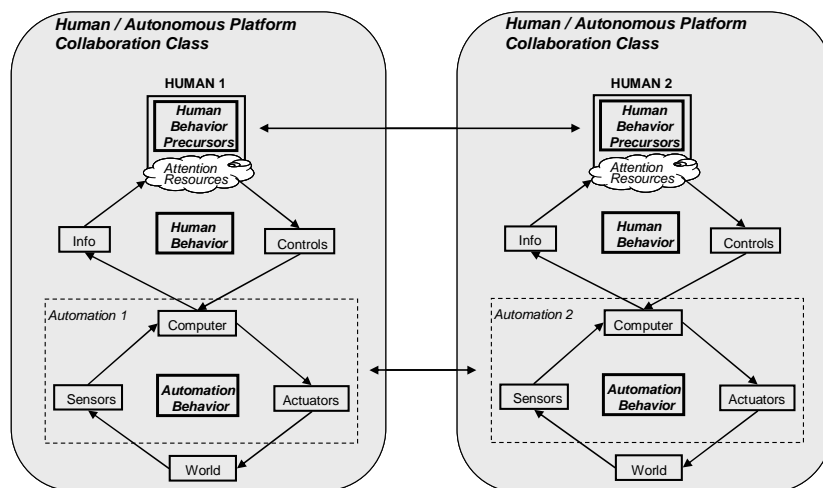
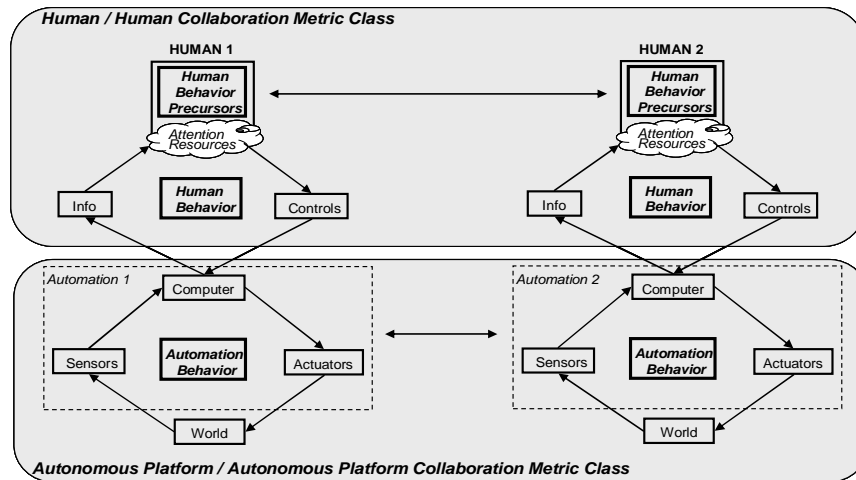


Figure 2: Human / Autonomous Platform Collaborative Metrics

**Human / Human Collaboration Metric Class:
measures collaboration among humans**



**Autonomous Platform / Autonomous Platform Collaboration Metric Class:
measures autonomous collaboration among platforms**

Figure 3: Human - Human & Autonomous Platform - Autonomous Platform Collaborative Metrics

3.2.1. Supervisory control of multiple independent or collaborative platforms

In a supervisory capacity, operators intermittently interact with autonomous platforms, so it is possible that an operator could control multiple vehicles, particularly as onboard automation increases. This section discusses a team of a single operator controlling multiple platforms. In these cases, the operator must continually shift attention among the platforms under his control, maintain situation awareness for the group of platforms, and exert control over a complex system [21].

We have adapted our single operator-single platform model above to demonstrate how these same metric classes would be characterized in a multiple autonomous platform scenario. However, single operator control of multiple platforms can be manifested in two ways: a) multiple platforms performing independent tasks (Figure 4), and b) multiple platforms performing collaborative tasks⁵ (Figure 6).

In the simplest case of an operator controlling two independent platforms, the operator monitors the environment and the platforms' status, decides on which one to focus attention, interacts with that platform, and returns to group monitoring, or decides to service another platform. In the independent multiple vehicle control case, no additional metric classes or subclasses are needed, but there are other considerations for various subclasses. In terms of the human behavior metric class, additional attention

⁵ In this research collaboration between platforms means two or more platforms working together to accomplish a shared goal under human supervision. Also, this research does not distinguish between coordination, cooperation and collaboration.

allocation metrics should be considered such as measuring task/vehicle switching frequency, platform prioritization strategies, and length and quality of vehicle interaction.

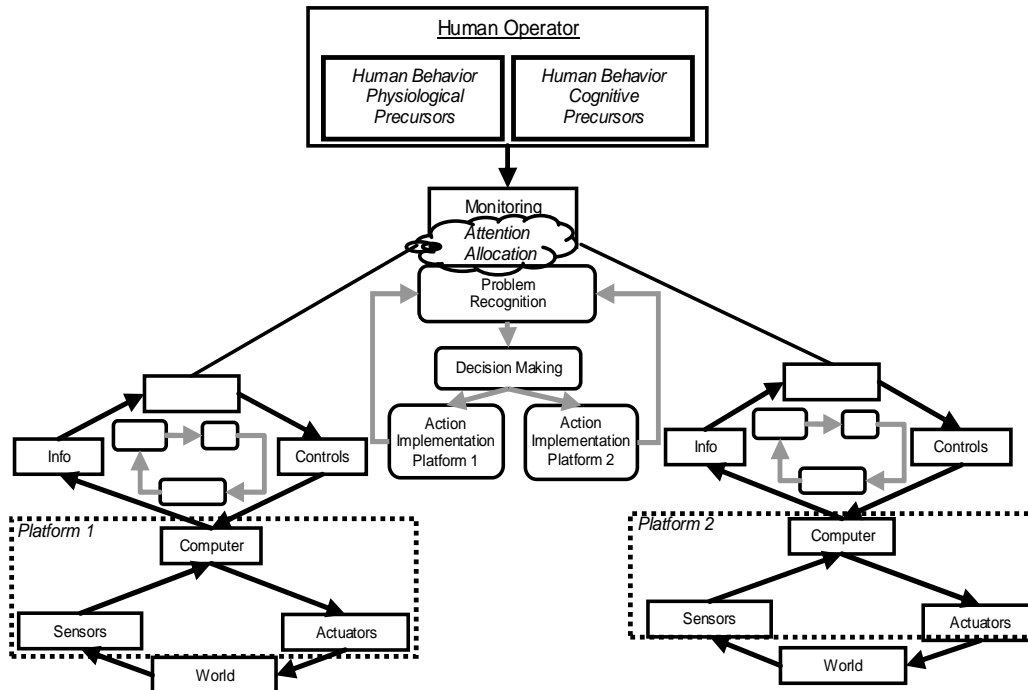


Figure 4: Supervisory Control of Independent Platforms.

In contrast with the independent multiple platform scenario, one operator can supervise multiple collaborative platforms. To perform dependent tasks or collaborative activities, platforms can autonomously collaborate or be manually coordinated by the operator. The case of an automated coordination layer is represented in Figure 5. In this example, the platforms directly coordinate among themselves and behave as a group without the operator's intervention. Collaboration only occurs at the level of the autonomous platform. This motivates the need for the autonomous platform - autonomous platform collaborative metric subclass that evaluates the efficiency of this autonomous collaboration layer. This metric subclass should be measured whenever there is an automated coordination layer among autonomous platforms independently of the number of operators controlling them.

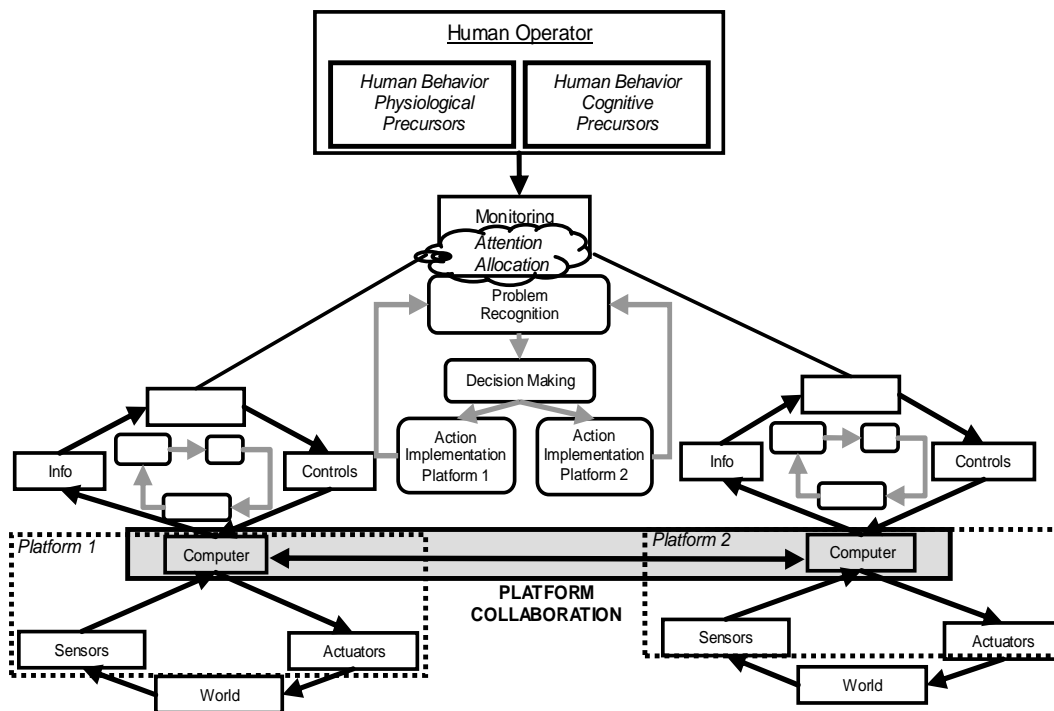


Figure 5: Automated Coordination between Autonomous Platforms.

In reality, coordination tasks are often shared between the operator and the automation because of the need to keep the human-in-the-loop, the mission complexity, the unpredictability of the environment dynamics, and the uncertainty of programming robots' behavior prior to the mission. Figure 6 illustrates the case of active human coordination, where the control loops for platform 1 and platform 2 are not independent and separated entities. Because the control loops of the platforms are no longer independent, servicing the platforms is inherently dependent. Controlling collaborative platforms 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 platform 1 in Figure 6 can involve acquiring and analyzing information related to platform 2, and implementing an action for platform 2 can require synchronizing it with another action for platform 1. 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 station level.

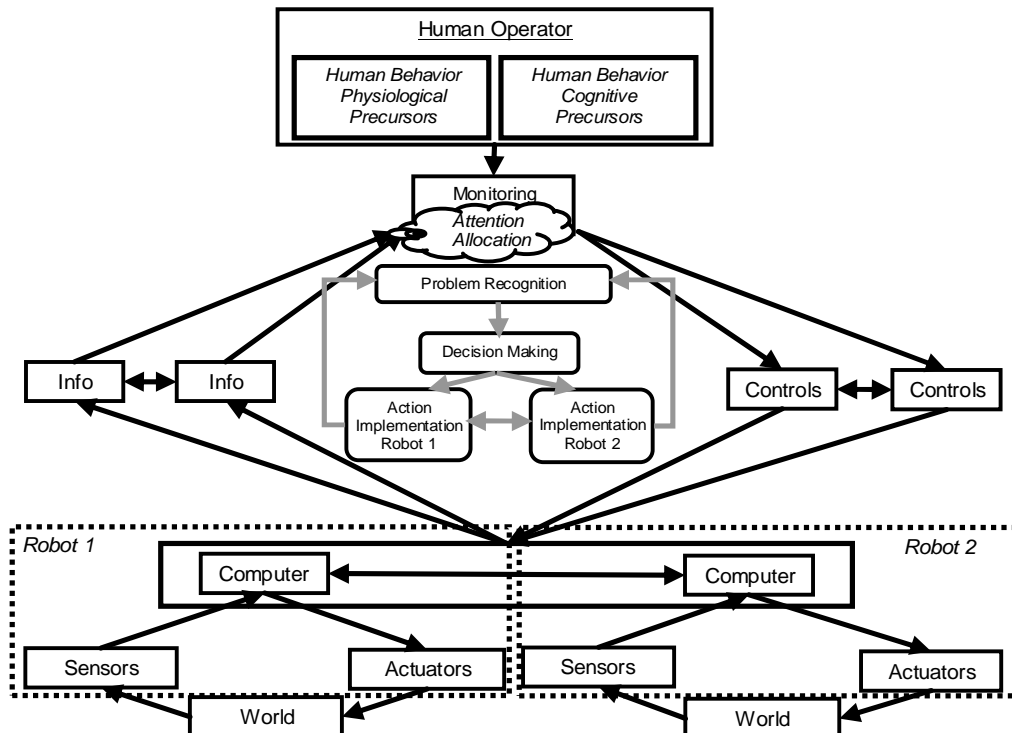


Figure 6: Supervisory Control of Collaborative Platforms.

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

To account for the inter-platform collaboration, a new subclass is needed in the collaboration metric class, which is the autonomous platform - autonomous platform collaboration. In the single operator-single vehicle and single operator-multiple independent platforms models, all collaboration took place just between the operator and the platforms. With collaborative platforms, both the quality and the efficiency of the collaboration among vehicles can also be measured (e.g., information sharing such as path obstacles and the presence of unexpected threats).

3.2.2. Human collaboration in supervisory control of multiple platforms

Given the inherent team nature of command and control operations, the single operator-multiple platform architecture is somewhat artificial and in most cases, will probably be a multiple operator-multiple platform scenarios. Thus, we extend our model to address this configuration (Figure 7). For the collaborative metric class, the previously discussed subclasses (human-autonomous platform and

autonomous platform-autonomous platform collaboration) also apply for the multiple operator, multiple platform system. However, because of the introduction of additional operators, we add the human-human collaboration subclass.

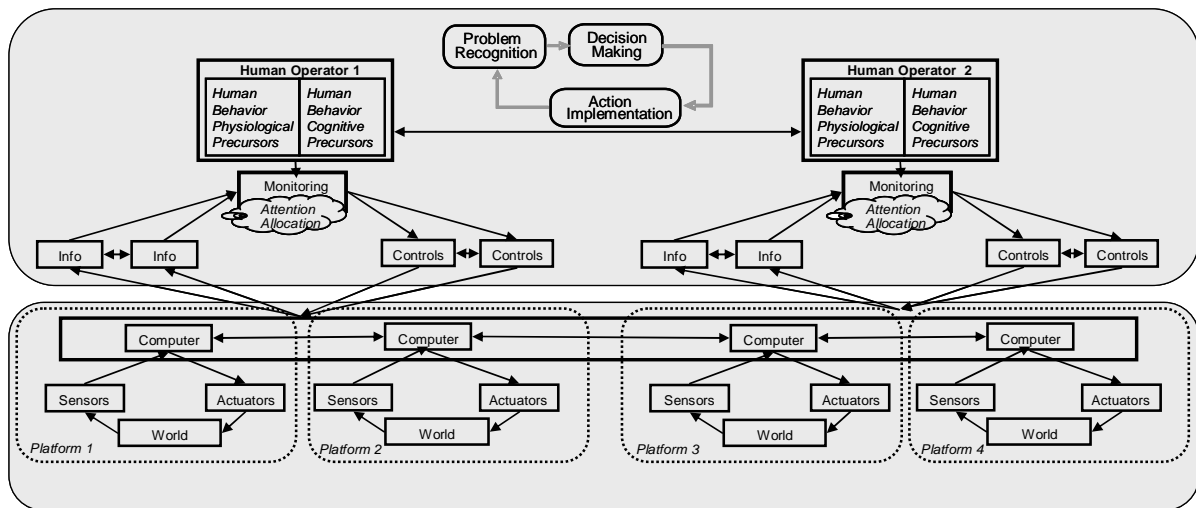


Figure 7: Human Collaboration in Supervisory Control of Platforms.

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 [22]. 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 among team members. Team coordination is generally assessed through communication analysis. Communication analysis can be characterized through two dimensions: physical data vs. content data, and static vs. sequential analyses [23]. Physical measures are relatively low-level measures such as duration of speech, whereas content measures account for what is actually said. Static measures are metrics of team communication at one point in time, or aggregate measures over some duration, whereas sequential analyses account for the ongoing stream of team interaction.

In addition to measuring team coordination for the human-human collaboration metric subclass, measuring team cognition, which refers to the thoughts and knowledge of the team, can be valuable in diagnosing team performance successes and failures, and identifying effective training and design interventions [22]. 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 [24], team mental models and team SA should be considered when evaluating the multiple operator, multiple platform architecture. Evaluating team mental models and SA requires assessing the similarity and consistency of the individual representations and understandings. 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 [25]. In order to capture this aspect, one can measure metrics such as the percentage of knowledge that is redundantly distributed across team members, the percentage of knowledge that is uniquely distributed, and the percentage of knowledge that is not covered.

In addition to team mental model and SA, understanding team cognition can also require evaluating workload distribution and social patterns and roles within the team. Evaluating workload distribution among team members is required in studies where team organization, configuration, or function allocation is explored. Generally, teams are designed so that workload is balanced among their members. Studies that explore team organization, configuration, or function allocation should also consider the existing social patterns, roles, and informal networks within the organization. The study of social patterns and roles is important because team dynamics are often driven by team roles. In addition, designing a team structure and organization that violates the existing social patterns and roles can have a detrimental effect on performance.

4. Generalizable Metric Classes for Human Supervisory Control

Based on the operator-autonomous platform models presented in this report, five generalizable metric classes were identified through a principled approach for human-automation team evaluation. Examples of sub-classes are included in brackets. We have shown that these metric classes apply to any systems of humans and autonomous platforms, regardless of the platform type, and the combination and degree of collaboration between humans and-or autonomous platforms. It is important to note that these classes are not independent, thus in many cases metrics will likely be correlated.

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 the human-automation team. Some metrics may be more valuable than others, and 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.

- 1) Mission Effectiveness (e.g., key mission performance parameters)
- 2) Autonomous Platform Behavior Efficiency (e.g., usability, adequacy, autonomy, reliability)
- 3) Human Behavior Efficiency
 - a) Attention allocation efficiency (e.g., scan patterns, prioritization)
 - b) Information processing efficiency (e.g., decision making)
- 4) Human Behavior Precursors
 - a) Cognitive Precursors (e.g., SA, mental workload, self-confidence, emotional state)
 - b) Physiological Precursors (e.g., physical comfort, fatigue)
- 5) Collaborative Metrics
 - a) Human / Autonomous Platform Collaborative Metrics (e.g., trust, mental models)
 - b) Human / Human Collaborative Metrics (e.g., coordination efficiency, team mental model, team SA, workload distribution, social patterns and roles)
 - c) Autonomous Platform / Autonomous Platform Collaborative Metrics (e.g., platforms' reaction time to situational events that require autonomous collaboration)

5. Populating Metric Classes with Existing Metrics

This section presents the literature review conducted on existing metrics and measuring techniques. The metrics and techniques are discussed in the context of the metric classes and subclasses presented previously.

It should be noted that there is no single metric and technique that is best across all supervisory control applications. Each method entails strengths and weaknesses that the researcher must consider in the context of application. Future research efforts will explore the application of a cost-benefit analysis framework for the selection of most appropriate metrics.

5.1. Mission effectiveness

Mission effectiveness is a measure of how well a team of humans and automation accomplishes some mission. These metrics are mission-specific and are directly identified from mission objectives and goals, and mission success criteria. These metrics are also known as Key Performance Indicators.

For example, Crandall and Cummings conducted an experiment where participants had to control multiple robots to remove as many objects as possible from a maze in a fixed time period [5]. The participants also had to ensure that when the time expired, all robots were out of the maze to avoid being destroyed. In this example, mission effectiveness was measured as the number of objects collected minus the number of robots lost. Another way of thinking about mission effectiveness metrics is to think about the mission objective function. An objective function is generally associated with an optimization problem and determines how good a solution is. In the previous example, the objective function of the game was to maximize the objects collected and minimize the robots lost.

In general, mission effectiveness metrics can be time-based metrics (e.g., speed of performance) in time-critical missions, error-based metrics in safety-critical missions (e.g., number of omission and commission errors), or coverage-based metrics that measure how much of some larger goal is achieved (e.g., percentage of targets destroyed).

5.2. Autonomous platform behavior efficiency

Based on the existing literature, the main metric subclasses for autonomous platform behavior efficiency are usability, adequacy, autonomy, and self-awareness.

5.2.1. Usability

Usability refers to several related attributes. It is traditionally associated with learnability, efficiency, memorability, errors, and satisfaction [9].

Usability techniques can be classified into three main categories: (1) predictive evaluation that usually involves design reviews performed by experts; (2) observational evaluation that is based on observation of users interacting with the system; (3) participative evaluation where information is collected from users based on their subjective reports [26].

Some of the most popular usability techniques are:

- Questionnaires. For example, a three-item questionnaire was developed by Lewis to measure the users' judgment of how easily and quickly tasks were completed [27]. Another popular questionnaire is the System Usability Scale (SUS), a 10-item questionnaire with a Likert scale format [28].
- Heuristic Evaluation. Usability specialists judge whether each dialogue element follows established usability principles [29].
- Cognitive Walkthrough. A group of evaluators inspect the user interface by going through a set of tasks [30].
- Contextual inquiry. This is a structured field interviewing method [31].
- Cognitive interviewing method. Think aloud technique and verbal probing techniques are included in this category.
- Focus Groups.

5.2.2. Adequacy

Automation can greatly affect operators' behavior and team performance, thus it is important to evaluate automation's adequacy to support and help operators complete the mission. This metric subclass contains objective measures of automation accuracy and reliability. In addition, subjective ratings are recommended as a complementary technique for this metric subclass. For example, the Modified Cooper Harper Scale for Unmanned Vehicle Displays (MCH-UVD) is a standardized technique developed for the Department of Defense to identify and categorize deficiencies in unmanned vehicle displays [32]. The key advantage of the MCH-UVD scale is that it guides operators to provide structured feedback about unmanned vehicle display deficiencies. Further, the deficiencies identified within the MCH-UVD 10 point scale are based on human factors design principles. A limitation to the scale is that it is not the "be all end all". It can guide discussion to find key problems, but additional detailed user-feedback and metrics are required for completeness. The MCH-UVD scale is shown in Figure 8.

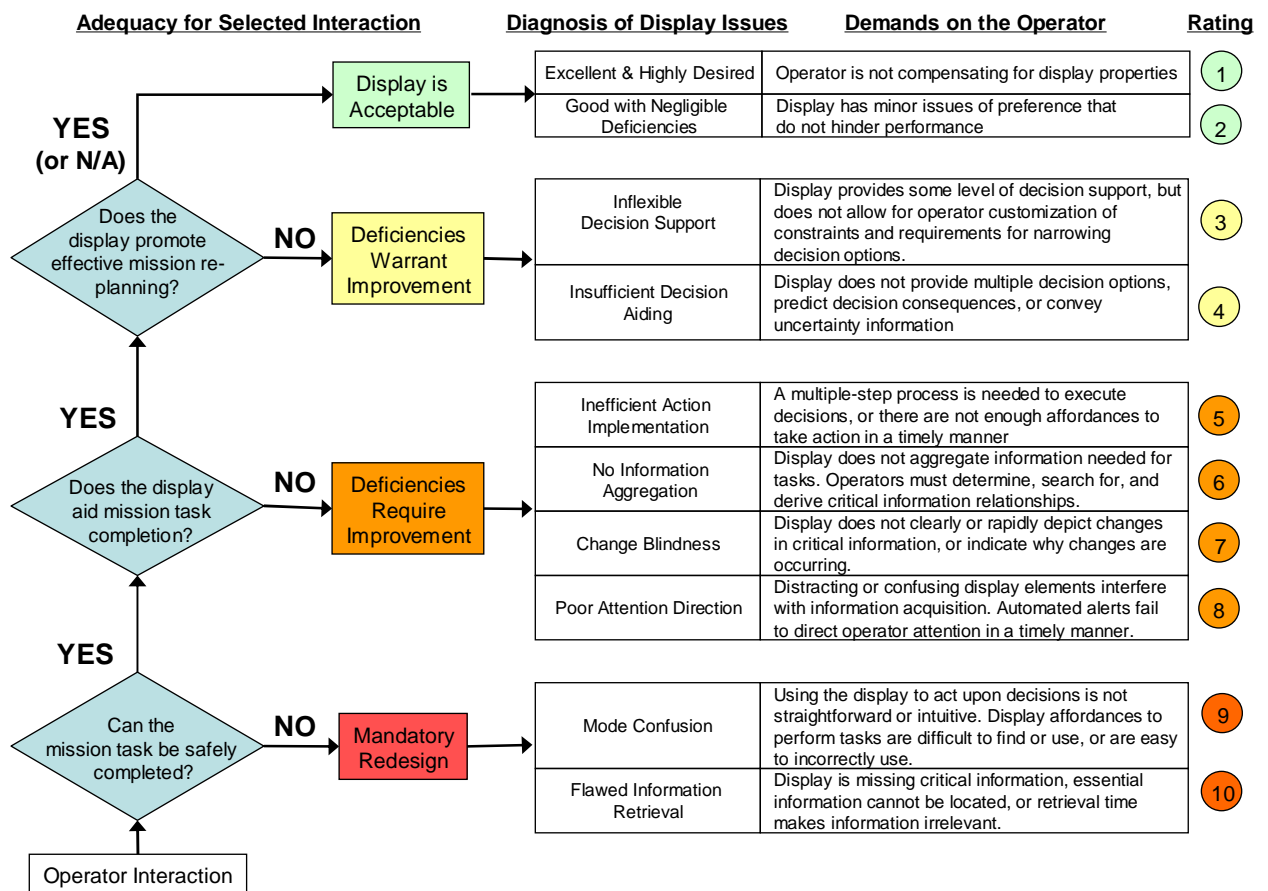


Figure 8: Modified Cooper-Harper Scale for Unmanned Vehicles.

5.2.3. *Autonomy*

Autonomy is the ability of automation to function independently. Neglect tolerance has been proposed as a useful metric for measuring autonomy [33]. Neglect tolerance measures the amount of time the robot can be neglected before performance drops below an acceptable level. Another potential autonomy metric is the automation execution efficiency, which is measured as the relative proportion of time the robot is executing instructions during a mission as opposed to the time it is waiting for directions [34].

A different approach to evaluate autonomy is based on the idea that humans function as a resource for automation to compensate for limitations of autonomy [35]. Therefore, the number of times an operator interrupts his current activity to assist automation can also be used as an autonomy metric.

5.2.4. Self-awareness

Self-awareness is the degree to which automation can accurately assess itself. To qualitatively measure self-awareness, Steinfeld et al. proposed assessing the following automation characteristics: (1) understanding of intrinsic limitations (mobility, sensor limitations, etc.); (2) capacity for self-monitoring (health, state, task progress) and recognizing deviations from nominal; and (3) effectiveness at detecting, isolating, and recovering from faults [3].

5.3. Human behavior efficiency

5.3.1. Attention allocation efficiency

In supervisory control applications, operators supervise a series of dynamic processes, sampling information from different channels and looking for critical events. Evaluating attention allocation efficiency involves not only assessing if the operator knows where to find the information or the functionality he needs, but also if he knows when to look for a given piece of information or when to execute a given functionality [36]. Attention allocation metrics help with understanding whether and how a particular element on the display is effectively used by operators. In addition, attention allocation efficiency metrics also measure operators' strategies and priorities. Main metrics and techniques for attention allocation efficiency are shown in Table 2.

Table 2: Overview of Metrics & Techniques for Attention Allocation Efficiency.

Metrics	Techniques	Measure Examples
Attention Allocation & Operators' Strategies and Priorities	Eye Movement Tracking	Proportion of time that the visual gaze spent within each "area of interest" of the interface
	Human-Computer Interactions	Average number of visits per min to each "area of interest" of the interface Switching time (if multiple autonomous platforms are controlled)
	TRACS	Frequency of use of low level vs. high level information detail per application
	Verbal Protocols	Operators' task and event priority hierarchy

As shown in Table 2, there are three main approaches to study attention allocation: eye movements (e.g., visual attention among the various elements of an interface), hand movements (e.g., human-computer interactions), and verbal protocols (e.g., operators verbalizing their thinking process).

TRACS is included in the hand movements' category because it is based on measuring human-computer interactions.

If operators are controlling multiple autonomous platforms, it can be relevant to study operator's attention distribution among platforms (e.g., percentage of time the operator is engaged with each platform). Other metrics that are found in the literature to measure attention allocation efficiency among multiple platforms are switching times and wait times caused by lack of operator SA (called WTSA) [39]. Switching times are defined as the time required to decide which autonomous platform the operator should service after he has completed an interaction with another platform; WTSA is the time an autonomous platform is in a degraded performance state due to lack of operator SA. This latter metric can be difficult to measure since it must be distinguished from those times that a platform is in a degraded performance state and the operator does not attend it because he is busy attending other tasks.

It should be noted that, in general, we are interested in comparing actual attention allocation strategies with optimal strategies, however, optimal strategies might ultimately be impossible to know. In some cases, it might be possible to approximate optimal strategies via dynamic programming or some other optimization technique. Otherwise, the expert operators' strategy or the best performer's strategy can be used for comparison.

Table 3 summarizes the main advantages, limitations, and recommended use of the metrics and techniques included in Table 2, which will be further defined and discussed in the following sections.

Table 3: Overview of Techniques for Attention Allocation Efficiency.

Technique	Main Advantages	Main Limitations	Recommended Use
Eye Movement Tracking	Continuous measure of visual attention allocation	Noise. Limited correlation between gaze and thinking. Equipment & training. Intensive data analysis	Research phase, in conjunction with other metrics
Human-Computer Interactions	Continuous measure of subjects' actions	Directing attention does not always result in an immediate action. Sensitive to other factors	For interactive interfaces (not for supervisory behavior)
TRACS	Visual representation eases pattern recognition and comparisons	Need to be customized for each interface and task	For interactive interfaces (not for supervisory behavior)
Verbal Protocols	Straight forward. Insight into operators' priorities and decision making strategies	Time intensive. Dependant on operator's verbal skills. Recall problems with retrospective protocols, and interference problems with real-time protocols	Research phase, in conjunction with other metrics

5.3.1.1. *Eye movement tracking*

Extensive research has been conducted with eye trackers and video cameras to infer operators' attention allocation strategies based on the assumption that the length and the frequency of eye fixations on a specific display element indicate the level of attention on the element [37][38].

Attention allocation metrics based on eye movement activity can be dwell time (or glance duration) and glance frequency spent within each "area of interest" of the interface. These measures can be normalized to mitigate strong differences between subjects and enable comparisons, as done by Janzen and Vicente in their study of operator attention allocation in real-time, interactive thermal-hydraulic process control [40].

Visual resources are not the only human resources available. However, as information processing starts with information acquisition, typically through our visual senses, visual attention can be used to infer operators' strategies on the employment of cognitive resources.

5.3.1.2. *Human-computer interactions*

The hand movements, or human-computer interactions, reflect the operators' physical actions, which are the result of the operators' cognitive processes. Thus operators' mouse clicking can be used to measure operators' actions and infer on operators' cognitive strategies. For example, to obtain a comprehensive understanding of their experiments on attention allocation in thermal-hydraulic process control, Janzen and Vicente measured the number of operators' visits to each display window, in addition to the dwell time [40].

It should be noted that some authors have proposed a method for tracking human operator's attention based on a single metric that combines operators' eye and hand movement behaviors [41]. The experimental studies revealed that the inter-relationships between the eye fixation and the mouse clicking, such as the eye fixation over the mouse clicking ratio and the differentiation between the eye fixation and the mouse clicking frequencies, are sensitive to operator's performance variations on different interfaces and to different operators. However, this study presents several limitations and further validation work is still required.

5.3.1.3. *TRACS: Tracking Resource Allocation Cognitive Strategies*

TRACS, a technique based on measuring human-computer interactions, provides a two-dimensional visual representation of operators' strategies during decision-support system interactions. TRACS depicts the user's thought process and actions, allowing for identification and evaluation of where individuals spend cognitive resources [42].

Figure 9 shows an example of TRACS visualization interface. As depicted in Figure 9, the two TRACS axes of MODE and Level of Information Detail (LOID) respectively correspond to the general

functionalities of an interface, as well as the information types available. TRACS assumes that every mouse click on the interface is a conscious decision of the operator to interact with the automation. Using a correspondence matrix for the two axes, each interface click is mapped to a specific MODE and LOID entry in the matrix, and linked to the previous and next clicks. For each click, a circle is added to the corresponding TRACS cell. The width of the circle is proportional to the number of times that particular action is repeated. Two cells are connected by a line when visited in sequence; the thickness of these lines increases each time a connection is repeated.

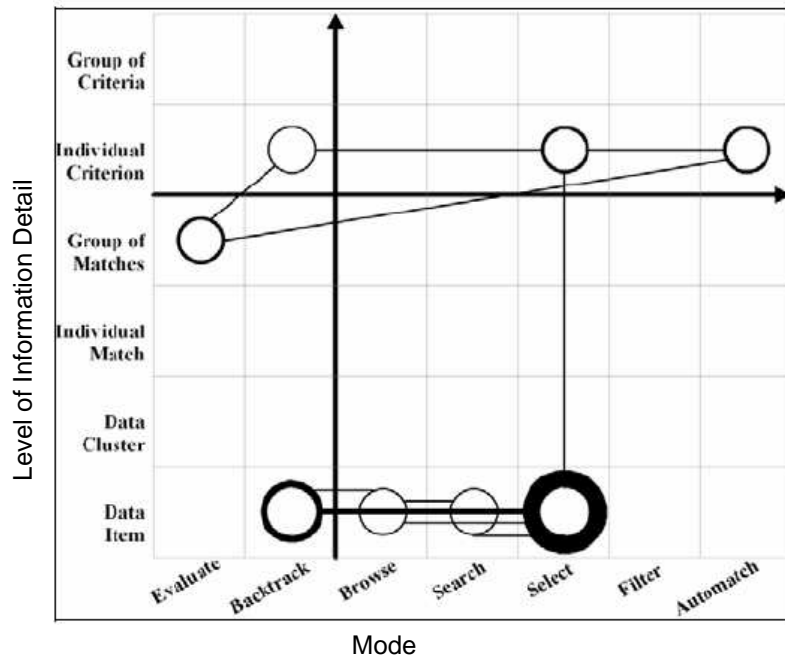


Figure 9: TRACS visualization interface.

Because TRACS is a standardized representation of an individual's cognitive strategy, it can be used to compare strategies between different users performing the same tasks, or to compare strategies across different interfaces.

5.3.1.4. Verbal protocols

Verbal protocols require the operators to verbally describe their thoughts, strategies, and decisions, and can be employed simultaneously while operators perform the task, or retrospectively after the task is completed.

Verbal protocols are usually videotaped so that researchers can compare what the subjects were saying and simultaneously observe the system state through the interface the subjects were using. This technique provides insights into operators' priorities and decision making strategy, but it can be time consuming and depends on operators' verbal skills.

5.3.2. Information processing efficiency

Human supervisory control applications require operators to monitor, process information, make decisions, take actions, and recover from errors if needed. Information processing efficiency metrics are grouped in three main categories: recognition efficiency –which includes task monitoring efficiency and error recovery–, decision making efficiency, and action implementation efficiency. The fourth generic subclass, task efficiency, is included for those cases in which disaggregating human behavior in the previously mentioned three categories is not possible or too costly.

Most popular metrics and techniques for information processing efficiency are summarized in Table 4.

Table 4: Overview of Metrics & Techniques for Information Processing Efficiency.

Metrics	Techniques	Measure Examples
Recognition Efficiency	Human-Computer Interactions	Reaction time Search time Correct recognitions vs. errors Error detection rate Speed of error recovery
	Expert Ratings	Severity of errors Quality of error recovery (impact)
Decision Efficiency	Human-Computer Interactions	Decision rate Number correct decisions / number errors
	Expert Ratings	Quality of decisions
Action Implementation Efficiency	Human-Computer Interactions	Control input activity Movement time
Task Efficiency	Human-Computer Interactions	Interaction time

Human-computer interactions (HCIs) are the observable outputs of human decisions, and they are commonly used to measure human behavior efficiency. However, these metrics can be insufficient to understand the decision making, or the error's magnitude. For that reason, expert ratings can be used as a complementary technique to the analysis of human-computer interactions.

Table 5 summarizes the main advantages, limitations, and recommended use of these three techniques.

Table 5: Overview of Techniques for Information Processing Efficiency.

Technique	Main Advantages	Main Limitations	Recommended Use
Human-Computer Interactions	Continuous measure of subjects' actions	Recognition does not always result in an immediate action.	For interactive interfaces with human manipulation (not for supervisory behavior)
Expert-ratings	Evaluate the impact of errors and decisions in complex systems when errors do not have an immediate effect	Dependant on the observer's expertise. Variability and inconsistency if multiple experts are used	To evaluate complex systems in conjunction with other metrics

5.3.2.1. Recognition, decision-making, and action implementation efficiency

Based on our model, human actions and decisions should 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 system performance. However, disaggregating HCIs may not always be possible.

Figure 10 illustrates a typical timeline for human information processing, which includes the recognition, decision making, and action implementation stages.

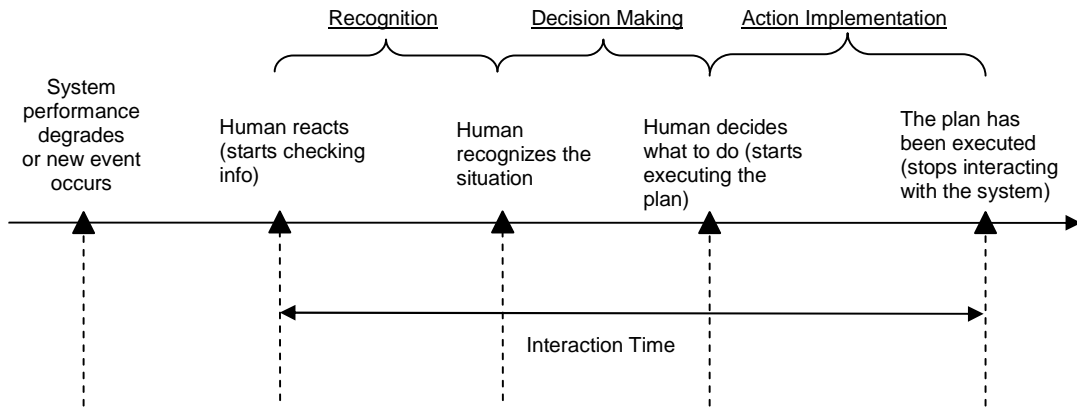


Figure 10: Example of Human Information Processing

Popular metrics from these categories can be found in Gawron's human performance measures handbook [43]. Some examples are:

Recognition Efficiency

- Search time, which is the length of time for a user to retrieve the desired information.
- Recognition time, which is the length of time required by a user to understand and recognize a problem or a given situation.
- Reaction time, which is the time elapsed between stimulus onset and response onset. In order to use this metric, the scenarios should contain observable stimulus (e.g., alarms, or off-nominal conditions or parameters). Reaction time can be seen as a particular case of recognition time.
- Correct recognitions vs. false recognitions.

Measuring recognition efficiency is important because supervisory control applications often require humans to be passive monitors of automated systems and humans are poor monitors by nature [4]. Recognition efficiency also includes the ability of the human to detect automation failures or other kind of errors or off-nominal conditions, which is also known as monitoring efficiency. Metrics for monitoring efficiency are, for example, error detection rate or miss rate. One potential problem with these metrics is that, in most domains, errors or critical signals are very rare, and operators can go through an entire career without encountering them. For that reason, it is not possible to include in an experiment such rare events with sufficient frequency to permit statistical analysis.

In safety-critical systems, error tolerance and error recovery are key issues. Therefore, considering metrics such as speed of error recovery, and quality of error recovery or impact can be important for these systems. However, as it occurs with the assessment of monitoring efficiency, the frequency of appearance of errors in an experiment has to be representative of real life applications, which generally is insufficient for a statistical analysis.

Most of the metrics mentioned above are calculated from human-computer interaction analysis. However, the employment of experts' ratings can be helpful to evaluate, for example, the severity of errors, and the quality of error recoveries.

Decision Making Efficiency

- Decision rate, which is the total number of decisions made divided by the interval elapsed time.
- Correct decisions rate, which is the rate of correct decisions over all decisions.
- Error rate.
- Ratio of number correct decisions / number errors.
- Correctness score, which is a five-point subjective rating developed to evaluate human's problem-solving performance.

Most of the metrics mentioned above are calculated from human-computer interaction analysis. However, the employment of experts' ratings can be helpful to evaluate, for example, the quality of operators' decisions (correct decisions vs. errors, impact of the decision) and how well operators recognized and diagnosed different situations.

Action Implementation Efficiency

- Execution time, the time length required for a user to execute a given plan.
- Movement time, which is the time from the initial touch on the control device to the final lift-off of the finger from the control. This metric can be seen as a particular case of execution time.
- Control input activity, which is the total number of control reversals in each controller axis divided by the interval elapsed time.
- Frequency / rate of tool usage.
- Amount of time no control input was given to the system.

5.3.2.2. Task efficiency

Disaggregating human behavior may not always be possible or efficient. In these cases, generic task efficiency metrics, such as performance metrics (e.g., the number of obstacles avoided by an autonomous vehicle, for example, if navigating is a main task of the mission) and time metrics (e.g., the time required to detect a deviation from the nominal route and correct it) would constitute the information processing metric subclass.

It should be noted that task time is generally not very useful in isolation, but can be useful for comparisons relative to other similar designs of displays or to measure improvements in a display. In particular, interaction time, which is the amount of time a human spends interacting with an autonomous platform to accomplish a given task, is a popular task efficiency metrics [44][45][46].

5.4. Human behavior cognitive precursors

5.4.1. Mental workload

Workload is a result of the demands a task imposes on the operator's limited resources. Thus, workload is not only task-specific, but also person-specific. The measurement of mental workload enables, for example, identification of bottlenecks in the system or the mission in which performance can but does not break down in a particular experiment, and to compare systems that lead to similar performance.

Mental workload metrics can be classified into three main categories: performance, subjective, and physiological metrics. Table 6 summarizes the existing metrics and techniques to measure mental workload.

Table 6: Overview of Metrics & Techniques for Mental Workload.

Metrics	Techniques	Measure Examples
Performance Measures	Primary-Task	Speed or accuracy completing the primary task
	Secondary Task	Time to respond to messages through an embedded chat interface
Subjective Measures	Unidimensional self-ratings	Modified Cooper-Harper scale for workload
	Multidimensional self-ratings	NASA TLX
Physiological Measures	Eye Movement Activity	Blink frequency Pupil diameter
	Electrocardiogram	Heart rate variability coefficient
	Electroencephalogram	Amplitudes of the N100 and P300 components of the event-related potential (ERP)
	Galvanic Skin Response	Skin electrical conductance (in Siemens)

Table 7 provides an overview of the main advantages, limitations, and recommended use of the metrics and techniques included in Table 6.

Table 7: Overview of Techniques for Mental Workload.

Technique	Main Advantages	Main Limitations	Recommended Use
Primary Task Performance	Direct measure on the performance of the system of interest	Insensitive in the “underload” region. Affected by other factors	In conjunction with other workload metrics
Secondary Task Performance	Sensitivity	Interference with the primary task performance	Embedded Secondary Tasks
Self-Ratings	High face validity, cheap, easy to administer	Recall problems	Not to be used with secondary task technique
Techniques to Measure Physiological Parameters	Continuous, real-time measure	Noise-to-signal ratio. Sensitivity to stress. Equipment & Training. Data analysis	Only in laboratory settings

5.4.1.1. *Performance measures: primary and secondary task*

Performance measures are based on the principle that workload is inversely related to the level of task performance [11]. Performance measures can be either on the primary task or on a secondary task.

- Primary Task Performance

Primary task performance should always be studied in any experiment, but this measure presents severe limitations as a mental workload metric. This metric is only sensitive in the “overload” region, when the task demands more resources from the operator than are available. Thus it does not discriminate between two primary tasks in the “underload” region (i.e., the operator has sufficient reserve capacity to reach perfect performance) that impose different demand levels on the operators. In addition, primary task performance is not only affected by workload levels but also by other factors, such as correctness of the decisions made by the operator.

Additional problems of this technique are the difficulty to determine adequate complexity levels for the scenarios so that the metric is sensitive, and the difficulty to make comparisons between different tasks if they differ in how they are measured or what those measures mean (e.g., compare reaction times with precision to follow a certain path).

- Secondary Task Performance

In this technique, a secondary task is imposed on operators as a measure of their residual resources or spare capacity [47]. Secondary task performance is assumed to be inversely proportional to the primary task demands imposed on the operators’ resources. This measure is sensitive to differences in task demands, practice, and other factors that are not reflected in the primary task performance. Some of the secondary tasks that have been proposed and employed are, for example, producing finger or foot taps at a constant rate, generating random numbers, or reacting to a secondary-task stimulus [11].

An advantage of this technique is that it is designed to predict the amount of residual attention an operator would have in case of an unexpected failure or event requiring his intervention. However, the main limitation of this technique is that it may interfere with and disrupt performance of the primary task. However, problems with obtrusiveness can be mitigated if embedded secondary tasks are used. In those cases, the secondary task is part of operators’ responsibilities but has lower priority in the task hierarchy than the primary task. For example, Cummings and Guerlain used an embedded chat interface as an embedded secondary tasking measurement tool [48].

Another potential problem of this technique is derived from the fact that humans have different type of resources (e.g., perceptual resources for visual signals vs. perceptual resources for auditory signals) [49]. Workload differences that result from changes in a primary task variable can be greatly underestimated if the resource demands of the primary task variation (such as reacting to a visual stimuli) do not match those of most importance for secondary task performance (such as reacting to an auditory stimuli).

5.4.1.2. *Subjective measures: self-ratings*

This technique requires operators to rate the workload or effort experienced while performing a task or a mission. Self-ratings have been widely utilized for workload assessment in multi-task environments, most likely due to its ease of use. Additional advantages are their non-intrusive nature, their low cost, and participant acceptability. Disadvantages include recall problems, and the variability of workload interpretations between different individuals. Self-ratings measure perceived workload rather than actual workload. However understanding how workload is perceived can be sometimes as important as measuring actual workload.

Another potential problem is the difficulty that humans can have to introspectively diagnose a multidimensional construct, and in particular to separate physical and mental workload [49]. In addition, it is unclear whether subjects' reported workload correlates with peak or average workload level. Self-ratings complement the information provided by other metrics, but can be of little diagnostic value in the evaluation of the cause of intensive workload in system design. These metrics are recommended in conjunction with other forms of metrics.

Self-rating scales can be unidimensional or multidimensional.

- Unidimensional Self-rating Scale

Unidimensional scale techniques involve asking the participant for a scaled rating of overall workload for a given task condition or at a given point in time. The most popular unidimensional self-rating scales are described in Appendix A.

- Multidimensional Self-rating Scale

Multidimensional scale techniques require the operator to rate various characteristics of perceived workload. This technique presents better diagnostic abilities than the unidimensional scale technique, and it can be used to diagnose causes and determine the nature of workload. However, different humans can differently understand and rate the same dimension, and moreover, they can have problems distinguishing and rating separately each of these dimensions. For example, the NASA-TLX multidimensional scale requests subjects to rate effort, mental demand, and physical demand. The difference between effort and demand can be unclear or even not understood by some subjects. In order to minimize these effects, multidimensional scales should be kept simple and with unambiguous wording, and experimenters should always provide definitions. The most popular multidimensional self-rating scales are described in Appendix B.

5.4.1.3. *Physiological measures*

Physiological parameters such as heart rate, heart rate variability, eye blink rate, galvanic skin response, and brain activity are measures of autonomic or central nervous system activity. Thus, these

measures are indicative of operators' level of effort and engagement, and have also been used to assess operator workload. However, it is important to notice that physiological measures do not necessarily assess workload. These metrics are also sensitive to changes in stress, alertness, or attention, and it is almost impossible to discriminate whether the physiological parameters vary as a consequence of mental workload or the changes are due to these other factors.

An advantage of physiological measures is the potential for a continuous, real-time measure of ongoing operator workload. Such a measure can be used to optimize operator workload, using times of inactivity to schedule less critical tasks or deliver non-critical messages so that they do not accumulate during peak periods [50]. Moreover, this type of measure could be used to implement adaptive automation, which is one technique for optimizing workload [51]. It should be noted that behavioral metrics might also be valid approaches for this type of implementations.

Some problems associated with physiological measures are noise, their sensitivity to emotional factors, and operators' opposition to wear equipment that imposes physical constraints. In addition, validation studies of physiological measures have reported contradictory results. Also, most of this validation work has been done in laboratory settings with controlled experiments with controlled stimuli, making it hard to generalize to real world settings.

The most popular techniques and metrics are:

- Eye Movement Activity

Eye activity measures, which can be obtained with an eye tracker, correlate with cognitive demands and have been used to measure real-time workload. Examples of workload metrics are blink rate and duration, dwell time, fixation frequency, pupil diameter, and saccadic extent. Findings indicate that blink rate, blink duration, and saccade duration all decrease with increased workload, while pupil diameter, number of saccades, and the frequency of long fixations all increase [52].

- Cardiac Functions: Electrocardiogram (ECG)

Heart rate variability is used as a measure of mental load since it is more sensitive to differences in workload than the actual heart rate. Heart rate variability is generally found to decrease as the workload increases [53].

- Brain Activity: Electroencephalogram (EEG)

The electroencephalogram is the only physiological signal that has been shown to accurately reflect subtle shifts in workload that can be identified and quantified on a second-by-second timeframe. However, it also reflects subtle shifts in alertness and attention, which are related to workload, but can reflect different effects. In addition, significant correlations between EEG indices of cognitive state changes and performance have been reported based on studies conducted in different domains and environments [54][55][56].

- Galvanic Skin Response (GSR)

Galvanic skin response (GSR) is the change in electrical conductance of the skin attributable to the stimulation of the sympathetic nervous system and the production of sweat. Perspiration causes an increase in skin conductance, thus GSR can be indicative of workload, as well as stress levels [57].

5.4.2. *Situation awareness*

Applying the framework developed by Drury et al. [10] to teams of humans and autonomous platforms, we can define human situation awareness⁶ as:

- the human understanding of
 - locations, identities, activities, status, and surroundings of the autonomous platforms,
 - and the overall goals of the joint human-automation activities and the moment-by-moment measurement of the progress obtained against the goals,
- and the certainty with which the human knows the afore mentioned information.

In addition, the term “understanding” refers to both the understanding of the current situation and dynamics, and the ability to anticipate future-situation events [15]. Presumably, good operators have a better understanding of the current state than poor operators do, but expert operators differ from intermediate operators because of their better predictions of the future [58].

Situation awareness metrics can be classified into two main categories: implicit and explicit metrics. The implicit metrics comprise the performance-based measures and the process-based measures. The explicit metrics comprise the subjective measures and the query-based measures. Table 8 summarizes the existing metrics and techniques to measure situation awareness.

⁶ In this research the term “situation awareness” refers to “human situation awareness”. In order to refer to the situation awareness that the autonomous platforms have we will use the term “platform situation awareness”.

Table 8: Overview of Metrics & Techniques for Situation Awareness.

Metrics		Techniques	Measure Examples
Implicit Metrics	Process-based Measures	Communication Analysis	Communication rate Anticipation ratio
		Verbal protocols	Operator feeling confused or lost Operator double-checking information
		Eye tracking	The point-of-gaze (EPOG) Fixation times
	Performance-based Measures	General Behavioral Measures	Time required to return to the original flight path after a deviation
		Testable Responses	Speed & accuracy of operators' response
		Global Implicit Measure	Performance score showing progress toward accomplishing task goals
Explicit Metrics	Subjective Measures	Observer ratings	Neutral expert rating participant's level of SA using a Likert-type scale ranging from "1" to "7"
		Self-ratings	Participants rating the amount of SA experienced using a Likert-type scale ranging from "1" to "7"
	Query-based Measures	Off-line query methods (memory-based)	Accuracy of operators' response
		On-line query methods (perception-based)	Operators' response time
		Post-experiment questionnaires	Accuracy of operators' response

Table 9 provides an overview of the main advantages, limitations, and recommended use of the metrics and techniques included in Table 8.

Table 9: Overview of Techniques for Situation Awareness.

Technique	Main Advantages	Main Limitations	Recommended Use
Communication Analysis	Continuous measure of SA	Time intensive. Communicativeness can be non-correlated with knowledge	To measure team SA, not individual members' SA
Verbal protocols	Insight into the operator's cognitive processes	Time intensive. Depends on operator's verbal skills. Recall problems with retrospective protocols. Interference problems with real-time protocols	Research phase, in conjunction with other metrics
Eye tracking	Insight into developing SA and processing info. Continuous measure of SA	Noise-signal ratios. Non-correlation between gaze and thinking. Equipment & training. Intensive data analysis	Research phase, in conjunction with other metrics
General Behavioral Measures	Evaluate ultimate consequences of a given knowledge state and users' recovery state	Sensitivity to other factors such as skill level, strategies, and workload	In conjunction with Testable Responses
Testable Responses	Evaluate ultimate consequences of a given knowledge state and users' recovery state	Not possible to test concurrently workload and performance. Repeatability. Interferes with subject's attention allocation	Final testing phases of an operational system
Global Implicit Measure (GIM)	Non-intrusive. Continuous metric	Cost. Sensitive to other factors such as workload	Not recommended
Observer ratings	Observer knows true state of affairs. Non-intrusive	Observer does not know operator's internal understanding	Field testing in conjunction with Testable Responses
Self-ratings	Inexpensive. Easy to use. Face validity. Non-intrusive	A metacomprehension ⁷ metric. Recall problems if administered post-trial	Operational tools, to promote user acceptance

⁷ Metacomprehension refers to knowing what is needed as knowledge and to knowing how much of that knowledge you have.

Technique	Main Advantages	Main Limitations	Recommended Use
Off-line query methods (memory-based)	Objective and direct measure. Assesses global SA	Recall problems. Intrusiveness of freezing scenarios (interferes with performance)	Research phase or studies where system performance is not measured
On-line query methods (perception-based)	Objective and direct measure	Interference with workload, attention, and primary task performance	Recommended if queries are embedded in the tasks
Post-experiment questionnaires	Easy to use. Non-intrusive	Recall problems. Punctual SA evaluation	Recommended in conjunction with other metrics

5.4.2.1. *Process-based measures*

Process measures examine the operator's cognitive processes upon which situation awareness is built.

– Communication Analysis

This technique requires examining the verbal exchanges between people involved in a task. Individual team members' SA awareness is inferred from what they say.

Some examples of SA metrics based on communication analysis are the total communication rate, the anticipation ratio (i.e., communications transferring information/communications requesting information), or communication content analysis (e.g. quality and frequency of communications by type).

Communication analysis can be very helpful to analyze team processes but not so much to understand team members' SA: an operator may know a lot but be uncommunicative. In addition, it does not tell us about what information is being processed, or how that information is being integrated and utilized. Also, this technique can be tedious and time intensive.

– Verbal Protocols

Verbal protocols require the operators to verbally describe their thoughts, strategies, and decisions while interacting with the system, and they are also known as "thinking aloud" protocols. This technique can be executed concurrently during the experiment or retrospectively requesting the subject to review a video recording of the experiment. Retrospective protocols are less intrusive but rely more on subjects' memory and require longer experimental sessions. The ability to infer operator's SA from this technique is limited. However, it can provide much insight into the cognitive processes employed to perform tasks. Its effectiveness is determined by the verbal skills of the operator. In addition, analyzing

verbal protocols can be very time consuming. Usually, this technique provides an incomplete picture of the situation.

Since verbal protocols are based on operators' thoughts and strategies, some examples of actual SA metrics can be the number of times the operator got confused or felt lost, or the number of times the operator needed to double check some information.

– Eye Tracking

This technique is based on tracking operators' eye movements during task performance. It is believed that eyes and eye-movements can tell us some things about the human mind. Some examples of SA metrics based on eye tracking are fixation times, visual scan patterns, and the eye point of gaze (EPOG). Identifying where someone is looking helps us infer the information being processed at each moment and understand the interaction between the operator and the display. Moreover, important information as well as badly displayed or confusing information can lead to long eye fixations.

Existing literature is contradictory in terms of the usefulness of this technique to assess SA. Durso et al. studied chess players' SA and concluded that the eye movements seemed to be the most complicated of all the SA methodologies studied and yielded the fewest insights [59]. However, Smolensky concluded that "evaluation of eye movements might be used to assess certain contributing constructs to SA [60]."

The main limitation of this technique is the difficulty in getting the critical signal-to-noise ratios. In addition, it is unclear whether where one looks tells you what one is thinking about. For example, focusing on, or tracking an object does not necessarily demonstrate a high level of awareness. Moreover, a subject can be aware of an event or an object in the field of vision even if it has never been focused or directly tracked. Another limitation of this technique is that these measures rely on visual information, and human awareness is also affected by other modalities, such as auditory information. The need for specialized equipment and training, and the extensive expertise and time required to analyze the data also limit the applicability of this technique.

Eye tracking is useful to conduct research and understand how people develop SA and process information in complex environments. However, its ability to infer SA as a state of knowledge for design evaluation is limited. This technique is recommended to understand operators' attention allocation strategy rather than their actual SA.

5.4.2.2. Performance-based measures

Performance measures examine the operator's observable response and actions and their impact on the system, given their knowledge state, rather than directly examining the operator's knowledge state. These measures are very useful to answer research questions such as "does the user have sufficient situation awareness?", "does operator-achieved SA lead to the desired system performance?"

The most popular metrics and measuring techniques are:

– General Behavioral Measures

Operator's behavior reflects his actual knowledge, as well as the perceived reliability of the knowledge, both important components of SA. However, these metrics are also sensitive to other factors such as workload, expertise, or decision strategy. Discerning whether people are behaving in a certain manner because of their SA or as the result of lack of skill, or poor strategies, or excessive workload is often difficult. Moreover, an expert participant may be able to achieve acceptable performance even when his SA is inadequate. Therefore, behavior measures are often used in combination with scenarios where situations have been created that, given an acceptable level of SA, will cause the operator to react in a predicted manner. It should also be noted that the ability to infer SA from performance measures depends highly on the scenarios constructed and the type of performance measures (e.g. nominal vs. non-nominal conditions) [61].

– Testable Responses

This technique consists on including SA revealing events embedded in the scenarios such as errors, unexpected incidents, or rare situations. Users are presented with realistic situations, which if they have sufficient SA, require decisive and identifiable actions [62]. This technique measures real-time responses of operators in time-critical situations. For example, Midkiff and Hansman in a flight simulator study allowed subjects to overhear communications which suggested that another aircraft had not departed the runway the subjects were very close to landing on [63]. In this case, action was required to avoid a collision; a lack of action was considered as a lack of situation awareness.

The most crucial aspect is the design and scripting of the situations so that a clear and unambiguous response is mandated if operators have sufficient SA. In addition, incidents should happen at a rate which is realistic and reasonable. In terms of subjects' behaviors, all probable actions should be evaluated. In the example of Midkiff's and Hansman's experiment, most subjects' reaction was to query ATC to confirm the information rather than immediately start a go-around procedure.

The SA metric can be the speed and/or the accuracy of operators' response. The number of strong reactions, uncertain and weak responses, and the lack of action are also SA metrics.

It should be noted that this technique can only be used in simulation environments. It can be combined with the observer rating technique so that domain experts rate the appropriateness of the subjects' reaction and actions.

A main limitation of this technique is having enough data so that a statistical analysis can be conducted. The main reasons are the difficulty of making situations repeatable for different subjects, and the limited number of incidents that the same user can be exposed to without biasing his attention allocation strategy. Finally, this technique should not be used during concurrent testing of workload or performance because the inclusion of unexpected or unusual events embedded in the scenario results in workload and performance not being representative of nominal conditions.

- Global Implicit Measure (GIM) developed by Brickman et al. [64]

The GIM technique provides an objective and real-time measure of situation awareness by comparing human performance during complex tasks against previously defined behavioral constraints (e.g. rules of engagement). It was originally developed for pilots. This technique is based on the assumption that the pilot is attempting to accomplish known goals at various known priority levels. Therefore, it is possible to consider the momentary progress toward accomplishing these goals [65]. In this approach, a detailed task analysis is used to derive rules that link measurable behaviors to the accomplishment of mission goals. Each of these rules is treated as an individual implicit probe and assessed at the frame rate of the simulation. For any time period during the mission, the SA metric can be calculated as the proportion of the implicit probes that have been correctly accomplished by the operator at to that moment.

The main limitation of this technique is cost since it requires multiple interface and scenario-specific GIM scoring algorithms to be developed and tested. In addition, one can argue that the GIM only provides a measure of performance and that deviations from the prescribed rules are not necessarily a result of poor SA.

5.4.2.3. *Subjective measures*

Subjective measures require the operator or a field expert to make judgments about their or other's knowledge state. For example, on a given scenario or task, a participant might be asked to use a Likert-type scale ranging from "1" to "7" in rating the amount of SA experienced. In the case of observer ratings, an unbiased, neutral expert is asked to observe a participant perform a task and rate the participant's level of SA.

Self-ratings measure metacomprehension rather than comprehension of the situation; it is unclear whether or not operators are aware of their lack of SA. In addition, humans seem to have different opinions on what SA actually is, estimating the SA experienced differently. Vidulich & Hughes found that about half of the participants in their experiments rated their SA by gauging the amount of information to which they attended; while the other half of the participants rated their SA by gauging the amount of information they thought they had overlooked [66].

However, it is important to evaluate both objective and subjective SA and make sure that both coincide [13]. Errors in perceived SA quality, over-confidence or under-confidence in SA, may be as harmful and affect individuals' or teams' decision making as errors in their actual SA [67]. In addition, subjective measures have the advantage of being non-intrusive and easy to use, and require minimal training. To ensure comprehensiveness subjective measures can be used in conjunction with other measures such as query-based measures, which are presented below.

In the case of observer ratings, observers have information regarding the true state of affairs but cannot observe the operator's internal understanding of the situation. For example, an operator could be aware of a piece of information but he could provide no observable evidence of this knowledge. Observer

rating techniques are most commonly used to assess SA during tasks performed “in-the-field”. In the case of a simulated environment, we recommend to use observer ratings in conjunction with the Testable Response Technique.

The most popular SA subjective metrics and techniques are described in Appendix C.

5.4.2.4. *Query-based measures*

Query measures are objective and direct measures, which require the operator to report pieces of task-relevant information. Query-based measures attempt to directly capture the operator’s state of knowledge and require a detailed analysis of SA requirements.

The most popular query metrics and techniques are:

Off-line query methods (memory-based)

The off-line query method is based on briefly halting the simulation at randomly selected intervals, blanking the displays, and administering a battery of queries to the operators. The SA metric assesses global SA by calculating the accuracy of operator’s responses compared to the reality.

The most popular off-line method is the Situation Awareness Global Assessment Technique (SAGAT). This technique has been validated and applied across a variety of domains including aviation, air traffic control, power plant operations, teleoperations, driving, and military operations [68].

According to the creator of the method, its main challenges are:

- administering the questions at the right time during the simulation so that the questions do not interfere with operator’s performance, workload, and attention, but still provide a comprehensive picture of operator SA,
- and determining the appropriate questions, which cover the entire range of relevant SA issues.

Due to the nature of this method, it can only be used in a simulation environment. In addition, some researchers have pointed out that recall problems can limit the applicability of this method: if the operator does not have a good picture of the situation when queried, that does not mean that he or she did not have the picture while performing the task [13].

On-line query method (perception-based)

In the on-line query method⁸, operators are presented with queries about the situation while the situation remains present and while they continue to perform the primary task [69]. This method leaves the operator in context and assumes that knowing where to find a piece of information is indicative of good SA. The SA metric is the operator’s response time. If an operator has the answer to the query in his active memory, response time should be short. If the information is not available, but the operator knows

⁸ This method is also known as the real-time probes method.

where to find it, then response time will be longer, but not as long as the case in which the operator does not know where to find the info [69].

The most popular on-line query method is the Situation Present Assessment Method (SPAM). According to the SPAM method developer, answers' accuracy tells us about SA when it fails while response time can help us in investigating what happens when SA succeeds [13]. A limitation of this method is the potential correlation between real-time probes and workload. Operators' speed of response is dependent on workload and spare capacity, thereby raising the concern that real-time probes may reflect measures of workload as well as SA [68][70].

Post-experiment questionnaires

In the case of post-experiment questionnaires, operators are presented with the queries after they finish the experiment. The SA metric is the accuracy of operators' response. This method only assesses SA at one point in time, after the experiment, which limits the ability of this metric to represent operators' SA along the experiment. In addition, operators are required to retrospectively recall their status of knowledge.

A questionnaire administered after an experiment can provide limited information about the process and operators' knowledge state. However, it is low-cost, easy to use, and non-intrusive, so it is recommended if used in conjunction with other metrics.

5.4.3. Self-confidence

Operators' self-confidence in their own abilities plays a major role in operators' effective use of automation [71]. Research results suggest that people are often overconfident in their abilities, both in forecasting future events [72] and in their general knowledge [73]. Overconfidence can result in operators less and less likely to delegate control to automation, and thus failing to benefit from the capabilities of the automation. On the contrary, lack of confidence can result in excessive reliance on automation and failure to intervene when needed.

Self-confidence is measured with subjective ratings. For example, Lee and Moray asked subjects to rate how high their self-confidence was in controlling the different parts of a simulated semi-automatic pasteurization plant [71].

5.4.4. Emotional state

Emotions and moods are temporary feelings that affect behavior. Thus, the state of a person's emotions and the mood can affect system performance. However, it is out of the scope of this research to measure individual emotional differences. This metric subclass is included in the model for completeness, but specific metrics and techniques to measure the emotional state of subjects are not discussed.

5.5. Human behavior physiological precursors

5.5.1. *Physical workload*

Physical workload is defined and measured in terms of energy expenditure. Traditionally, human physical work is measured in kilocalories and oxygen consumption [74].

However, measures of physical workload are becoming less and less relevant, in particular in human supervisory control domains, where automation is assuming the functions that require large forces and that once were the responsibility of the human operator.

While it is still possible for an operator to become physically fatigued, specially during an emergency when some of the automation fails, it is far more likely for designers to worry about mental, rather than physical, overload [75].

5.5.2. *Fatigue*

Fatigue tests are important, for example, to determine appropriate shifts by answering questions such as 'how tired do you have to be before your performance might be appreciably affected?'

A considerable number of studies have measured fatigue effect by measuring the related reductions in performance on tests like Simple Reaction Time and vigilance tasks, like the Mackworth Clock test developed to evaluate vigilance in British Air Force radar technicians during World War II [76], both in laboratory and field studies [77][78][79].

Physiological measures, such as adrenaline and noradrenaline production, cortocosteroid production, brain electrical activity, eyelid closure, eye position/eye movement, heart rate, and gross body movement have also been found sensitive of the onset and detection of fatigue [80]. Electroencephalography (EEG), which provides insight into cerebral arousal, is one of the most popular physiological fatigue measures.

Despite the variety of objective measures available, fatigue remains essentially a subjective experience [80]. Thus, self-ratings of fatigue are very popular in the literature. However, it should be noted that physiological fatigue may not be detected until extreme subjective fatigue is reported [81]. Fatigue due to lack of sleep can be underestimated, and fatigue due to physical activity can be overestimated; activity, such as muscular effort, can also alter the perception of sleepiness.

5.5.3. *Physical comfort*

Humans are sensitive to the conditions in their workplace, such as temperature, light, noise or even the chemicals being used. If these sensitivities are unable to be managed within the workplace, performance may decline, thus it can be important to evaluate operators' physical comfort. However, it is

out of the scope of this research to discuss metrics and techniques to measure physical comfort. This metric subclass is included in the model for completeness.

5.6. Collaborative metrics

5.6.1. Human - autonomous platform collaborative metrics

5.6.1.1. Autonomous platform - human awareness

Autonomous platform - human-awareness is the degree to which the autonomous platform is aware of humans, including humans' commands and any human-originated constraints that may require a modified course of action or command noncompliance. Depending on the application, automation may need to have knowledge of humans' expectations, constraints, and intents. "Awareness violations" that occur during the execution of the task, for example, robots running into victims in a search and rescue mission, has been proposed as a metric to measure human-awareness [16].

5.6.1.2. Trust

Trust is a human attitude toward automation that affects reliance. Trust concerns an expectancy regarding the likelihood of favorable responses [19]. People tend to rely on and use the automation they trust and tend to reject the automation they do not. Operators' lack of trust in automation often forms a barrier, thwarting the potential that a new technology offers. On the other hand, excessive trust results in complacency and the operator failing to intervene when the technology fails.

The metrics available to measure trust can be classified into two main categories: implicit and explicit metrics. The implicit metrics refer to measures based on operators' use of automation. The explicit metrics are subjective measures. Table 10 summarizes the existing metrics and techniques for trust.

Table 10: Overview of Metrics & Techniques for Trust.

Metrics		Techniques	Measure Examples
Implicit Metrics	Use of Automation	Human-Computer Interactions	Frequency and duration of manual control Actions abstraction level and information integration level
Explicit Metrics	Subjective Measures	Unidimensional Rating Scale	Lee and Moray trust scale
		Multidimensional Rating Scale	Human-Computer Trust (HCT) scale

Table 11 provides an overview of the main advantages, limitations, and recommended use of the metrics and techniques included in Table 10.

Table 11: Overview of Techniques for Trust.

Technique	Main Advantages	Main Limitations	Recommended Use
Human-Computer Interactions	Objective measure	Sensitive to other factors such as workload and self-confidence	For interactive interfaces with human manipulation (not for supervisory behavior)
Unidimensional trust rating scale	Direct measure of a purely psychological state	Difficult to capture the complexity of trust in one single dimension	If no particular barriers for automation adoption, such as system reliability and cultural issues, are expected
Multidimensional trust rating scale	Direct measure of a purely psychological state	Limited validation evidence	Research phase or early stages in the design process

5.6.1.2.1. *Implicit measures: use of automation*

Lee and Moray's research showed a strong relationship between operators' trust and their reliance on automation [71]. Thus, operators' use of automation can be an indirect indication of operator's level of trust.

Simple implicit trust metrics such as the frequency of use of certain tools, and the frequency and duration of manual versus automated control can be obtained from human-computer interactions. The abstraction level of the actions performed⁹, and the integration and processing level of the information accessed by the operator can also be indicative of operator's trust levels. Low reliance on automation can result in actions with low abstraction levels, such as manually flying an aircraft, or in accessing information with low processing and integration levels, such as constantly checking the raw data. On the contrary, performing actions with high abstraction levels, or accessing information with high processing and integration levels can indicate excessive reliance on automation.

The main disadvantage of implicit measures is that the use of automation is not only affected by trust but also by other factors such as workload, time criticality of the situation, and self-confidence. Lee &

⁹ The abstraction level of an action refers to the degree the action is described in terms of physical processes or system instances; actions at the lowest level are defined in terms of physical components and processes, whereas high level actions are described in terms of purposes and goals. For example, *aviate* is at a lower abstraction level than *navigate*, but *navigate* is at a lower abstraction level than *payload management*.

Moray demonstrated that shifts between automation versus manual control model can be predicted by the ratio between trust and self-confidence [71].

5.6.1.2.2. Explicit measures: subjective ratings

Some authors consider that trust is a purely psychological state and can be assessed only by subjective rating [82]. Thus, the use of subjective questionnaire-based rating scales is the most common means of measuring trust. These rating scales usually include several levels because, as Muir and Moray concluded, trust is not a discrete variable, but variable levels of trust can exist between none and total [83].

The most popular subjective rating techniques are:

Single rating scale to evaluate operators' overall trust [71]

Lee and Moray used a simple ten point rating scale from "not at all" to "completely" in their experiments to evaluate operators' overall trust. After completing their task, subjects were asked questions such as "how much did you trust the automatic controller of the steam pump?"

The main drawback of this technique is the difficulty of capturing the complexity of a multidimensional construct such as trust with a unidimensional rating scale.

Multiple rating scales to elicit dimensions of trust

Multidimensional rating scales present better diagnostic abilities than the unidimensional scales. However, humans can have problems distinguishing and rating separately individual dimensions such as integrity, reliability, accuracy, dependability, or confidence. These terms refer to different automation aspects, but often generate the same responses from participants. In addition, the same word can have different meanings for different subjects, in particular words such as reliability, accuracy, and even trust that seem to depend on the cultural and national background of the subject. Thus, multidimensional scales should be kept simple and with unambiguous wording, and experimenters should always provide definitions.

Muir and Moray questioned their subjects about three aspects of automation to estimate trust: the degree of trust in automation's display, the degree of trust in automation responding accurately, and the overall degree of trust in automation [83]. Other popular and more complex multiple rating scales are described in Appendix D.

5.6.1.3. Mental model efficiency

Mental models allow people to describe and understand phenomena, draw inferences, make predictions, and decide which actions to take. Evaluating humans mental models on automation and the

mission is important because it can feed directly into the automation design and the content of training materials.

Since individual's mental model reflects the individual's perception of reality, mental models vary in their accuracy and coherence [84]. In general, evaluating mental model efficiency comprises assessing their accuracy and coherence.

Main metrics for mental model efficiency are shown in Table 12.

Table 12: Overview of Metrics for Mental Models Efficiency.

Metrics	Measure Examples
Accuracy of a Mental Model	Similarity of participants' mental model and that of subject matter experts
	Similarity of participants' mental model and that of the best performer
Coherence of Mental Models	Similarity of participants' mental model and system architecture
	Similarity of participants' mental model before and after the experiment
	Similarity among participants' mental models

Assessing the accuracy of mental models is important because those of experts differ from those of non-experts, and moreover, mental models can predict individual performance. For example, Rentsch et al. found in their experiments that participants who reported high experience tended to use fewer categories or dimensions to describe a concept, used more abstract definitions, and represented their knowledge more consistently than those reporting low experience [85].

Regarding the coherence of mental models, it is important to evaluate the consistency between participants' mental models and system architecture, and the stability of mental models. The similarity of participants' mental models before and after the experiment can provide an indication of mental models' stability. Furthermore, for new systems and concepts of operation, it can be important to identify and understand the mental model of the participant who performed best in the experiment so that effective training programs can be designed.

Mental model evaluation is composed of two main stages: knowledge elicitation, and mental model representation and analysis. Knowledge elicitation is the practice of explicating the domain-related knowledge held by an individual; it includes the concept generation (i.e., determining the concepts that the participants will develop mental models of) and the concept rating (i.e., describing the relationships between the concepts). The mental model representation and analysis includes deriving the actual mental models and determining their similarity to other mental models.

The concepts to be rated can be generated by subject matter experts or participants. In the latter case, participants are not restricted by the concepts the subject matter experts generate, and more realistic mental models can be elicited. However, comparisons across participants can be difficult.

Table 13 includes the most popular techniques for knowledge elicitation, and Table 14 the most popular ones for mental model representation and analysis. These techniques are described in Appendices E and F.

In general, mental model elicitation techniques are time consuming, labor intensive, imply cognitively draining tasks for the participants, and relatively complicated analyses. Most recommended techniques are: visual card sorting technique, causal mapping (coupled with the distance ratio formula), and pairwise ratings (coupled with multidimensional scaling or Pathfinder algorithm) [86].

Table 13: Overview of Techniques for Knowledge Elicitation.

Technique	Main Advantages	Main Limitations	Recommended Use
Cognitive Interviewing	Straight- forward	Relies on interviewer's abilities and interpretations	Initial research stage to familiarize with the domain
Verbal Protocols	Explicit elicitation of participants' strategies	Incomplete picture. Difficulty to compare among participants	To study operators' cognitive strategies, in conjunction with behavioral metrics
Pairwise Rating	Time efficiency	Repetitive nature of pairwise ratings can induce a response set	When research time is constrained
Causal Mapping	Versatility	Focus on causal relationships between concepts	In domains driven by causal relations
Card Sorting	Quick, inexpensive, easy-to-administer, and flexible	Captures "surface" characteristics Content-centric rather than task-centric	Initial phases of information architecture design
Repertory Grid Method	High validity and reliability	Prohibitive amount of time required	Only when ample research time is available

Table 14: Overview of Techniques for Mental Model Representation and Analysis.

Technique	Main Advantages	Main Limitations	Recommended Use
Multidimensional Scaling (MDS)	Relevant dimensions for users are identified. Pictorial representation of concept clusters	Not always easy to identify the appropriate variation of the technique	Coupled with pairwise ratings
Distance Ratio Formula (DR)	It can isolate 3 different type of differences	Cannot be generalized to maps of different types	Coupled with causal mapping
Pathfinder Algorithm	Applied to a variety of domains (e.g., aviation, HCIs, education)	Arbitrary layout of items in a Pathfinder network	Coupled with pairwise ratings

5.6.2. Human - human collaborative metrics

5.6.2.1. Team coordination efficiency

Team coordination is generally assessed through communication analysis. Communication analysis can be characterized through two dimensions: “physical” data vs. “content” data, and “static” vs. “sequential” analyses [23]. Physical measures are relatively low-level measures such as duration of speech, whereas content measures account for what is actually said. Static measures are metrics of team communication at one point in time, or aggregate measures over some duration, whereas sequential analyses account for the ongoing stream of team interaction.

In general, communication analysis is very time consuming, thus it requires automation in measurement and analysis. Kiekel et al. proposed to use Latent Semantic Analysis (LSA) to assess communication content, either statically or sequentially, Procedural Networks (PRONET) to address either physical or content-based sequential data, and Clustering Hypothesized Underlying Models in Sequence (CHUMS) to address sequential physical data [23].

LSA is a model of human language that can be used to code communication content and to determine similarity among utterances [87]. CHUMS is a clustering approach to determine pattern shifts in sequential data. PRONET is a sequential analysis that relies on the network modeling tool, Pathfinder [88]. PRONET can be used to determine what events “typically” follow one another, for a given lag, and to identify “typical” chains of events. For example, PRONET helps identify events that tend to co-occur in time such as, pilot begins speaking--navigator interrupts--pilot finishes speaking.

5.6.2.2. Team situation awareness

The metrics and techniques to measure team situation awareness are similar to those used to evaluate individual situation awareness. For example, it is possible to extend query methods such as

SAGAT and SPAM to the team situation. Cooke et al. have applied the SPAM method, in which the environmental cues remain present on the display, in several studies [22]. The authors typically query each individual on the team about aspects of the task environments in the present or future. Team awareness accuracy is estimated through aggregation of individual accuracy scores. In addition, the similarity or agreement of individual responses is also calculated as another metric of team situation model.

In addition to query-based metrics, metrics of team situation awareness based on communication analysis are popular in the literature. These metrics are based on the idea that team communication (particularly verbal communication) supports the knowledge building and information processing that leads to SA construction [89]. The main advantage of these measures is that they can be used to measure team situation awareness in the field, in real time, and unobtrusively.

For example, Gorman et al. applied a communication-based measure of team situation awareness in an experiment where command and control teams had to overcome a communication channel glitch [25]. Their team situation awareness metric included the number of times team members had independently noted the glitch, the number of times team members discussed the glitch, and the number of times team members coordinated actions to circumvent the glitch.

Furthermore, Bolstad et al. are developing a tool called Automated Communication Analysis of Situation Awareness (ACASA) that combines computational linguistics and machine learning techniques coupled with LSA to automatically analyze team communication and predict team situation awareness [90].

5.6.2.3. *Team mental model*

As in the case of individual mental models, the evaluation of team mental models consists of a two-step process: knowledge elicitation, and mental model analysis and metric calculation. The techniques to elicit knowledge are explained in Appendix E. One of the most popular one is to ask the users to estimate the pairwise relatedness of task-relevant concepts. It should be noted that in the case of teams, the concept pairs are selected so that it is possible to discriminate between team members with accurate and poor mental models or between team members with different task roles.

Most popular team mental model metrics are accuracy metrics, coherence metrics, heterogeneous accuracy metrics, and knowledge distribution metrics. Accuracy and coherence metrics are similar to those discussed for individual mental model evaluation; they refer to the similarity of mental models across participants and between each participant and the reality. Heterogeneous accuracy metrics are based on the idea that team members are given specific roles and that this division of labor corresponds to specific portions of the knowledge base. Cooke et al. propose to “chunk” or partition the knowledge base into units associated with each team member's role to measure heterogeneous accuracy

[91]. Chunks can also consist of the same concepts but represent a different view or perspective on that information.

Knowledge distribution metrics are based on the concept that the manner in which specific knowledge is distributed among team members can be a critical factor in team performance [92]. Coverage is a popular metric for knowledge distribution; coverage refers to how much of the knowledge is shared among the team. An example of how to measure coverage is to determine the percentage of knowledge that is redundantly distributed across team members, the percentage of knowledge that is uniquely distributed, and the percentage of knowledge that is not covered.

5.6.2.4. Workload distribution

Evaluating workload distribution among team members is required in studies where team organization, configuration, or function allocation is explored. Generally, teams are designed so that workload is balanced among their members. The metrics and techniques to evaluate workload were discussed in section 5.4.1.

In the initial phases of such studies, operator utilization can also be used. Utilization is defined as the ratio of the time the operator spends interacting with the system, or servicing events, to total mission time, and it can be estimated by using discrete-event simulation [93]. Utilization measures temporal load, and it is not a workload measure per se because the time the operator spends interacting with the system depends on many different factors. However, it can be a useful metric to detect potential overloads in the team and to explore the tasks and responsibilities distribution in a team that better balance the temporal load among the members.

5.6.2.5. Social patterns and roles

Studies that explore team organization, configuration, or function allocation should also consider the existing social patterns, roles, and informal networks within the organization. The study of social patterns and roles is important because team dynamics are often driven by team roles. In addition, designing a team structure and organization that violates the existing social patterns and roles can have a detrimental effect on performance.

A role is the typical behaviors that characterize a person in a social context [94]. It is normal that various members come to play different roles in the social structure. In general, three roles commonly emerge in groups: the task-oriented role, the socioemotional role, and the self-oriented role [95]. The roles that people adopt are often related to their personalities. An important aspect of social patterns is informal networks. Networks of, for example, trust, advice, communication, and respect exist within any organization, and their understanding can be useful to maximize system performance.

However, it is out of the scope of this research to discuss metrics and techniques to measure social patterns and roles. This metric subclass is included in the model for completeness.

5.6.3. Autonomous platform - autonomous platform collaborative metrics

The efficiency of automated collaboration among multiple platforms can be measured by the platforms' speed and reaction time to situational events without the operator's intervention. In addition, one can also measure performance metrics for the specific activities that are based on autonomous collaboration among platforms without the human intervention. These metrics will highly depend on the application and the mission.

Another alternative approach is to measure the four complexity measures for characterizing multi-robot distributed algorithms proposed by McLurkin and Kaelbling: accuracy, physical running time, communication complexity, and configuration complexity. Accuracy measures how well the robots achieve the desired physical configuration. The physical running time takes into account the robot's velocity, the speed at which messages propagate throughout the network, and the complexity of the environment. Communication complexity measures communication range, available bandwidth between neighboring robots, and messaging rate needed to adapt to changing network topology. Configuration complexity quantifies the minimum number of robots required for an algorithm, the amount of information stored in their configuration, and the algorithmic cost of storing and retrieving this information. This is still an on-going research, and the authors are working to define these complexity metrics and use them to predict the performance of a library of multi-robot algorithms¹⁰.

¹⁰ Further information can be found at <http://publications.csail.mit.edu/abstracts/abstracts07/jamesm/jamesm.html>

6. Case Studies: Past Research

6.1. An ecological perceptual aid for precision vertical landings

6.1.1. Background

The Vertical Altitude and Velocity Indicator (VAVI) is an integrated flight instrument display component intended for a heads-up display (HUD). The VAVI is designed to aid astronauts and pilots with precision vertical landing and hover operations. The VAVI conveys altitude and vertical velocity information to indicate unsafe situations and hover maneuvers in an integrated form. The display instrument takes advantage of direct-perception interaction by leveraging ecological perception and emergent features to provide quick perception and comprehension of critical flight parameters in an integrated fashion. The VAVI display is shown in Figure 11.

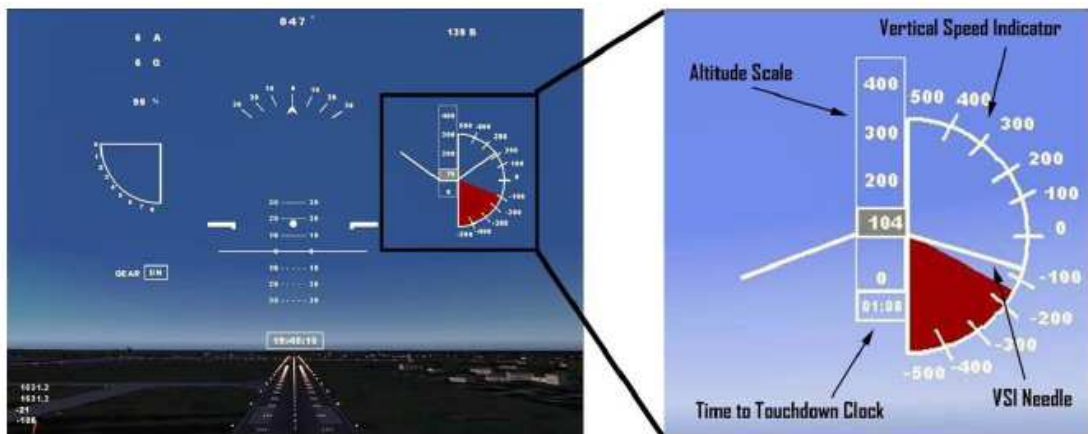


Figure 11: The Vertical Altitude and Velocity Indicator (VAVI)

To test the effectiveness of the VAVI, an experiment was conducted, in which participants flew a simulated Harrier vertical landing flight profile using Microsoft Flight Simulator (MSFS) 2004. Participants were recruited for their helicopter pilot experience or PC flight simulator experience. Two heads-up displays were implemented: one which included the VAVI, and another which displayed altitude and vertical speed information consistent with operational V/STOL aircraft head-up displays. A 2x2 ANOVA design was utilized in which the heads-up display was a between-subjects factor, and flight task, which included hovering and landing, was a within-subjects factor. Participants completed two test scenarios which involved hovering at a specified altitudes and descending using either a static or dynamic vertical speed heuristic. Further information on this research can be found in [96].

6.1.2. Metrics

The metrics used in this experiment are:

- Hover accuracy

Hover accuracy is measured as the difference between the commanded hover altitude and the actual hover altitude. The actual hover altitude, defined as the altitude that the participant tried to maintain, was determined to be the altitude at the time that the participants indicated they were beginning their hover.

- Hover precision

Hover precision addresses the ability to maintain a precise hover. Deviations from the actual hover altitude are captured using a root mean square error (RMSE). In the measure of hover precision, the desired variable corresponds to the altitude at the first second of the self-initiated 20 second hover and actual corresponds to the altitude at every second during that 20 second interval.

- Vertical Speed Precision

The ability to maintain a static descent rate was also measured using a RMSE. The RMSE of vertical speed from the completion of the hover to landing was calculated. Participants were not penalized for being within 10% of the commanded static descent rate, and participants were also not penalized for having a positive (> 0 fpm) vertical speed indicating that they were climbing and not descending.

- Descent Duration Error

The descent duration is a measure of the comparison between the time that it should have taken the participant to descend, had they descended according to the commanded dynamic or static vertical speed heuristic, versus the actual time of descent.

- Workload Measures

Participants rated their perceived mental workload on a ten-point scale, with 1 corresponding to minimal or no mental workload, and 10 corresponding to the highest mental workload the participant has experienced.

Table 15 summarizes the actual metrics and the corresponding metric classes used in this experiment.

Table 15: Metric classes and actual metrics used to evaluate the VAVI

Metric Class	Metric Subclass	Metric
Mission Effectiveness	N/A	Hover accuracy Hover precision Vertical speed precision
Autonomous Platform Behavior Efficiency	N/A	N/A (manual control, research question about data visualization)
Human Behavior Efficiency	Information Processing Efficiency (task efficiency)	Descent duration error
Human Behavior Precursors	Workload	Workload self-rating
Collaborative Metrics	N/A	None

6.1.3. Results

Table 16 summarizes the results obtained for each metric.

Table 16: Results of the VAVI evaluation

Metric Class	Metric	Results
Mission Effectiveness	Hover accuracy	No significant difference
	Hover precision	No significant difference (marginal significance for expert users)
	Vertical speed precision	Significantly lower with the VAVI
Human Behavior Efficiency	Descent duration error	No significant difference
Human Behavior Precursors	Workload self-rating	Marginal significance, lower workload with the VAVI (significant difference for experts)

In addition to the global analysis, the subgroup of best performers was separately studied. Hover precision showed marginally significant improvement with use of the VAVI, while workload results strongly indicated a lower mental workload associated with the VAVI. These results indicate improved hover performance and reduced perceived workload with use of the VAVI when used by an expert subset of participants.

6.1.4. Discussion and conclusions

Mission effectiveness and information processing efficiency metrics were important in this experiment. However, the experimenters also measured workload in order to better understand the VAVI effect and the reasons behind subjects' behavior. This latter metric proved to be essential.

In this experiment, the metric subclasses of attention allocation efficiency, situation awareness, or mental model efficiency were not considered, but they could have provided additional insight. Attention allocation efficiency, and, in particular, subjects' visual behavior, can help in understanding how a particular display directs and affects users' attention allocation strategy. Also, it can be interesting to measure the effect of the display on situation awareness, and to use retrospective think-aloud to capture subjects' cognitive strategies and mental models to understand how the display is used. But there is the risk that the benefits obtained from measuring these additional metrics do not justify their cost. A rigorous cost-benefit analysis is required to determine which of these metrics provide enough added value for this particular experimental setting and constraints. This type of analysis will be conducted in the second phase of this research effort.

Finally, the experimenters in this example concluded that the hover performance was a difficult metric to capture because the 20 seconds during which the participants indicated to the experimenter that

they were hovering, did not necessarily capture the participants' best hover performance throughout the cycle. It is important to carefully select performance metrics so that they capture the relevant performance aspects. For example, experimenters should think about whether one wants to capture the user performance during a given period of time or the user performance when a given event or behavior happens. In conclusion, prior to define performance metrics, experimenters should clearly identify and define the relevant mission parameters, behaviors, and success criteria.

6.2. Decision support for lunar and planetary exploration

6.2.1. Background

This research effort studied decision support for planetary-surface traversals. Different automation levels and visualization interfaces were evaluated. Figure 12 shows the interface, with which the participants were able to make, modify and submit the least-costly paths they had planned.

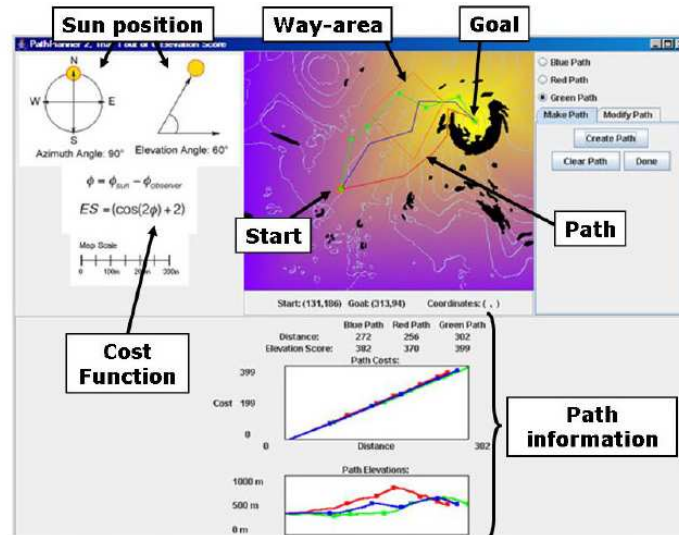


Figure 12: PATH Interface

The PATH Interface is a decision support tool to plan and optimize paths based on objective functions important to planetary-surface traversals. In particular, this research studied visualization of equal cost contours (LOEC) showing areas on the map that had equal cost for the objective function. These LOEC were represented with colored gradients as shown in Figure 12. In addition, this research effort also explored displaying terrain-elevation contours. PATH Interface displayed elevation contours in two ways: grayscale-filled contour map and contour lines overlaid on top of levels of equal cost visualization. For the grayscale map, shown in Figure 13, white was the highest elevation while dark gray was the lowest elevation, black remains obstacles.

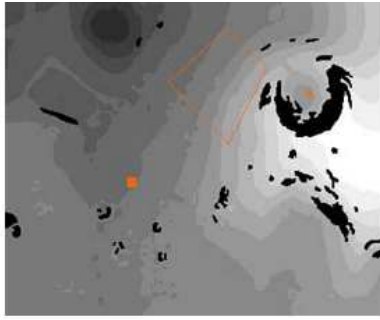


Figure 13: Grayscale-filled contour map

Human-in-the-loop testing was employed to understand the effects of the automated assistance and different visualizations on path planning performance across multivariate cost functions. In two separate experiments, participants were tasked to make obstacle-free, least-costly paths based on given cost functions. In the first experiment, three independent variables were tested: type of visualization (3 types: elevation contours, levels of equal cost (LOEC), and combination of elevation contours and LOEC); level of automation (2 levels: passive and active); cost function (2 functions, Elevation Score, and Sun Score). While visualization type was a between-subject variable, cost functions and automation type were within-subject variables, resulting in a 2 x 2 x 3 repeated measures design.

In the second experiment three independent variables were tested, but under two experimental matrices. Within one matrix, the variables were type of visualization (the same 3 as those of experiment 1) and cost function (4 increasingly complex cost functions: Distance, Time, Metabolic, and Exploration). Within the second matrix, the variables were type of visualization (the same 3 types as those of experiment 1), cost function (Time and Exploration), and type of scenario (nominal and off-nominal). The off-nominal scenario represented a degraded automation condition. In the nominal scenario, participants were told they could rely on the PATH interface to provide them with accurate path cost based on the cost functions. In the off-nominal scenario, participants were informed that PATH's cost function models were inaccurate. Since only a subset of the nominal cost functions were tested in the off-nominal case, two experimental matrices were required for the analysis and statistical models. While the visualization type was a between-subjects variable, the cost function and the scenario type were within-subjects variables. Thus, there was a 3 x 4 repeated measures design for the first experimental matrix, and a 2 x 2 x 3 repeated measures design for the second matrix. Further information on this research can be found in [123].

6.2.2. Metrics

The metrics used in this experiment are:

- Path Cost Error

Path planning performance was measured by path cost errors that were calculated by comparing the path cost generated by the participant to the automation's minimum path cost.

- Total Time

The total time to complete trial was a main performance metric in this experiment.

In the second experiment, the total time to complete task, was converted to time penalty. Participants were asked to complete the task as well and as fast as possible. The time pressure was imposed by showing the participant a timer, an incremental clock. The penalty time started after 4 minutes, which was not told to the participants.

- Percent Time Spent Modifying Path

The percent of time spent modifying a path was the time spent by subjects, after having created an initial path, on modifying this path by moving, adding, and deleting waypoints.

- Path Cost Profiles: true time, differential cost, and non-optimal satisficing

In the second experiment, the path cost profiles were also analyzed. In particular, the measures used were: true time, differential cost, and non-optimal satisficing.

True time is how long a participant took to arrive to the minimum path cost found. It differs from total time or time penalty because it reflects the actual time that it took a participant to optimize a path, excluding any time spent afterwards attempting to find another solution. Short true times indicate that participants were able to optimize the path quickly.

Differential cost is the path cost error difference (in percent) between the first path cost the participant made for a particular trial and the submitted path cost. This metric helps to assess how much path cost error decreased during the optimization process. A negative differential cost would indicate that the participant submitted a path that was worse than the first one they had made. A small differential cost would indicate that the participant's first path attempt was close in cost to the submitted least-costly path.

Non-optimal satisficing refers to actions taken by a participant attempting to find a lower path cost after a minimum was already achieved. Specifically, non-optimal satisficing is defined as 1) cost surplus: the difference between the minimum path cost achieved and the submitted path cost, and 2) time surplus: the percent of time spent between those ($[(\text{total time} - \text{true time})/\text{total time}]$).

- Situation Awareness

Situation awareness was measured through the number of situation awareness (SA) questions that were answered correctly. Multiple choice questions were used as a global measure of the participants SA. After every trial, participants were asked two questions about the previous trial. Specifically, participants were asked about the elements in the display (e.g., sun positions), and the cost functions (e.g., how path costs would be affected by changes in variables). There were a total of eight multiple-choice questions asked, four per automation level.

In the second experiment, performance metrics in off-nominal conditions can be considered as performance measures of situation awareness. If subjects could not rely on automation, they had to rely

on their own knowledge and understanding of the environment, thus subjects' performance in off-nominal mode was considered as an implicit measure of their situation awareness.

Table 17 and 18 summarize the metrics used in the first and the second experiment.

Table 17: Metrics used to evaluate the PATH Interface in the first experiment

Metric Class	Metric Subclass	Metric
Mission Effectiveness	N/A	Path cost error Total time
Autonomous Platform Behavior Efficiency	N/A	None
Human Behavior Efficiency	Information Processing Efficiency (task efficiency)	Percent time spent modifying path
Human Behavior Precursors	Situation Awareness	Post-experiment Questionnaire
Collaborative Metrics	N/A	None

Table 18: Metrics used to evaluate the PATH Interface in the second experiment

Metric Class	Metric Subclass	Metric
Mission Effectiveness	N/A	Path cost error Total time
Autonomous Platform Behavior Efficiency	N/A	None
Human Behavior Efficiency	Information Processing Efficiency (task efficiency)	Percent time spent modifying path True time Differential cost Non-optimal satisficing
Human Behavior Precursors	Situation Awareness	Performance metrics in off-nominal conditions
Collaborative Metrics	N/A	None

6.2.3. Results

Table 19 and Table 20 summarize the results obtained for each metric.

Table 19: Results of the first evaluation of the PATH Interface

Metric Class	Metric	Result
Mission Effectiveness	Path cost error	Significant reduction for active automation
	Total time	Significant reduction (by about 1.3 minutes) when using active automation
Human Behavior Efficiency	Percent time spent modifying path	Significant difference between automation levels. With passive automation, participants spent a large portion of time modifying paths
Human Behavior Cognitive Precursors	Situation awareness	Significant reduction for active automation

Table 20: Results of the second evaluation of the PATH Interface

Metric Class	Metric Subclass	Metric
Mission Effectiveness	Path cost error	Significant differences among cost functions
	Total time	No significant differences, but the LOEC group spent fewer time to optimize the Time cost function (the one with the longest times)
Human Behavior Efficiency	Percent time spent modifying path	No main effect between visualization groups and cost functions. The trend is that participants that had the LOEC visualization tended to spend longer time modifying paths
	True time	Significant difference between cost functions
	Differential cost	In the cost function with the smallest differential cost, Exploration, there is a marginal main effect of visualization. Participants with just the LOEC visualization had a significantly smaller differential cost than the participants in the elevation contours & LOEC visualization group.
	Non-optimal satisficing	Participants did not submit sub-optimal path costs relative to the achieved minimum. But, they spent on average between 20 – 35% of their time conducting non-optimal satisficing. With the Time cost function, participants had a significantly lower percent time surplus than with Metabolic and Exploration functions.
Human Behavior Cognitive Precursors	Situation awareness	In the off-nominal cases, the LOEC participants emerge as being the better performers.

In conclusion, the results of the first experiment showed that the effect of the level of automation was strong and consistent across all dependent variables. The second experiment focused on testing only passive automation in order to further examine visualization effect. Results showed that, for the Exploration cost function, the most complex one because of the highest number of variables, the levels of equal cost (LOEC) visualization helped participants initially make paths that were close to their optimal.

The best performers under the nominal conditions were mostly in the elevation-contour group. However, in the off-nominal cases, the LOEC participants emerge as being better performers. The reasons behind this may be that while visualization did not have a main effect on performance, it did influence the choice of subjects' strategies. Participants with the additional LOEC visualization tended to spend more time modifying the Time function, which assisted them during degraded automation conditions. The levels of equal cost visualization, which aggregates all variables into one cost map, helped reduce the complex problem, in terms of providing an efficient optimizing strategy, and promoted sensitivity analysis for difficult problems.

For the task of path optimization, humans perform best when they leverage sensitivity analysis. As the presence of the LOEC visualization promoted sensitivity analysis, the conclusion of this research was that visualization is a desirable attribute in decision support aids during the optimization of paths.

6.2.4. Discussion and conclusions

In the first experiment, only one human behavior efficiency metric was considered, whereas in the second experiment three more metrics from this class were measured, enabling a more comprehensive understanding of the experiment. These additional metrics complement the information provided by the original ones.

However, it is an open question whether the three additional metrics were necessary, or the same results would have been obtained adding only one or two metrics. Correlation among these metrics should be further analyzed to understand the relationships among them. Based on the results obtained, non-optimal satisficing turned to be an important metric, while true time and differential cost revealed less about subjects' behaviors. A rigorous cost-benefit analysis would be needed to answer whether or not, non-optimal satisficing and percent time spent modifying path would have been sufficient metrics for this study.

It should be noted that the percent time spent modifying paths only describes a method of conducting sensitivity analysis. This method was the most frequent one, the strategy chosen by 26 participants. However, there is another type of sensitivity analysis, creating multiple paths, that was conducted by the remaining 8 participant. It is important that performance metrics capture all possible behaviors and strategies, not only the most popular or straight-forward ones.

Regarding performance metrics, there were no significant correlations between errors and time measures. This indicates that both type of metrics can reflect different effects and provide complementary information. However, some of the time metrics were correlated. For example, there was a significant correlation within the Time function between non-optimal satisficing time surplus and true time (Pearson correlation = -0.48, $p = 0.005$). This might indicate that after spending a long time attempting to solve the Time function path, participants did not spend additional time conducting non-optimal satisficing once a minimum path was found. Correlation between metrics should be carefully studied since it can indicate that both metrics refer to the same phenomenon.

No metrics from the attention allocation efficiency were included in this experiment. Analyzing interactions between the participant and the PATH interface (mouse clicks) per interface element (i.e., a particular button, or area of interest) could potentially provide further insight into the individual elements of the interface. Regarding human behavior precursors, only situation awareness was measured. However, workload measures might have provided additional insights on the complexity of each cost function, and on how certain visualizations support better decision making and partially mitigate task complexity. Finally, collaborative metrics such as trust or mental model efficiency would have helped understanding whether the subjects understood and trusted the decision-support tool.

6.3. Assessing the impact of auditory peripheral displays for UAV displays

6.3.1. Background

The focus of this experiment was audio cues in unmanned aerial vehicle (UAV) interfaces. The objective was to determine whether sonifications maximize the information conveyed to UAV operators more efficiently than typical discrete alarms used in current ground control stations. In addition, the impact of continuous versus discrete alerting on operators was also explored.

Participants had to control the UAVs, ensuring that they did not deviate from their paths and they did not arrive late to their targets. The test bed used for the experiment is the Multiple Autonomous Unmanned Vehicle Experimental (MAUVE), which has been developed at HAL.

The experiment was a 4x2 fixed factor repeated measures model, with two independent variables: the audio condition (a between-subjects treatment), and the number of vehicles under control (a repeated within-subjects factor). The audio conditions were: 1) a threshold audio condition¹¹ for both the late arrivals and course deviations, 2) a continuous oscillating course deviation audio condition¹² with threshold alerts for the late arrivals, 3) a continuous modulated late arrival audio condition¹³ with a threshold alert for course deviations, and 4) an oscillating course deviation alert and a modulated late arrival alert.

The second independent variable, the number of vehicles under control, had two levels: single UAV and multiple UAV. In the single level, the participant supervised only one UAV, while in the multiple factor level, the participant supervised four UAVs. Further information on this research can be found in [124].

¹¹ The threshold course deviation alert consisted of a single beep.

¹² The oscillating alert consisted of comb filters that were applied to a mix of pink noise and the ambient signal.

¹³ The modulated alert consisted of discrete harmonic signals continuously playing.

6.3.2. Metrics

The metrics that were used in this experiment are:

- Course Deviation Reaction Time

The course deviation reaction time indicates how quickly the participant responded to the audio cue that warned him about a UAV being deviated from its path. The scenarios included four triggered course deviations that require the participant to respond.

- Course Deviation Errors

Course deviation errors of omission are the number of times the participant failed to respond to one of the four triggered course deviations.

- Late Arrival Reaction Time

Late arrival reaction time indicates how quickly the participant responded to the audio cue that warned him about a UAV being late to its target. Four late arrivals were present in each test scenario and were caused when a UAV slowed down because of headwinds.

- Late Arrival Errors

Late arrival errors of omission are the number of times the participant failed to respond to one of the four triggered late arrivals.

- Workload Metrics

Secondary-Task performance workload measure: The number of radio calls missed was measured as an indication of the operator's level of mental workload. The secondary task consisted of participants monitoring air traffic radio communications and acknowledging the word "Push" by clicking "Acknowledge Push" button on the display.

NASA Task Load index: The NASA TLX, a subjective workload measure, was administered after each condition.

Table 21 summarizes the metrics classes and actual metrics used in this experiment.

Table 21: Metrics used in the Auditory Peripheral Displays study

Metric Class	Metric Subclass	Metric
Mission Effectiveness	N/A	Course deviation errors Late arrival errors
Autonomous Platform Behavior Efficiency	N/A	None
Human Behavior Efficiency	Information Processing Efficiency	Course deviation reaction time Late arrival reaction time
Human Cognitive Precursors	Workload	Secondary task performance NASA TLX
Collaborative Metrics	N/A	None

6.3.3. Results

Table 22 summarizes the results obtained for each metric.

Table 22: Results of the Auditory Peripheral Displays study

Metric Class	Metric	Result
Mission Effectiveness	Course deviation errors	No significant differences
	Late arrival errors	No errors
Human Behavior Efficiency	Course deviation reaction time	Oscillating course deviations alerts promoted the best performance
	Late arrival reaction time	Modulated late arrival alerts promoted the best performance
Human Behavior Cognitive Precursors	Secondary task performance	No significant differences
	NASA TLX	No significant differences

In conclusion, the experiment showed that sonifications for two different events, the combination audio scheme, promoted the best performance for the two reaction time dependent variables. This combination scheme consisted of oscillating course deviations alerts and modulated late arrival alerts. The workload equality was not a surprise, because the two scenarios in this experiment were designed to have comparable workloads, with the multi UAV scenario dividing the monitoring tasks over four vehicles instead of just one. The subjective ratings, the NASA TLX confirmed that the participants felt the two scenarios were comparable in workload.

6.3.4. Discussion and conclusions

In this experiment, selected metrics are directly related to the mission (i.e., monitoring course deviations and late arrivals) and reflect the most important aspects of controlling a UAV. The omission of course deviations and late arrivals can often prove disastrous and affect the health of the UAV.

The main problem with error metrics is that often participants do not commit sufficient errors to conduct a statistical analysis. As shown by this case study, error metrics are required, but the experimenter should not rely on them and should measure other metrics, such as reaction times. In addition, it is important to design the experiment with the appropriate complexity level so that participants do not commit too few or too many errors.

Reaction time, a metric from the human behavior efficiency class, was also measured because lower reaction times affect control efficiency even if late responses do not automatically imply harmful results. Furthermore, the goal of alarms is to attract users' attention so that they respond quickly to potential threads. Thus, reaction times are a direct measure of the efficiency of alarms. It should be noted

that reaction times to different events are often correlated, reflecting related effects that should not be reported independently.

Regarding attention allocation efficiency, the experimenters could have included a post-experiment survey or use a verbal protocol to better understand how the different alarms affected subjects' attention allocation. Collaborative metrics were not included either; however, in this experiment, the reliability of the auditory alarms was not simulated, thus measuring trust was not applicable. In addition, auditory alarms are simple concepts that do not require evaluating any related mental model.

Finally, workload was measured through two different metrics. The same conclusions were derived from each metric. This indicates that there was no added value on measuring two different workload metrics for this experiment.

7. Case Studies: On-going Research

7.1. Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU)

7.1.1. Background

This research effort will conduct two experiments in which subjects have to supervisory control heterogeneous unmanned vehicles. The first experiment will compare remote testing through internet to controlled testing environments with the physical presence of the experimenter. The second experiment will validate a human-supervisory control model of unmanned vehicle operators that predicts system performance.

Subjects' task in both experiments is to conduct a surveillance and reconnaissance mission, navigating the unmanned vehicles, operating video cameras, and recognizing targets.

7.1.2. Metrics

The metrics that will be used in this experiment are:

- Number of targets correctly identified by operator.
- Operator utilization, defined as the time the user is interacting with the interface divided by total mission time.
- Vehicle wait, defined as the time a vehicle has to wait from the moment it needs operator's attention until it receives it from the operator.
- Response time to threat areas, defined as the amount of time it takes for the operator to respond to threat areas that might be in the paths of vehicles. This is considered an indication of situation awareness of the operator.

Table 23 summarizes the metrics classes and actual metrics to be used in these experiments.

Table 23: Metrics to be used in the RESCHU research study

Metric Class	Metric Subclass	Selected Metric
Mission Effectiveness	N/A	Number of targets correctly identified
Autonomous Platform Behavior Efficiency	N/A	None
Human Behavior Efficiency	Information Processing Efficiency	Operator utilization
	Attention Allocation Strategy	Vehicle wait
Human Behavior Precursors	Situation Awareness	Response time to threat areas Awareness self-rating
	Workload	Workload self-rating
Collaborative Metrics	N/A	None

7.1.3. Discussion and conclusions

The metrics selection in this experiment is driven by the model that wants to be validated. However, in a more generic experiment, additional metrics from human information processing efficiency could help in understanding the individual components of the mission and provide additional insight. Some examples of this type of metrics are: time required to identify a target, number of times that the shortest goal is assigned to an unmanned vehicle, time spent modifying unmanned vehicles' paths, number of objects incorrectly identified. Regarding attention allocation efficiency, since the operator has to control several vehicles, metrics such as switching times and homogeneity of attention allocation among vehicles could also be measured.

Regarding human behavior precursors, response time to threat areas will be measured as an indication of situation awareness. This performance metric can be very helpful, however, it evaluates a very particular aspect of situation awareness and it is sensitive not only to variation in SA, but also to variations in other factors, such as workload, skill level, or operators' strategies. Additional metrics that would complement this response time are, for example, the number of times that the shortest goal is assigned to an unmanned vehicle, or the number of times a target is overpass without identifying it while operating the video camera. Subjective metrics of situation awareness are easy to administer and can provide interesting results. However, humans are very often unaware of their lack of awareness. The on-line query methods and the testable responses are also recommended techniques to measure situation awareness, but they might interfere with primary task performance and, therefore, confound the validation of the model of interest.

8. Conclusions from the Case Studies

The lessons learned and conclusions that can be extracted from these case studies are:

- Mission Effectiveness, Human Behavior Efficiency, and Human Behavior Precursor metric classes are the most popular ones. Including a metric from each of these classes seems to be the minimum recommended set of metrics. This hypothesis and the added value of incorporating additional metrics for a given experiment should be evaluated with a cost-benefit analysis. Such methodology will be developed in the next stage of this research effort.
- Collaborative Metric and Autonomous Platform Behavior Efficiency classes were not included in any of the experiments analyzed. The added value of these metric classes is still to be evaluated since there was no case study to discuss it.
- Attention allocation efficiency metrics were only included in the on-going experiment, from which there are no available results yet. Metrics to capture users' strategies can add value to an experiment. However, it is still to be demonstrated the actual value of measuring the interactions of participants with an interface (e.g., mouse clicks), administering post-experiment surveys, and conducting verbal protocols to better understand how the different elements of an interface affect subjects' attention allocation and strategies.
- Performance metrics should be chosen so that they capture the most relevant performance aspects of a mission. Prior to defining performance metrics, experimenters should carefully think about the mission parameters and behaviors of interest that need to be captured.
- Performance metrics should capture all possible behaviors and strategies, not only the most popular or straight-forward ones.
- Performance metrics should be analyzed taking into account operators' state or initial conditions prior to a given response or behavior.
- Errors and time metrics tend to be uncorrelated, indicating that these metrics can reflect different effects and provide complementary information.
- Some time metrics tend to be correlated, for example, reaction times to different events. These effects should not be reported independently.
- Participants tend to commit insufficient errors to conduct a statistical analysis on this type of metrics. It is recommended to use other metrics in addition to error-based metrics. For example, in research studies related to the effectiveness of alarms, reaction times should also be measured.
- Experiments should be designed with the appropriate complexity level so that participants do not commit too few or too many errors if error-based metrics are measured.
- The use of two different metrics to measure the same aspect of the system is not always justified. Most of the time, the same conclusions are derived from both set of data resulting in wasted resources and an inflation of the type I error.

9. Preliminary Evaluation Criteria for Supervisory Control Metrics

As discussed in this report, there are multiple metrics and measuring techniques that can be used to evaluate supervisory-control applications. All of these metrics and techniques have advantages and limitations, and the actual set of metrics that provide the most value for a given experiment will depend on the context and the application. Based on the results and conclusions presented in this report, we have identified a preliminary list of evaluation criteria for supervisory-control metrics. This list will be refined in the next stage of the project, and it will be used as the basis to develop a cost-benefit methodology to select supervisory control metrics. The preliminary evaluation criteria that have been identified are:

Experimental Constraints:

Time: How much time is required for the measurement and analysis of this metric?

Cost: How costly is the metric to measure and analyze? There will also be a cost for time.

Experimental setting: Which experimental setting is this metric appropriate for? (e.g., field testing vs. simulation environment)

Development phase: How far in the system development does this metric require us to be?

Construct validity: How well is each metric measuring what I want to measure?

Power to discriminate between similar constructs: How well does the metric discriminate between abstract constructs that are hard to measure such as workload, or attentiveness (e.g., Does galvanic skin response really measure workload, or does it measure stress?).

Intra-subject reliability: Does the metric assess the same construct for every subject? Subjective responses have lower intra-subject reliability.

Comprehensive understanding gained: How much does this set of metrics explain the whole phenomenon?

Proximity to the primary research question: How much does this metric help me answer what I ultimately want to learn? For example, a workload metric may not tell much without a mission effectiveness metric.

Amount of additional understanding gained: Given that I have other metrics, how much more do I learn from this metric?

Causal relations with other metrics: How well does this metric help explain underlying reasons for other metrics collected?

Statistical validity/efficiency: How much does this set o metrics support the statistical analysis?

Correlation with other metrics: How correlated is this metric with others we are collecting? Need to be cautious when collecting more than one metric in a specific metric class. We may measure the same phenomenon, which will result in inflated type I errors, and also wasted resources.

Effect size: Based on my research questions, am I expecting a good separation between different conditions for this metric?

Frequency of observations: Will this metric result in enough observations to enable us extract meaningful information? This will be driven by time and cost related to collecting observations.

Measuring technique: How appropriate is this measuring technique?

Non-intrusiveness: How intrusive is the equipment needed to collect this metric?

Realism: How much does the measuring technique interfere with the nature of the task?

References

- [1] Jean Scholtz, Jeff Young, Jill L. Drury, and Holly A. Yanco. Evaluation of Human-Robot Interaction Awareness in Search and Rescue. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), New Orleans, April 2004.
- [2] Olsen, R., O. and Goodrich, M.A. 2003. Metrics for evaluating human-robot interactions. In Proc. NIST Performance Metrics for Intelligent Systems Workshop.
- [3] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz and M. Goodrich. 2006. Common Metrics for Human-Robot Interaction. In Proceedings of the 1st Annual IEEE/ACM Conference on Human Robot Interaction (Salt Lake City, Utah, USA, March 2 - 3, 2006). HRI'06. ACM Press, New York, NY.
- [4] Sheridan, T.B. 2002. Humans and Automation: System Design and Research Issues. A John Wiley & Sons, Inc., Publication. New York, NY.
- [5] 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.
- [6] Sheridan T.B. 1992. Telerobotics, Automation, and Human Supervisory Control. The MIT Press. Cambridge, MA.
- [7] Autonomous Vehicles in Support of Naval Operations by Committee on Autonomous Vehicles in Support of Naval Operations (Author), National Research Council (Author). Publisher: National Academies Press (August 5, 2005).
- [8] Donmez, B., Ng Boyle, L.; Lee, J.D. 2006. The impact of distraction mitigation strategies on driving performance. Human Factors 2006, vol.48, n. 4, pp. 785-804.
- [9] Nielsen, J. 1993. Usability engineering. Academic Press / AP Professional, Cambridge, MA.
- [10] J. L. Drury, J. Scholtz and H. A. Yanco. Applying CSCW and HCI techniques to human-robot interaction. Proceedings of the CHI 2004 Workshop on Shaping Human-Robot Interaction, April 2004, Vienna, pp. 13-16.
- [11] Wickens, C.D. and Hollands, J.G. (1992). Engineering psychology and human performance. Third Edition. New York: HarperCollins.
- [12] 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.
- [13] Klein, G., & Klingner, D. (2000). Naturalistic Decision Making. Human Systems IAC GATEWAY, 11 (3), 16-19.
- [14] Durso, F.T., Rawson, K.A., Giroto, S.(2007). Comprehension and Situation Awareness. Handbook of Applied Cognition. Second Edition. Edited by Francis Durso.
- [15] Endsley, M.R. and Garland D.J. (Eds.) (2000) Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum Associates.
- [16] 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.
- [17] Rouse, W.B. and Morris N.M. (1986). On looking into the black box: Prospects and limits in the search for mental models. Psychological Bulletin, 100, 349-363.
- [18] Norman, D.A. (2002). The design of everyday things. New York: Basic Books.
- [19] Lee, John D.; See, Katrina A. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, Volume 46, Number 1, Spring 2004 , pp. 50-80(31).

- [20] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.
- [21] Lewis, M., Wang, J., and Hughes, S. (2007). USARSim: Simulation for the Study of Human-Robot Interaction, *Journal of Cognitive Engineering and Decision Making*, (1)1, 98-120.
- [22] 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).
- [23] Kiekel, P. A., Cooke, N.J., Foltz, P.W., Gorman, J. C., & Martin, M.J. (2002). Some promising results of communication-based automatic measures of team cognition. *Proceedings of the Human Factors and Ergonomics Society's Annual Meeting, USA*, 46, 298-302.
- [24] 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.
- [25] 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*.
- [26] Hilbert D M, Redmiles D F. Extracting usability information from user interface events *ACM Computing Surveys (CSUR)*, December 2000: Volume 32 Issue 4, pp. 384-421.
- [27] Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: *The ASQ. SIGCHI Bulletin*, 23, 1, 78-81.
- [28] Brooke, J. (1996) SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & A. L. McClelland (eds.) *Usability Evaluation in Industry*. London: Taylor and Francis.
- [29] Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces. In *proceedings of ACM CHI'90 Conference on Human Factors in Computing System (Seattle, WA, 1-5 April)*, 249-256.
- [30] Lewis, C., Polson, P., Wharton, C., and Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *proceedings ACM CHI'90 Conference on Human Factors in Computing System. (Seattle, WA, April 1-5)*, 235-242.
- [31] Beyer, H. Holtzblatt, K. (1998) *Contextual Design: Defining Customer-Centered Systems*, Academic Press, San Diego CA.
- [32] Graham, H.D., *Developing General and Specific Modified Cooper Harper Scales for Assessing Unmanned Vehicle Displays, (HAL2008-03)*, MIT Humans and Automation Laboratory, Cambridge, MA.(2008)
- [33] M. A. Goodrich and D. R. Olsen, Jr. Seven Principles of Efficient Interaction. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, October 5-8, 2003, pp. 3943-3948.
- [34] De Visser, E.;Parasuraman, R.;Freedy, A.;Freedy, E.;Weltman, G. 2006. A Comprehensive Methodology for Assessing Human-Robot Team Performance for Use in Training and Simulation. *Human Factors and Ergonomics Society 50th Annual Meeting Proceedings, Training* , pp. 2639-2643(5).
- [35] T.W. Fong, C. Thorpe, and C. Baur. 2003. Robot, Asker of Questions. *Robotics and Autonomous Systems*, 42 (2003), 235-243.
- [36] Wickens, C. D., Helleberg, J., Goh, J., Xu, X., and Horrey, W. J. (2001). Pilot task management: Testing an attentional expected value model of visual scanning (Technical Report ARL-01-14/NASA-01-7). Moffett Field, CA: NASA Ames Research Center
- [37] Wickens, C. D., Dixon, S. R., Goh, J., Hammer, B. (2005) Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. *Proceedings of the 13th International Symposium on Aviation Psychology*.

- [38] Talleur, D. A., Wickens, C. D. (2003) The effect of pilot visual scanning strategies on traffic detection accuracy and aircraft control. Proceedings of the 12th International Symposium on Aviation Psychology.
- [39] J. W. Crandall and M. L. Cummings. Attention Allocation Efficiency in Human-UV Teams. AIAA Infotech@Aerospace Conference, 2007.
- [40] Janzen, M. E., and Vicente, K. J., "Attention allocation within the abstraction hierarchy," *International Journal of Human-Computer Studies*, vol. 48, pp. 521-545, 1998.
- [41] Y. Lin, W.J. Zhang, and R.J. Koubek. Effective attention allocation behavior and its measurement: a preliminary study. *Interacting with Computers*, Volume 16, Issue 6, December 2004, Pages 1195-1210.
- [42] Bruni, S., Marquez, J., Brzezinski, A., & Cummings, M.L., "Visualizing Operators' Cognitive Strategies In Multivariate Optimization", Proceedings of HFES 2006: 50th Annual Meeting of the Human Factors and Ergonomic Society, San Francisco, CA, USA, October 16-20, 2006.
- [43] V. J. Gawron. 2000. Human Performance Measures Handbook. LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS 2000 Mahwah, New Jersey London
- [44] D. R. Olsen and S. B. Wood, "Fan-out: Measuring human control of multiple robots," in Proc. SIGCHI Conf. Human Factors Comput. Syst., Vienna, Austria, Apr. 2004, pp. 231–238.
- [45] J. W. Crandall, M. A. Goodrich, D. R. O. Jr., and C. W. Nielsen, "Validating human–robot interaction schemes in multitasking environments," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 35, no. 4, pp. 438–449, Jul. 2005.
- [46] Cummings, M.L., and P.J. Mitchell. In press. Predicting controller capacity in remote supervision of multiple unmanned vehicles. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*.
- [47] Ogden, G.D., Levine, J.M., and Eisner, E.J. (1979). Measurement of workload by secondary tasks. *Human Factors*, 21: 529-548.
- [48] Cummings, M.L. and S. Guerlain, "Using a Chat Interface as an Embedded Secondary Tasking Tool," 2nd Annual Human Performance, Situation Awareness, and Automation conference, March 2004.
- [49] O'Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance: Vol. II. Cognitive processes and performance* (pp. 42-1 – 42-49). New York: Wiley Interscience.
- [50] Iqbal, S.T., P.D. Adamczyk, X.S. Zheng and B.P. Bailey. Towards an Index of Opportunity: Understanding Changes in Mental Workload During Task Execution. Proceedings of the ACM Conference on Human Factors in Computing Systems, Portland, Oregon, USA, 2005, pp. 311-320.
- [51] Parasuraman, R., Hancock, P.A. 2001. Adaptive control of workload. In P.A. Hancock & P. Desmond (Eds.), *Stress, Workload, and Fatigue*. Mahwah, NH: Erlbaum.
- [52] Ahlstrom, U., Friedman-Berg, F. 2005. Subjective Workload Ratings and Eye Movement Activity Measures. Technical Report DOT/FAA/ACT-05/32. U.S. Department of Transportation, Federal Aviation Administration.
- [53] Tattersall, A. J.;Hockey, G. R. J. 1995. Level of Operator Control and Changes in Heart Rate Variability during Simulated Flight Maintenance. *Human Factors*, Volume 37, Number 4, pp. 682-698.
- [54] Berka, C, Levendowski, DJ, Cvetinovic, M, Petrovic, MM, Davis, GF, Lumicao, MN, Popovic, MV, Zivkovic, VT, Olmstead, RE, Westbrook, P. Real-Time Analysis of EEG Indices of Alertness, Cognition and Memory Acquired with a Wireless EEG Headset. Special Issue of the *International Journal of Human-Computer Interaction on Augmented Cognition* 2004; 17(2): 151-170.
- [55] Brookhuis, K.A., & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics*, 36, 1099-1110.
- [56] Brookings JB, Wilson GF, Swain CR. Psychophysiological responses to changes in workload during simulated air-traffic control. *Biol Psychol* 1996; 42: 361-77.

- [57] Levin S, France DJ, Hemphill R, Jones I, Chen KY, Rickard D, Makowski R, Aronsky D. 2006. Tracking workload in the emergency department. *Human Factors*. 2006 Fall;48(3):526-39.
- [58] Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J. M., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6(1), 1-20.
- [59] Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., Ohrt, D., & Manning, C. A. (1995). Expertise and chess: A pilot study comparing situation awareness methodologies. In D.J. Garland & M.R. Endsley (Eds.), *Experimental analysis and measurement of situation awareness* (pp. 295-303). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
- [60] Smolensky, M. W., 1993. Toward the physiological measurement of situation awareness: The case for eye movement measurements. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Santa Monica, Human Factors and Ergonomics Society.
- [61] Wickens, C.D. 2000. "The tradeoff of design for routine and unexpected performance: Implications of situation awareness" In *Situation Awareness Analysis and Measurement*. eds. Endsley, M. R. and Garland D. J. Mahwah, NJ: Lawrence Erlbaum Associates.
- [62] Amy R. Pritchett and R. John Hansman. 2000. "Use of testable responses for performance-based measurement of situation awareness" In *Situation Awareness Analysis and Measurement*. eds. Endsley, M. R. and Garland D. J. Mahwah, NJ: Lawrence Erlbaum Associates.
- [63] A.H. Midkiff & R.J. Hansman, "Identification of Important 'Party Line' Information Elements and Implications for Situational Awareness in the Datalink Environment", *Air Traffic Control Quarter Iv*. Vol 1. Number 1, 1993.
- [64] Brickman, B.J., Hettinger, L. J., Roe, M. M., Stautberg D.K., Vidulich, M. A., Haas, M. W., & Shaw, R.L. (1995). An assessment of situation awareness in an air combat task: The Global Implicit Measure approach. In D.J. Garland & M.R. Endsley (Eds.), *Experimental analysis and measurement of situation awareness* (pag. 339-344). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
- [65] Brickman, B.J., Hettinger, Stautberg D., Vidulich, M. A., Haas, M. W., & Shaw, R.L. (1999). The Global Implicit Measurement of Situation Awareness: Implications for Design and Adaptive Interface Technologies. In Mark W. Scerbo & Mustapha Mouloua (Eds.), *Automation Technology and Human Performance: Current Research and Trends*. Mahwah, NJ: Lawrence Erlbaum Associates.
- [66] Vidulich, M.A. & Hughes, E.R. (1991). Testing a subjective metric of situation awareness. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1307-11). Santa Monica, CA: The Human Factors and Ergonomics Society.
- [67] Endsley, M.R., Selcon, S.J., Hardiman, T.D., & Croft, D.G. (1998) A comparative evaluation of SAGAT and SART for evaluations of situation awareness In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 82-86). Santa Monica, CA: Human Factors and Ergonomics Society.
- [68] Mica R. Endsley, Betty Bolte, Debra G. Jones. 2003. *Designing for Situation Awareness: An Approach to User-Centered Design*. CRC Press, Taylor & Francis Group. Boca Raton, FL.
- [69] Frank T. Durso and Andrew R. Dattel. 2004. "SPAM: The Real-Time Assessment of SA." In *A Cognitive Approach to Situation Awareness: Theory and Application*, eds. Simon Banbury, Sébastien Tremblay. Ashgate Publishing.
- [70] E. Jeannot, C. Kelly, and D. Thompson. 2003. *The Development of Situation Awareness Measures in ATM Systems*. Document HRS/HSP-005-REP-01. Brussels, Belgium. European Organisation for the Safety of Air Navigation.
- [71] J. D. Lee and N. Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *international Journal of Human-Computer Studies*, Volume 40, Issue 1, Pages 153-184.
- [72] Fischhoff, B., MacGregor, D., 1982. Subjective confidence in forecasts. *Journal of Forecasting* 1, 155-172.

- [73] Fischhoff, B., Slovic, P., & Lichtenstein, S. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, 3, 522-564.
- [74] McCormick, E.J., & Sanders, M. (1982). *Human factors in engineering and design* (6th ed.). New York: MacGraw-Hill.
- [75] Kantowitz, B.H., & Casper, P.A. (1988). Human Workload in Aviation. In Wiener, E.L., and Nagel D.C. (Eds.), *Human Factors in Aviation*. Academic Press, San Diego, CA.
- [76] H. Mackworth, The breakdown of vigilance during prolonged visual search, *Quart. J. exp. Psychol.*, 1, 1948, 6-21.
- [77] Angus, R.G., & Heselgrave, R.J. (1985). Effects of sleep loss on sustained cognitive performance during a command and control simulation. *Behavior Research Methods, instruments, & Computers*, 17:1, 55-67.
- [78] Dinges, D.F., Whitehouse, W.G., Orne, E.C, & Orne, M.T. (1988). The benefits of a nap during prolonged work and wakefulness. *Work and Stress*, 2, 139-153.
- [79] Kribbs, N.B., & Dinges, D. (1994). Vigilance decrement and sleepiness. In R.D. Ogilvie & J.R. Harsh (Eds.). *Sleep onset*. Washington, D.C.: American Psychological Association. (pp.113- 125).
- [80] Belz, Steven M.;Robinson, Gary S.;Casali, John G. 2004. Temporal Separation and Self-Rating of Alertness as Indicators of Driver Fatigue in Commercial Motor Vehicle Operators. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Volume 46,Number 1, Spring 2004 , pp. 154-169(16).
- [81] Akerstedt T, Gillberg M. Subjective and objective sleepiness in the active individual. *Int J Neurosci*. 1990;52:29-37.
- [82] Wickens, C. D., & Xu, X. (2002). Automation trust, reliability and attention HMI 02-03 (AHFD-02-14/MAAD-02-2). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- [83] Muir and Moray, 1996. Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*. v39. 429-460.
- [84] Lim, B. C. & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, 27, 403-418.
- [85] Rentsch JR, Heffner TS, Duffy LT. 1994. What you know is what you get from experience: team experience related to teamwork schemas. *Group and Organization Management* 19, 450-474.
- [86] Janice Langan-Fox, Sharon Code, Kim Langfield-Smith. (2000). Team Mental Models: Techniques, Methods, and Analytic Approaches. *Human Factor*, Vol. 42, No.2, 242-271.
- [87] Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2&3), 259-284.
- [88] Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996). Procedural network representations of sequential data. *Human-Computer Interaction*, 11, 29-68.
- [89] Endsley, M. R. & Jones, W. M. (1997). Situation awareness, information dominance, and information warfare (No. AL/CF-TR-1997-0156). Wright-Patterson AFB, OH: United States Air Force Armstrong Laboratory.
- [90] Cheryl A. Bolstad, Peter Foltz, Marita Franzke, Haydee M. Cuevas, Mark Rosenstein, Anthony M. Costello. Predicting Situation Awareness from Team Communications. In *Proceedings of the Human Factors and Ergonomics Society (HFES) 51st Annual Meeting*, October 1-5, 2007.
- [91] Cooke, N.J., Salas, E., Cannon-Bowers, J.A., and Stout, R. (2000). Measuring team knowledge. *Human Factors*, 42, 151-173.
- [92] Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MIT Press.
- [93] Nehme, C. E., Crandall, J. W., Cumming, M. L., Using Discrete-Event Simulation to Model Situational Awareness of Unmanned-Vehicle Operators, *Virginia Modeling, Analysis and Simulation Center Capstone Conference*, Norfolk, VA, April 2008.

- [94] Biddle, B. J. 1979. *Role Theory: Expectations, Identities, and Behaviors*. New York: Academic Press.
- [95] Greenberg, J. 2004. *Managing Behavior in Organizations*. New Jersey: Pearson Prentice Hall.
- [96] Smith, C.A., (2006) *An Ecological Perceptual Aid for Precision Vertical Landings*, Thesis, (MS) MIT Aeronautics and Astronautics, Cambridge, MA.
- [97] Marquez, J.J., (2007) *Human-Automation Collaboration: Decision Support for Lunar and Planetary Exploration*, Unpublished Dissertation, MIT Aeronautics and Astronautics, Cambridge, MA.
- [98] Graham, H.D., Cummings, M.L. (2007). *Assessing the Impact of Auditory Peripheral Displays for UAV Operators*, (HAL2007-09), MIT Humans and Automation Laboratory, Cambridge, MA.
- [99] Wierwille, W.W., & Casali, J.G. (1983). A validated rating scale for global mental workload measurement applications, *Proceedings of the Human Factors Society 27th Annual Meeting*, Santa Monica, CA, 129-133.
- [100] Skipper, J. H., Rieger, C.A., and Wierwille, W.W. 1986. Evaluation of decision tree rating scales for mental workload estimation. *Ergonomics*, 29, 585-599.
- [101] Roscoe, A. H.; Ellis, G. A. 1990. *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use*. (Tech. Rep. No. TR90019). Farnborough, England: Royal Aeronautical Establishment.
- [102] Reid, G.B., & Nygren, T.E. 1988. *The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload*. In *Human Mental Workload*, P. Hancock & N. Meshkati (Eds.). Amsterdam, The Netherlands: North Holland.
- [103] Hart, S.G., and Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In *Human Mental Workload*, P. Hancock, N. Meshkati (Eds.) pp. 139-183. Amsterdam, The Netherlands: North Holland B.V.
- [104] Taylor, R.M. *Situational awareness rating technique (SART): the development of a tool for aircrew systems design*. *Proceedings of the NATO Advisory Group for Aerospace Research and Development (AGARD) Situational Awareness in Aerospace Operations Symposium (AGARD-CP-478)*; October 1989, p. 17.
- [105] Fracker, M.L. (1991). *Measures of situation awareness: Review and future directions*. (AL-TR-1991-0128). Wright-Patterson AFB OH: Armstrong Laboratory, Crew Systems Directorate.
- [106] Waag, W. L. & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine*, 65(5, Suppl.), A13-A19.
- [107] Dennehy, K. (1997). *Cranfield - Situation Awareness Scale, User Manual*. Applied Psychology Unit, College of Aeronautics, Cranfield University, COA report N0 9702, Bedford, January.
- [108] Matthews, M. D., Beal, S. A., & Pleban, R. J. (2002). *Situation Awareness in a Virtual Environment: Description of a Subjective Measure*. (Research Report 1786). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- [109] McGuinness, B., & Foy, L. (2000). A subjective measure of SA: The Crew Awareness Rating Scale. In D. B. Kaber and M. R. Endsley (Eds). *Human performance, situation awareness and automation: User centered design for the new millennium*. Atlanta, GA: SA Technologies.
- [110] Adams, S. *Practical considerations for measuring Situational Awareness*. *Proceedings for the Third Annual Symposium and Exhibition on Situational Awareness in the Tactical Air Environment*, 1998, 157-164.
- [111] J. Jian, A. M. Bisantz, C. G. Drury. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 2000, Vol. 4, No. 1, Pages 53-71.
- [112] Madsen, M. & Gregor, S. (2000). *Measuring human-computer trust*. In *Proceedings of Eleventh Australasian Conference on Information Systems*, Brisbane, 6-8 December.
- [113] P. Goillau, C. Kelly, M. Boardman, and E. Jeannot. (2003). *Guidelines for Trust in Future ATM Systems: Measures*. Document HRS/HSP-005-GUI-02. Brussels, Belgium. European Organisation for the Safety of Air Navigation.

- [114] Cavaleri, S., & Sterman, J. D. (1997). Towards evaluation of systems thinking interventions: A case study. *System Dynamics Review*, 13, 171-186.
- [115] Redding, R. E., & Cannon, J. R. (1992). Expertise in air traffic control (ATC): What is it, and how can we train for it? In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1326-1330). Santa Monica, CA: Human Factors and Ergonomics Society.
- [116] Rusbult, C. E., Onizuka, R. K., & Lipkus, I. (1993). What do we really want? Mental models of romantic involvement explored through multidimensional scaling. *Journal of Experimental Social Psychology*, 29, 493-527.
- [117] Wyvman, B. G., & Randell, T. M. (1998). The relation of knowledge organization to performance of a complex cognitive task. *Applied Cognitive Psychology*, 12, 251-264.
- [118] Markoczy, L. & Goldberg, J. (1995). A method for eliciting and comparing causal maps. *Journal of Management*, 21(2), 305-333.
- [119] Von Hecker, U. (1997). How do logical inference rules help construct social mental models? *Journal of Experimental Social Psychology*, 33, 367-400.
- [120] Carpendale, J. L., McBride, M. L., & Chapman, M. (1996). Language and operations in children's class inclusion reasoning: The operational semantic theory of reasoning. *Developmental Review*, 16, 391-415.
- [121] Daniels, K., de Chematony, L., & Johnson, G. (1995). Validating a method for mapping managers' mental models of competitive industry structures. *Human Relations*, 48, 975-991.
- [122] Reger, R. K. (1990). Managerial thought, structures and competitive positioning. In A. Huff (Ed), *Mapping Strategic Thought*. Chichester, Wiley.
- [123] Young, F. W. and Hamer, R. M. (1994). *Theory and Applications of Multidimensional Scaling*. Erlbaum Associates. Hillsdale, NJ.
- [124] Markoczy, L., & Goldberg, J. (1995). A method for eliciting and comparing causal maps. *Journal of Management*, 21, 305-333.
- [125] Schvaneveldt, R. W. (Ed.) (1990) *Pathfinder Associative Networks: Studies in Knowledge Organization*. Norwood, NJ: Ablex.

Appendix A: Unidimensional Workload Self-Rating Scales

– Modified Cooper-Harper Scale [99]

This technique consists of a structured rating to elicit a single dimensional rating. Participants are requested to follow a binary decision tree containing increasingly specific questions in order to reach a final rating between 1 and 10. The scale is shown in Figure 14.

It is argued that conventional scales can suffer from operator judgment and selection variability, whereas a decision tree flowchart scale may reduce the variability due to its tighter structure [100].

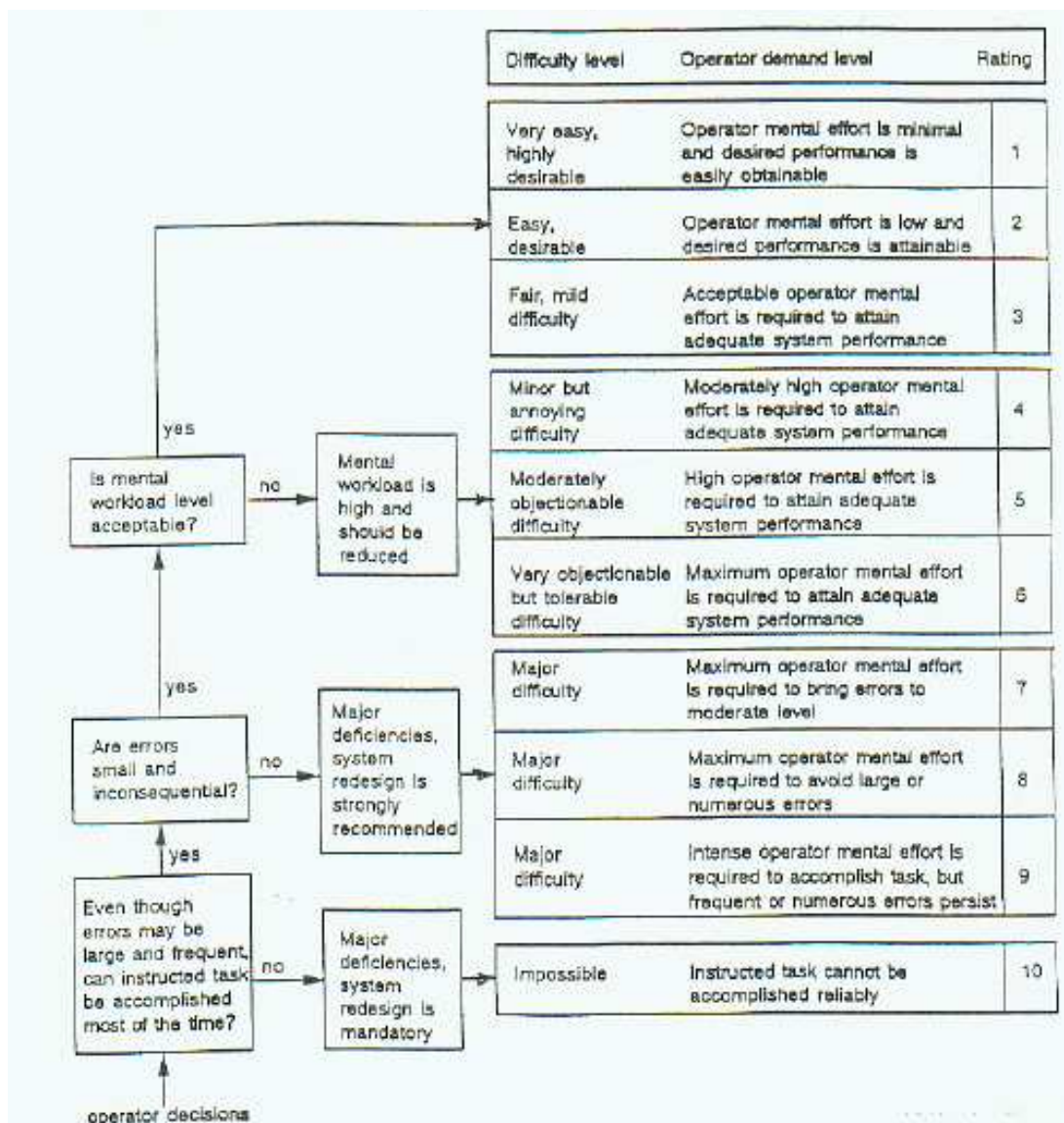


Figure 14: Modified Cooper-Harper Scale.

– Bedford Workload Scale [101]

This scale was created by trial and error with the help of Royal Aircraft test pilots to assess pilot workload. The Bedford scale also involves the use of a hierarchical decision tree to assess participant workload; participants follow the decision tree to derive a workload rating for the task under analysis. The scale is shown in Figure 15.

This scale is similar to the Modified Cooper-Harper scale, as it also uses a hierarchical decision tree. However, this scale was specially developed for pilot workload assessment, whereas the Modified-Cooper Harper scale was developed for applications beyond aircraft environment.

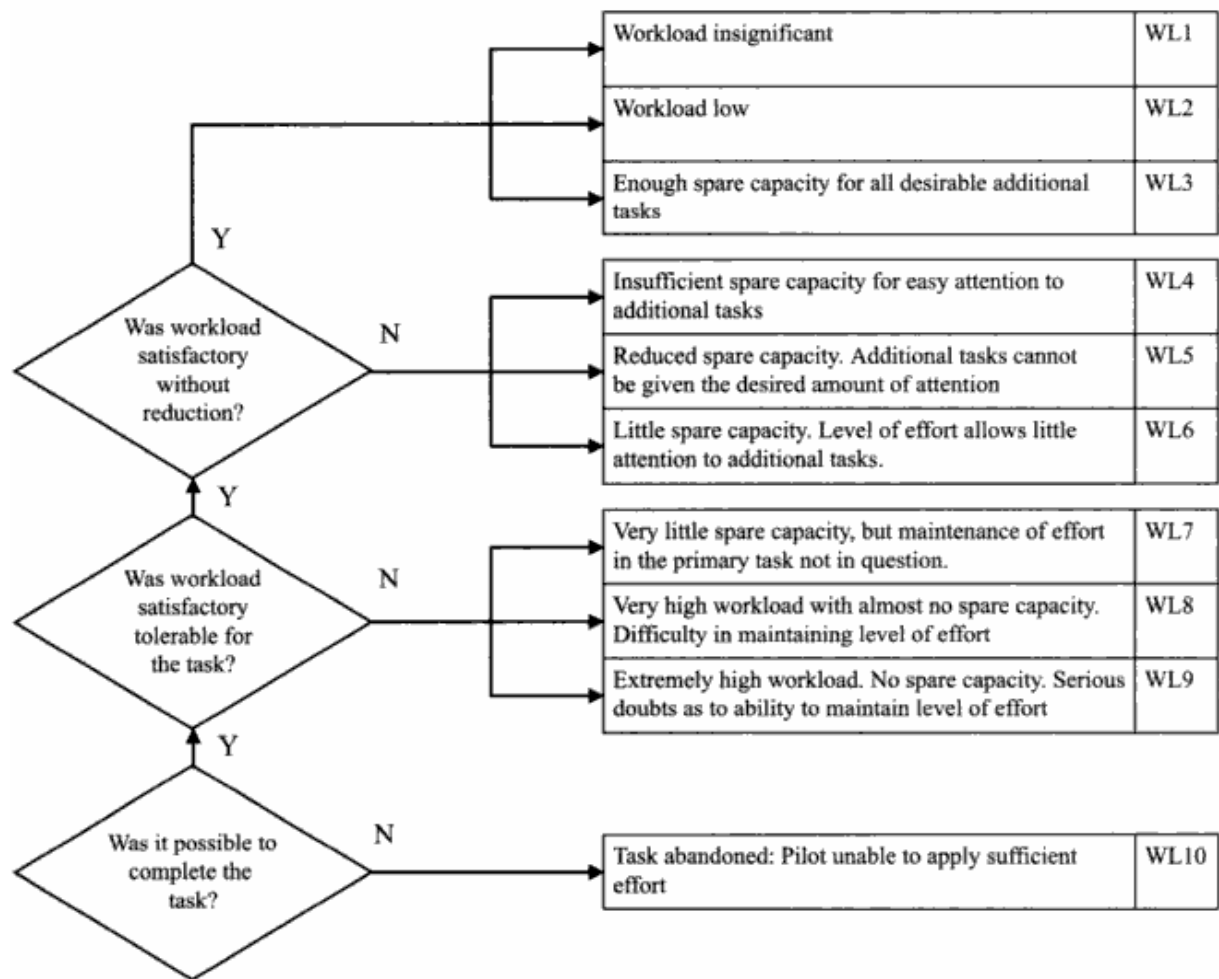


Figure 15: Bedford Scale.

Appendix B: Multidimensional Workload Self-Rating Scales

– Subjective Workload Assessment Technique (SWAT) [102]

SWAT rates three major workload dimensions: time load, mental effort, and psychological stress, through a 3-point scale. This technique involves a two-step procedure: scale development, and event rating. In the first step, participants rank all possible 27 combinations of the 3 levels of the 3 workload dimensions, based on what they consider to be the lowest to highest workload. Then the SWAT scale is developed by calculating the corresponding score (1 to 100; 0 represents virtually no perceived workload and 100 represents high workload) for every combination of ratings on the three subscale. Then, participants perform the task under analysis and report the perceived workload, rating each of the workload dimensions. The researcher then maps the set of ratings to the SWAT score (1 to 100), which has been calculated during the development phase. This data is transformed by means of conjoint measurement, into an interval scale of workload. Table 24 shows the SWAT rating scales.

Table 24: SWAT Rating Scales.

TIME LOAD	MENTAL EFFORT LOAD	STRESS LOAD
1. Often have spare time: interruptions or overlap among activities occur infrequently or not at all.	1. Very little conscious mental effort or concentration required: activity is almost automatic, requiring little or no attention.	1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated
2. Occasionally have spare time: interruptions or overlap among activities occur frequently	2. Moderate conscious mental effort or concentration required: complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity; considerable attention is required.	2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload: significant compensation is required to maintain adequate performance.
3. Almost never have spare time: interruptions or overlap among activities are very frequently, or occur all the time	3. Extensive mental effort and concentration are necessary: very complex activity requiring total attention.	3. High to very intense stress due to confusion, frustration, or anxiety: high to extreme determination and self-control required.

– NASA TLX [103]

In this technique participants are required to rate six subscales: mental demand, physical demand, temporal demand, effort, performance, and frustration. Then, participants are repeatedly asked to choose which of a pair of subscales contributes more to their overall workload, until all possible pairs of subscales have been compared. In order to calculate the workload metric, the ratings from the six subscales are combined into a single weighted measure of workload using the number of times a particular subscale

was preferred as its weight. Examples of the NASA TLX rating scales and pair scale-rating questions are shown in Figures 16-17.

An advantage of this technique is that there is a free automated computer application of the NASA-TLX available for download at the NASA website.

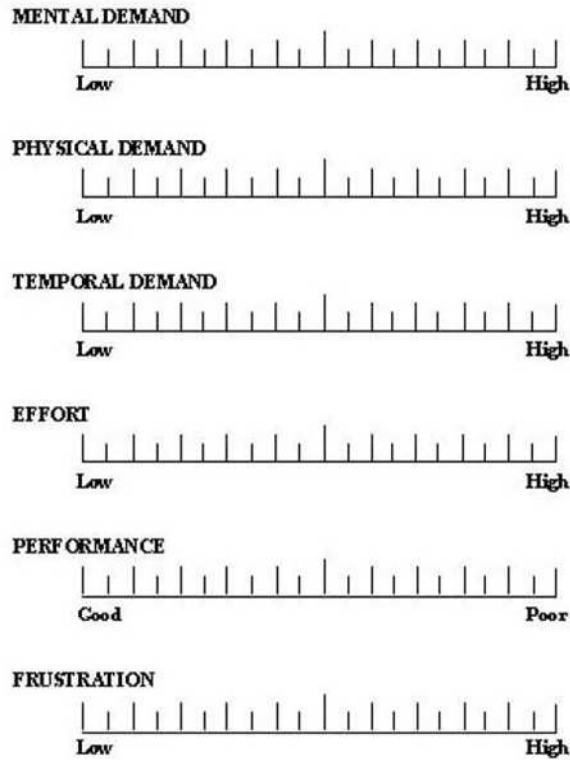


Figure 16: NASA TLX Rating Scales.

Which rating contributed more to your personal workload level?

<input type="checkbox"/> Frustration or <input type="checkbox"/> Effort

Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent you felt during this scenario.

Effort: How hard you had to work (how much effort did you put in mentally and physically) to accomplish your level of performance.

Figure 17: Example of NASA TLX Pair Rating Questions.

Appendix C: SA Subjective Metrics and Techniques

- Situation Awareness Rating Technique (SART) developed by Taylor [104]

SART requires the operator to rate the 3 or 10 dimensions shown in Table 25¹⁴. Operators are asked post-trial to rate each dimension on a Likert scale of 1 to 7. Alternatively, specific categories (low vs. high) or pairwise comparisons can also be used. The SA metric is calculated based on the total scores obtained for each of the three dimensions. The formula to calculate the SA metric is “Understanding - (Attentional Demand – Attentional Supply)”.

Table 25: SART Dimensions.

Domains	Construct	Definition
Attentional Demand	<i>Instability of situation</i>	Likelihood of situation to change suddenly
	<i>Variability of situation</i>	Number of variables that require attention
	<i>Complexity of situation</i>	Degree of complication of situation
Attentional Supply	<i>Arousal</i>	Degree that one is ready for activity
	<i>Spare mental capacity</i>	Amount of mental ability available for new variables
	<i>Concentration</i>	Degree that one's thoughts are brought to bear on the situation
	<i>Division of attention</i>	Amount of division of attention in the situation
Understanding	<i>Information quantity</i>	Amount of knowledge received and understood
	<i>Information quality</i>	Degree of goodness of value of knowledge communicated
	<i>Familiarity</i>	Degree of acquaintance with situation experience

SART dimensions are generic and so can be applied to many different domains. SART is a widely used method and has a number of associated validation studies. However, the SART dimensions only reflect a limited portion of SA, and the rating is sensitive also to performance and workload differences. Testing of the technique often reveals a correlation between SA and performance, and also between SA and workload.

- Situation Awareness – Subjective Workload Dominance Technique (SA-SWORD) developed by Vidulich & Hughes [66]

This technique requires participants to do comparative self-ratings, comparing self-assessed SA from one trial to another. Therefore, SA-SWORD can only be used with within-participants experimental designs. A judgment matrix comparing each task to every other task is filled in with each subject's evaluation of the tasks. The SA metric is calculated using a geometric means approach.

The SA-SWORD technique can be useful to compare two different interface design concepts and their effect upon operator SA. In addition, comparative self-ratings encourage within-participant consistency but in some situations, the number of comparisons required can become quite large, making

¹⁴ . The 3-D SART requires the subject to rate only the 3 domains but the 10-D SART requires rating the 10 constructs.

the use of these measures impractical [105]. Finally, there is limited evidence of the use of this technique in the literature.

- SA Rating Scale (SARS) developed by Waag & Houck [106]

SARS can be used as a self-rating or an observer rating technique. It was developed for aviation and it measures SA rating the 31 behaviors presented in Table 26 in a 6-point scale with acceptable and outstanding as anchors. The 31 SARS behaviors are divided into 8 categories representing phases of mission performance. These categories and associated behaviors were developed from interviews with experienced pilots. The SA metric is obtained by calculating an average score for each category and also a total SARS score (sum of all rating).

Table 26: SARS Behaviors.

General traits	Information interpretation
Discipline	Interpreting vertical situation display
Decisiveness	Interpreting threat warning system
Tactical knowledge	Ability to use controller information
Time-sharing ability	Integrating overall information
Spatial ability	Radar sorting
Reasoning ability	Analysing engagement geometry
Flight management	Threat prioritisation
Tactical game plan	Tactical employment – BVR
Developing plan	Targeting decisions
Executing plan	Fire-point selection
Adjusting plan on-the-fly	Tactical employment – Visual
System operation	Maintain track of bogeys/friendlies
Radar	Threat evaluation
Tactical electronic warfare system	Weapons employment
Overall weapons system proficiency	Tactical employment -- General
Communication	Assessing offensiveness/defensiveness
Quality (brevity, accuracy, timeliness)	Lookout
Ability to effectively use information	Defensive reaction
	Mutual support

This scale combines assessments on many dimensions besides SA, including decision making abilities, flight skills, performance, and the subjective impressions of a person’s personality traits. Moreover, the scale is closely tied to the particular aircraft type and mission, so the applicability of this measure to other domains is doubtful.

- Cranfield Situation Awareness Scale (C-SAS) developed by Dennehy [107]

C-SAS can be used as a self-rating or an observer rating technique. It requires to rate each of these five subscales: knowledge; understanding and anticipation of future events; management of stress, effort and commitment; capacity to perceive, attend, assimilate and assess information; and overall situation awareness. The SA metric is calculated by adding all the subscales scores together. A high score indicates a high level of SA. The technique can be used during or after the experiment. C-SAS has been subjected to only limited use.

- Mission Awareness Rating Scale (MARS) developed by Matthews, Beal, and Pleban [108]

MARS is a development of the CARS technique¹⁵, but it was designed specifically for use in the assessment of SA in military exercises. The technique comprises two separate sets of four questions. The first set of four questions pertain to assessing SA content, for example, how well the respondent thinks he or she understands the situation. The second set of questions addresses workload, for example, how much mental effort is required to achieve understanding in a given situation. In both sets, the first three questions are about the ease of identification, understanding, and projection of mission critical cues; the fourth assesses how aware the participant felt during the mission. Figure 18 illustrates the first part of this questionnaire.

Content subscales

Please rate your ability to identify mission-critical cues in this mission.

<input type="checkbox"/>	Very easy- able to identify all cues
<input type="checkbox"/>	Fairly easy – could identify most cues
<input type="checkbox"/>	Somewhat difficult – many cues hard to identify
<input type="checkbox"/>	Very difficult – had substantial problems identifying most cues

How well did you understand what was going on during the mission?

<input type="checkbox"/>	Very well – fully understood the situation as it unfolded
<input type="checkbox"/>	Fairly well – understood most aspects of the situation
<input type="checkbox"/>	Somewhat poorly – had difficulty understanding much of the situation
<input type="checkbox"/>	Very poorly – the situation did not make sense to me

How well could you predict what was about to occur next in the mission?

<input type="checkbox"/>	Very well – could predict with accuracy what was about to occur
<input type="checkbox"/>	Fairly well – could make accurate predictions most of the time
<input type="checkbox"/>	Somewhat poor – misunderstood the situation much of the time
<input type="checkbox"/>	Very poor – unable to predict what was about to occur

How aware were you of how to best achieve your goals during this mission?

<input type="checkbox"/>	Very aware – knew how to achieve goals at all times
<input type="checkbox"/>	Fairly aware – knew most of the time how to achieve mission goals
<input type="checkbox"/>	Somewhat unaware – was not aware of how to achieve some goals
<input type="checkbox"/>	Very unaware – generally unaware of how to achieve goals

Figure 18: MARS Questionnaire (Part 1: ability to detect and understand important cues).

The technique was developed for field trials instead of simulation trials and it is normally administered after the trial or mission completion. MARS can be used across domains with minimal modifications. A potential disadvantage is the construct validity of this technique; it could be argued that MARS rates the difficulty in acquiring and maintaining SA rather than the actual SA because the second set of questions is formulated as “how difficult –in terms of mental effort– was to”. In addition, the technique has only limited validation evidence.

¹⁵ Crew Awareness Rating Scale (CARS) was developed by McGuinness and Foy [109].The technique elicits self-ratings of SA post-trial from participants and it also consists on two sets of four questions each.

- China Lake SA (CLSA) developed by Adams [110]

CLSA is a five-point rating scale based on the Bedford Workload Scale. It was designed at the Naval Air Warfare Center at China Lake to measure SA in flight. CLSA is a unidimensional scale, which can be insufficient for capturing SA's richness and complexity. In addition, the technique has only limited validation evidence. The rating scale is presented in Figure 19.

SA SCALE VALUE	CONTENT
VERY GOOD 1	<ul style="list-style-type: none"> • Full knowledge of a/c energy state/tactical environment/mission; • Full ability to anticipate/accommodate trends
GOOD 2	<ul style="list-style-type: none"> • Full knowledge of a/c energy state/tactical environment/mission; • Partial ability to anticipate/accommodate trends; • No task shedding
ADEQUATE 3	<ul style="list-style-type: none"> • Full knowledge of a/c energy state/tactical environment/mission; • Saturated ability to anticipate/accommodate trends; • Some shedding of minor tasks
POOR 4	<ul style="list-style-type: none"> • Fair knowledge of a/c energy state/tactical environment/mission; • Saturated ability to anticipate/accommodate trends; • Shedding of all minor tasks as well as many not essential to flight safety/mission effectiveness
VERY POOR 5	<ul style="list-style-type: none"> • Minimal knowledge of a/c energy state/tactical environment/mission; • Oversaturated ability to anticipate/accommodate trends; • Shedding of all tasks not absolutely essential to flight safety/mission effectiveness

Thresholds – 1 (very good) to 5 (very poor).

Figure 19: China Lake SA Rating Scale.

Appendix D: Multiple Rating Scales to Elicit Dimensions of Trust

Jian, Bizantz, and Drury developed from empirical evidence a twelve-item questionnaire to measure trust in automation [111]. This questionnaire incorporates a seven point rating scale in the range from “not at all” to “extremely”. Subjects are requested to rate the degree of agreement or disagreement of with these twelve trust-related statements. This measure represents the first attempt at empirically generating a scale to measure trust in automation. Figure 20 shows the questionnaire.

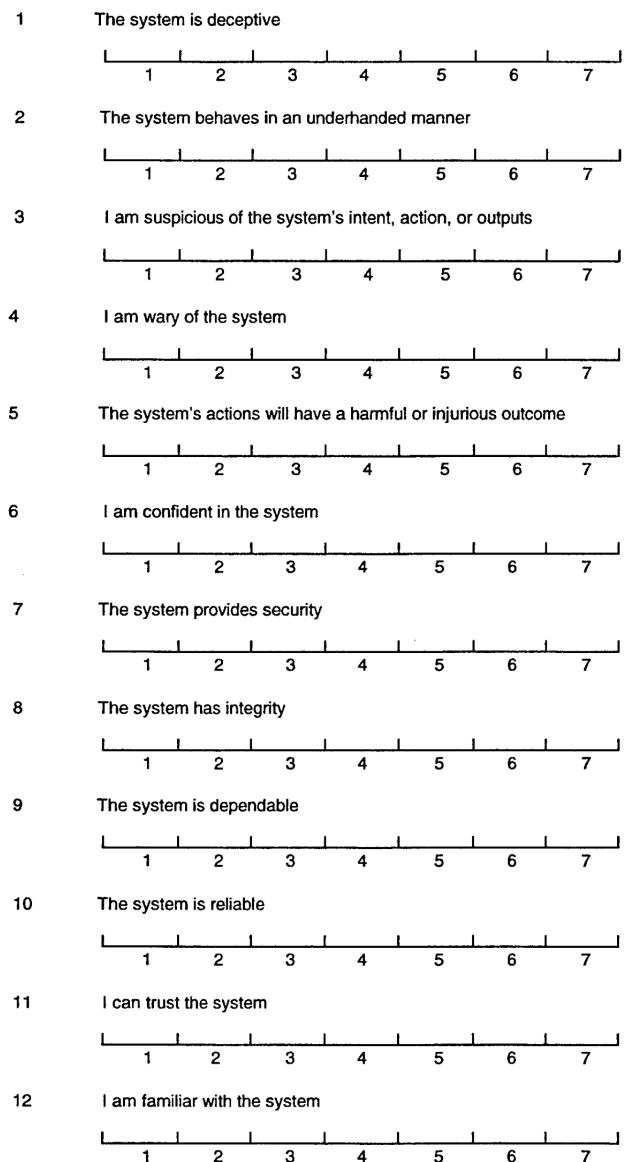


Figure 20: Twelve item questionnaire developed by Jian et al. [111]

Madsen and Gregor developed the Human-Computer Trust (HCT) scale, which consists of five main constructs each with five sub-items as shown in Figure 21 [112]. These five items are drawn from an original list of ten trust constructs as having the most predictive validity. Madsen and Gregor claim that the HCT has been empirically shown to be valid and reliable.


1. ***Perceived Reliability***
 - R1 - The system always provides the advice I require to make my decision.
 - R2 - The system performs reliably.
 - R3 - The system responds the same way under the same conditions at different times.
 - R4 - I can rely on the system to function properly.
 - R5 - The system analyzes problems consistently.
2. ***Perceived Technical Competence***
 - T1 - The system uses appropriate methods to reach decisions.
 - T2 - The system has sound knowledge about this type of problem built into it.
 - T3 - The advice the system produces is as good as that which a highly competent person could produce.
 - T4 - The system correctly uses the information I enter.
 - T5 - The system makes use of all the knowledge and information available to it to produce its solution to the problem.
3. ***Perceived Understandability***
 - U1 - I know what will happen the next time I use the system because I understand how it behaves.
 - U2 - I understand how the system will assist me with decisions I have to make.
 - U3 - Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
 - U4 - It is easy to follow what the system does.
 - U5 - I recognize what I should do to get the advice I need from the system the next time I use it.
4. ***Faith***
 - F1 - I believe advice from the system even when I don't know for certain that it is correct.
 - F2 - When I am uncertain about a decision I believe the system rather than myself.
 - F3 - If I am not sure about a decision, I have faith that the system will provide the best solution.
 - F4 - When the system gives unusual advice I am confident that the advice is correct.
 - F5 - Even if I have no reason to expect the system will be able to solve a difficult problem, I still feel certain that it will.
5. ***Personal Attachment***
 - P1 - I would feel a sense of loss if the system was unavailable and I could no longer use it.
 - P2 - I feel a sense of attachment to using the system.
 - P3 - I find the system suitable to my style of decision making.
 - P4 - I like using the system for decision making.
 - P5 - I have a personal preference for making decisions with the system.

Figure 21: Human-Computer Trust (HCT) Scale developed by Madsen & Gregor [112]

A limitation of this technique is that it assumes that the user has already several months of experience with the system.

Figure 22 shows another multidimensional rating scale, the SHAPE Automation Trust Index [113]. This technique was developed by Eurocontrol to measure human trust in Air Traffic Management systems. This measure is primarily concerned with human trust of ATC computer-assistance tools and other forms of automation support, which are expected to be major components of future ATM systems. This questionnaire is based on Madsen & Gregor work. This questionnaire was developed in close collaboration with air traffic controllers, updating and simplifying the original Madsen's & Gregor's questionnaire to make it easy to use and understand. In particular usability evaluation trials, and construct validity feedback from air traffic controllers were collected. A main advantage of this technique is that their creators did actively work to create their scales to reflect how air traffic controllers understand trust.


1. What did you think of the simulation? (Please mark the scale with an 'X').

Bad		Good
------------	--	-------------

2. Were you prepared to trust the simulated system?

No	Yes
-----------	------------

3. How much confidence did you have in the simulated system? (Please mark the scale with an 'X').

None		Full
0%	50%	100%

4. Please give your reasons. If your trust or level of confidence in the system has changed since the start of the day, please explain why.

Would you work live traffic with the tools? In your opinion, what changes would the automation need so that your trust and confidence would be increased? If there are any other factors which influence your trust in an ATC system, or if you have any general comments, please write them here.

5. Please judge each automation tool against the following factors (mark each scale with an 'X').

Name of automation tool: _____			
1. Is the automation tool useful?	☹️ <i>Not useful</i>	- 5 0 + 5	<i>Useful</i> ☺️
2. How reliable is it?	☹️ <i>Not reliable</i>	- 5 0 + 5	<i>Reliable</i> ☺️
3. How accurately does it work?	☹️ <i>Not accurate</i>	- 5 0 + 5	<i>Accurate</i> ☺️
4. Can you understand how it works?	☹️ <i>Not understand</i>	- 5 0 + 5	<i>Understand</i> ☺️
5. Do you like using it?	☹️ <i>Dislike</i>	- 5 0 + 5	<i>Like</i> ☺️
6. How easy is it to use?	☹️ <i>Difficult</i>	- 5 0 + 5	<i>Easy</i> ☺️

6. Please rank these factors in order of relative importance. Number them from 1 (*least important*) to 6 (*most important*). Please use each number once only.

Name of automation tool: _____	
Usefulness	<i>ranking:</i>
Reliability	<i>ranking:</i>
Accuracy	<i>ranking:</i>
Understanding	<i>ranking:</i>
Liking	<i>ranking:</i>
Ease of use	<i>ranking:</i>

7. Please rate your amount of confidence in each of these five dimensions. Please mark each scale with an 'X'.

1. Confidence in automation tools	0 50 100 %
2. Confidence in simulation	0 50 100 %
3. Self-confidence	0 50 100 %
4. Confidence in controller colleagues	0 50 100 %
5. Confidence in pilots	0 50 100 %

Figure 22: SHAPE Automation Trust Index developed by Eurocontrol [113]

Appendix E: Knowledge Elicitation Techniques

– Cognitive Interviewing

Cognitive interviewing techniques, such as open interviews, question-answer interviews, and inferential flow analysis, can be used to elicit mental models. For example, in the open interview form, the researcher engages the participant in an open conversation to elicit domain concepts and the relationships between them. This technique has been used, for example, by Cavaleri & Sterman [114], and Redding & Cannon [115].

Although cognitive interviewing is straight-forward, it relies heavily on the interviewer's interpretations. In addition, this technique only captures information that can be expressed verbally. This technique is recommended at the very initial stage of a research, as a starting point for obtaining information about the domain of interest.

– Verbal Protocols

In this technique, participants are asked to think aloud and the researcher identifies the relationships between concepts from participants' verbalizations. This technique is used to obtain information about decision making strategies and general reasoning processes. It is particularly useful for uncovering decision making errors [86].

Its limitations are the labor-intensive process of collecting and analyzing data, and the difficulty to systematically make comparisons among participants. In addition, humans have limited consciousness of their actual thought structures and this technique usually provides an incomplete picture. However its use is recommended in domains in which verbalization is a normal part of task performance. Furthermore, it can provide valuable insights into participants' cognitive strategies if used in conjunction with other behavioral metrics for attention allocation and information processing efficiency.

– Pairwise Ratings

In this technique, participants are presented with all possible pairs of concepts one pair at a time, and they are requested to provide a similarity or relatedness ratings for each pair. Then, a matrix of pairwise ratings is created that can be analyzed with multidimensional scaling [116] or the Pathfinder algorithm [117].

This method is time efficient, it requires little reading or writing, and the mental model is not articulated by the participant, but inferred through statistical analysis. A disadvantage is that the repetitive nature of pairwise ratings can induce a response set. The pairwise rating method is recommended when research time is constrained.

- Causal Mapping

This is the most commonly used technique in an organizational or management setting. Participants are presented an expert-generated list of concepts and they are requested to select the “n” most important concepts. These “n” concepts are then placed in a “nxn” matrix and participants indicate if one concept causes the other. Participants need to identify the direction of the causal relation (concepts in the rows are sources while those in the columns are targets), its nature (+ or -), and its strength (usually -3 to 3). To compare two causal maps, the causal mapping matrices can be analyzed using a Distance Ratio (DR) Index [118]. This technique is recommended for the elicitation of information about the causal relations between concepts.

- Card Sorting

This technique involves sorting a series of cards, each labeled with a concept, into groups that make sense to participants. Participants also have to explain why they arranged the cards in those groups. This technique has been used to examine various types of models, including social mental models [119], reasoning mental models [120], and mental models of competitive industry structures [121]. In addition, card sorting is used by many information architects as an input to the structure of a site or product.

Card sorting is quick, inexpensive, easy-to-administer, and flexible. However, in contrast to other techniques, visual card sorting might capture only “surface” characteristics, eliciting knowledge that is easily and often accessed from short-term memory. Another disadvantage is that card sorting is an inherently content-centric technique that neglects users’ tasks. In addition, the sorting is quick, but the analysis of the data can be difficult and time consuming, particularly if there is little consistency between participants. This technique is recommended when research time is restricted.

- Repertory Grid Method (RGM)

In this technique, 3 concepts are presented at a time to participants, and they are requested to describe how 2 of these concepts differ from the third on a particular dimension. After several trios of concepts, a list of dimensions or constructs with opposing poles is obtained. Next, the concepts are rated using these dimensions. The data elicited through this method is usually analyzed with multidimensional scaling. An example of the application of this method can be found in [122].

Advantages of this method include high validity, it is well grounded theoretically, George Kelly’s Personal Construct Theory, and reliability, it produces similar representations over time) [86]. A disadvantage is the prohibitive amount of time required to administer the technique. RGT should be used only when there is ample research time available.

Appendix F: Mental-Model Analysis and Representation Techniques

- Multidimensional Scaling (MDS)

MDS generates spatial configurations that give a pictorial representation of how concepts are clustered within a multidimensional space. The strength of a dimension in a mental model is calculated in terms of a structural ratio (i.e., the ratio of the mean distance between concepts in the same category to the mean distance between concepts in different categories). More information about this technique can be found in [123].

This technique can be used to identify the dimensions that an individual uses to judge the similarity between clusters of concepts and the dominance of a particular concept within an individual's mental model. A potential problem with this technique is that there are a number of variations of scaling techniques to choose from, and the most appropriate technique is not always easy to identify [86].

- Distance ratio Formula (DR)

DR calculates the degree of similarity between two maps, represented as expanded association matrices. The idea is to sum the differences between the two maps and then divide that sum by the greatest possible difference (if DR = 0, then the maps are identical; if DR = 1, then the distance between the maps is maximum). More information about this technique can be found in [124].

This technique can be used to isolate 3 types of differences: differences in the strengths of commonly held beliefs, differences attributable to the existence or nonexistence of beliefs involving common concepts, and differences attributable to beliefs consisting of unique concepts. A disadvantage is that the formula treats the absence of a link between two concepts the same as the absence of a link attributable to the absence of a concept. Another problem is that the formula cannot be generalized to maps of different types [86].

- Pathfinder Algorithm

Pathfinder is a computerized networking technique that is used to derive associative networks based on perceived relatedness among a selected set of concepts. It takes in raw scores (i.e., pairwise comparisons) in a form of upper or lower triangle matrix and generates a network. Concepts that are highly related are separated by a few links and appear close together in the Pathfinder network. More information about this technique can be found in [125].

A disadvantage of this technique is that the layout of items in a Pathfinder network is arbitrary (i.e., it represents associative but not semantic information about conceptual relationships) [86].