

Measurement of Trust in Human-Robot Collaboration

Amos Freedy, Ph.D.
Ewart DeVisser
Gershon Weltman, Ph.D.
Perceptronics Solutions, Inc.
info@percsolutions.net
Nicole Coeyman
U.S. Army RDECOM-STTC
Nicole.Coeyman@us.army.mil

ABSTRACT

We describe a Collaborative Performance Model that captures the critical performance attributes of the distinctive human-robotic decision and control environment. The literature and our initial experimental studies show that the element of trust in human-robot collaboration is an extremely important factor in the Performance Model, and accordingly we have focused much of our attention on deriving suitable and practical measures of this variable. In this paper we describe the formulation of a decision-analytical based measure of trust as well as the results of two initial experiments designed to examine trust in a tactical human-robot collaborative task performed in our new Mixed Initiative Team Performance Assessment System (MITPAS) simulation environment.

KEY WORDS: Human Robot Collaboration, Mixed-Initiative Systems, Human-Robot Performance Modeling, Measurement of Trust

1. INTRODUCTION

Unmanned vehicles and other robotic systems are being introduced into Army systems to extend manned capabilities and act as “force multipliers” (Barnes, Parasuraman & Cosenzo, in press). Because of the resulting increase in the cognitive workload demands on the soldier, increasing levels of robot autonomy, as well as adaptive automation aids, may be required to enhance soldier and system performance. The resulting human-robot team represents a *collaborative mixed*

initiative system — an intermediate stage between the far-off goal of full robot autonomy and the unacceptable use of continuous manual teleoperation by the soldier.

Collaborative mixed initiative systems can in principle yield significant benefits in terms of mission effectiveness. However, appropriate and sensitive metrics are needed for evaluating collaborative human-robot team performance in order to determine the appropriate levels of robot autonomy and automation that will lead to enhanced system performance [1]. This is because mixed initiative introduces a new and unique aspect to the psychology of team performance: the interaction of two cognitive systems -- human and autonomous or semi-autonomous unmanned robot. In addition to the critical performance factors associated with human teams -- which include information exchange, communication, supporting behavior and team leadership -- the mixed manned/unmanned team adds a number of challenging new dimensions. Foremost among these is the ability of the human team to predict, collaborate and develop trust with unmanned systems that may sometimes exhibit fuzzy responses in unstructured and unpredictable environments [3] [5].

The major challenge in our work has been to develop system-specific measures of behavior on which to base assessment of the mixed initiative team performance. As reported by Freedy et al [8], we have developed a performance model for mixed initiative tasks. The performance model represents a critical challenge since the measures must be unique to the information and decision environment associated with human-robot

teams and must also link together behavioral processes important to tactical outcomes.

The measures need to provide feedback for skill improvement in collaboration as well as adaptation to stress and workload, and they should help define the training needs themselves. The performance model formed the basis for implementation of the measurement system that was used in conjunction with our initial simulation environment.

2. PERFORMANCE MODEL

Our approach to human robot collaborative performance measurement has been to develop a multi-dimensional Collaborative Performance Model that captures the critical performance attributes of the distinctive human-robotic decision and control environment.

The Performance Model draws on four separate research areas that have been pursued independently in the past but which are integrated here to establish meaningful criteria of team performance: (1) Psychology of Team Performance, (2) Unmanned Systems, (3) Mixed Initiative Systems and (4) War Fighting Behavior.

and training. Most critical are variables related to the decision-making behavior of the unmanned systems, such as behavior transparency to the human collaborators, human trust in robot decisions, and human ability to synergize the autonomy of robots so as to add to the capability of the total team.

Our objective is to identify the dimensions of performance which contribute to effective outcomes of collaborative manned-unmanned tasks and, in particular, to formulate measures to evaluate training in processes that are unique to the collective team of humans and robots. In essence, our approach has been to create a taxonomy of specific processes which can be decomposed into explicit behavioral objectives side-by-side with measures of effectiveness based on actual outcomes. Our focus is on process measures that are closely linked to outcomes, because it is these measures that will provide the feedback necessary for training.

The three levels of team processes critical to training evaluation and remediation are: (1) individual human; (2) team human; and (3) collective human/robot team. We decomposed the processes into these three levels and developed taxonomy of measures for each level. We narrowed the performance measures to the simplest factor structure that adequately covered the dimension

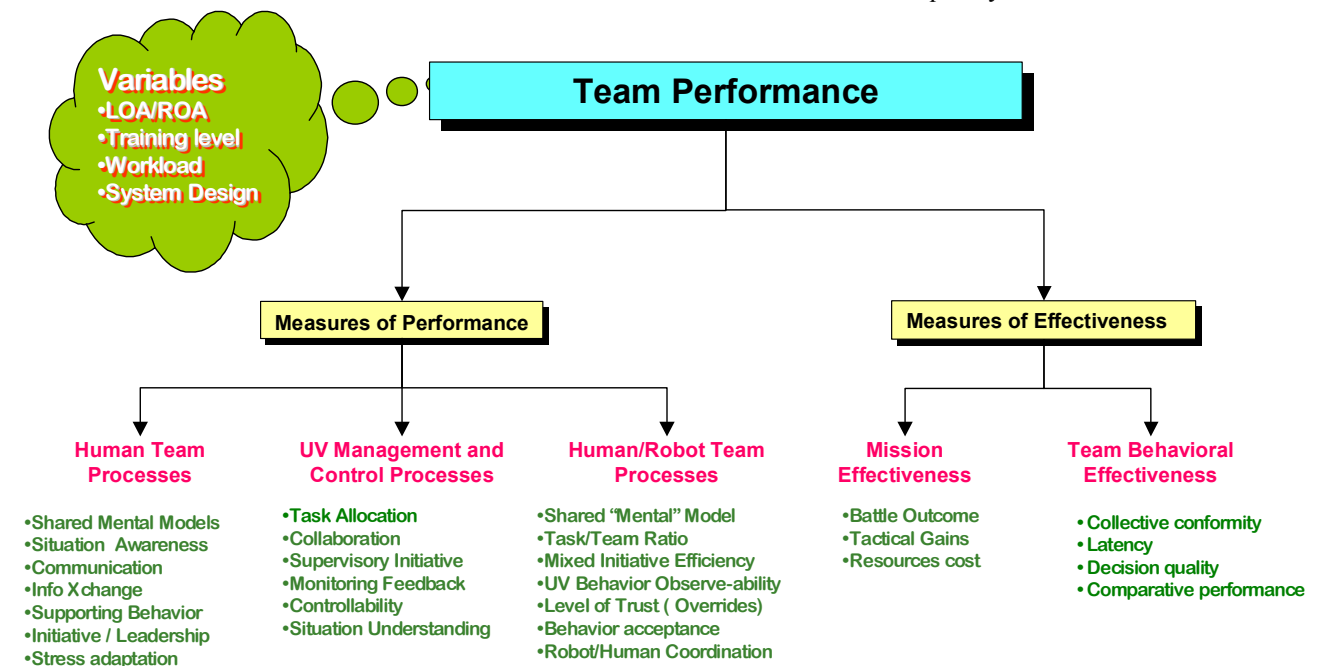


Figure 1 Collaborative Performance Model

The Performance Model has adapted and combined theories and concepts in these areas to processes associated with manned/unmanned team performance

of teamwork as was found in previous investigations [15]. The final Performance Model shown in Figure 1 consists of a multi-dimensional task process performance schema which aggregates the performance factors at each level.

The schema serves as a multi-attribute discriminate function to determine an overall level of proficiency as well as a “pass-fail” score. In practice, the weights of the attributes are established by a process of simulation experiments in which the linkage between specific task performance measures and outcomes can be estimated.

We have adopted two types of measures for the various elements of the Performance Model:

1. Measures of Performance (MOP), which are observable and derived measures of the operators’ task skills, strategies, steps or procedures used to accomplish the task. They consist of the cognitive and interactive processes of the individual and team in collaborating together and controlling the robotic entities in a coordinate manner.
2. Measures of Effectiveness (MOE), which measure the “goodness” of the composed behavior in the quality and execution of the war-fighting tasks. MOEs are influenced by more than human performance alone. These measures also contain variance accounted for by system design, the surrounding environment and luck (Smith-Jets et al 2000).

Our preliminary experimental studies, reported in DeVisser et al [4] showed that the element of trust in human-robot collaboration is an extremely important factor in the Performance Model, and accordingly we have focused much of our attention on deriving suitable and practical measures for the trust variable.

3. HUMAN-ROBOT TRUST

Trust of humans in robots’ autonomous decision capabilities is a known major issue that significantly impacts the effectiveness of human-robot collaboration, particularly in the willingness to share and allocate tasks as well as to exchange of information and create an impetus for supportive behavior.

Trust in automation is a multi-dimensional construct that is influenced by the human as well as the automated unit. Measurement of the level of trust of the humans is a major factor in the ability to train operators to develop advanced skills in human-robot collaborative activities. Training can facilitate this process by reducing initial biases, providing knowledge about system capability, and applying a risk assessment based on the behavior of the automation.

That is, mixed initiative teams will become truly successful only if humans know how to appropriately trust and hence appropriately rely on the automation. According to Lee and Moray [9], trust is “the attitude that an agent will help achieve an individual’s goals in a

situation characterized by uncertainty and vulnerability.” Humans can use trust to direct their reliance on the agent that could be an autonomous robot, for instance. They determine their trust by observing the characteristics of the system, such as its performance (how well does it accomplish the individual’s goals) and the manner in which the process of accomplishing goals is transparent [9].

3.1 Previous Research on Automation

Trust in automation is not a new concept and has been researched thoroughly [9][12]. Similar to the significance of trust in human-human teams, trust in human-robot or *mixed initiative* teams is a key factor in determining the success of such a team. However, a mixed initiative team creates unique challenges for developing trust because the robotic element may not be able to convey the necessary cues to develop trust as humans would among one another.

Lee and See [11] propose that humans develop trust in a combination of three interplaying processes: analytic, analog, and affective methods. The analytic process is a rational approach to trust and assumes that when humans make decisions in uncertainty they use a cost-benefit analysis to determine appropriate trust for the system. If, for example, a robot makes a potentially costly error the operator may be less trusting and take over control to avoid future automation mistakes. Analytic methods for trust involve categorizing observed characteristics of the system and generalize them to a broader set of assumptions about a certain group. Applied to the mixed initiative team, such assumptions could be formed based on experiences with a certain type of robot that can do a particular task well.

Finally, trust also forms purely affectively. Emotions even supersede rational thinking and therefore can play an important role in relying on automation.

From these descriptions it becomes clear that trust is a complicated and multidimensional construct influenced by types of information received by humans and their approaches to developing and determining trust as well as external influences such as system capability and reliability. Some of these factors are based on biases and inexperience of the operator and therefore it may be possible to train humans on how to appropriately trust the automation leading to improvements in appropriate reliance of the automation.

3.2 Trust and Self-Confidence

According to the definition of trust described above, trust in automation refers to the trust in the capability of

a system to accomplish an individual's goals. Trust in oneself, often referred to as self-confidence, is trust in the capability of completing the task yourself.

Various studies have subjectively measured trust in the system and self-confidence of the operator and found that they correlate [10]. As self-confidence goes down, trust in automation goes up, resulting in an increased use of the automation. When self-confidence goes up, trust in automation goes down, resulting in a decreased use of automation. Although this relationship was significantly correlated, some operators deviated from this trend by engaging in manual control even though trust in automation exceeded self-confidence. The authors explain this behavior as 'exploratory' because operators could be trying different strategies for optimizing system performance. Indeed, individual differences between operators influences trust in ways that may not be related to the characteristics of automation [11]. This reasoning is important to keep in mind when measuring trust and self-confidence subjectively.

A second model describes the relationship between trust and automation and trust in the human or self slightly differently [17]. In this paradigm, dynamic task allocation can be facilitated by suggesting when humans are preferred over automation and vice versa based on determining the levels of satisfactory trust in each agent. This approach could be used for training operators to apply appropriate trust in specific situations in which costs and uncertainty play a crucial role in determining success. Unfortunately, this theory is not backed with any data, however their intentions to conduct an experiment varying the reliability of the machine, difficulty of the task, levels of autonomy, costs, and time pressure sound promising.

3.3 Trust Over Time

Pilots who use automation frequently tend to trust automation more than students (Riley, 1996). Trust seems to change over time and this could possibly be measured across trials in an experiment. In addition, initial trust is an important factor to consider. Some studies investigating use of automated decision aids found that levels of trust will not be as high for initial low reliability, than for initial high reliability [7]. For training purposes, operators should experience an intermediate level of reliability prior to training with extreme high/low levels to prevent any biases from occurring.

3.4 The Objective Measurements of Trust

Creation of an objective measure of trust in collaborative human-robot tasks is a challenging problem. The main issue is how to represent mathematically the concept of trust and measure the necessary parameters for computing a single score. Our preliminary experiments used subjective measures of trust adapted from a common trust and self-confidence measure used in earlier automation research [9]. Following is a suggested objective score based on the measurements of observed operator and robot behavior.

The basic hypothesis is that "rational" trust behavior is reflected by the expected value of the decisions whether to allocate control to the robots on the basis of past robot behavior and the risk associated with autonomous robot control. The use of rational decision models for human-automation task allocation was proposed earlier using an expected-value analysis similar to a cost-benefit approach to improve diagnostic decisions (Swets et al 2000). Sheridan and Parasuraman (2000) propose several equations that can help improve decisions about allocating control to the human or the automation.

Using a rational decision model it is possible to establish a compound "goodness" score that collectively transforms the observed human task allocation decision behavior, risk and observed robot performance into an a relative expected loss score CEL as shown in equation (1):

$$(1) \quad REL = \frac{\sum_{i=1}^{i=n} el_i}{K}$$

Where el_i is the expected loss of robot autonomous control at trial i ; el_i can be computed as

$$(2) \quad el_i = P_i * C_i$$

and

n is the number of trials

K is the number of operator overrides or intervention to take over control from the robot

The Relative Expected Loss can be written as

$$(3) \quad REL = \frac{\sum_{i=1}^{i=n} P_i * C_i}{K}$$

This objective measure of trust was tested empirically using the MITPAS Simulation Environment described below. MITPAS (for Mixed-Initiative Team Performance Assessment System) has been developed under a SBIR Phase I project sponsored by US Army RDECOM STTC.

4. MITPAS SIMULATION

MITPAS is a new simulation environment for performing experiments that will measure and assess the performance of mixed human-robot teams in a variety of military and non-military situations. The total environment combines several novel components:

1. A Mixed Initiative Team Performance Assessment System (MITPAS) consisting of a methodology, tools and procedures to measure the performance of mixed manned and unmanned teams in both training and real world operational environments. It is directed toward supporting the operational and training needs of future military forces that will use mixed manned and unmanned forces for a broad variety of functions.
2. A simulated robotic command and control system based on the U.S. Army's OneSAF simulation. OneSAF is a generalized set of models and tools designed to supply semi-autonomous entities to a variety of simulations.
3. An event-based test scenario characteristic of anticipated military operations. Our event-based scenario reflects mid-term plans for DOD's Future Combat System (FCS) as well as current uses of robotic systems in combat.

The purpose of the new simulation environment is to provide a self-contained, easy to use test bed for obtaining meaningful measures of both human and unmanned system performance to be used to test proposed mixed initiative configurations and to identify new training requirements.

5. EXPERIMENTAL STUDIES

We studied collaborative human-robot behavior in a scenario that took place at a simulated location with typical tactical features of roads, forests and a small village or built-up area, shown in Figure 2.

The operator controlled an unmanned ground vehicle (UGV) as part of a reconnaissance platoon, whose mission was to ensure that a route through the area was safe for passage by eliminating all surrounding enemies. To do this, the operator had to move the UGV to a checkpoint where it could commence targeting an enemy, firing on the targeted enemy, and repeating this process for each of 6 attacking enemies. Operators also had to monitor and evaluate the autonomous targeting and firing capabilities of the UGV and take over control of the vehicle if these autonomous processes would cause a delay or fail completely.

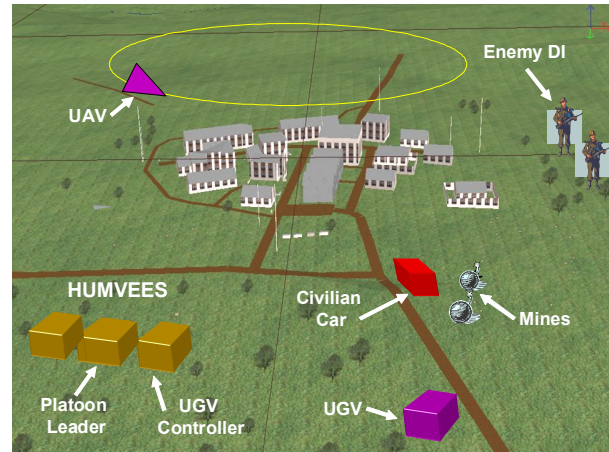


Figure 2 MITPAS Experimental Environment

The UGV could be set to one of three levels of targeting and firing competency – low, medium and high. Operators were trained to recognize the distinct symptoms associated with competency failures in autonomous targeting and UGV firing. Operators could improve targeting by either manually selecting an enemy on the battlefield or rotating the UGV manually to face the enemy. Firing competency could also be enhanced by moving the UGV closer to the enemy. Participants were instructed to perform an override only when they thought the overall mission time would decrease as a result of their intervention. A secondary task included reporting each enemy kill, target overrides, and check point arrivals to the Battlemaster. Operators completed the mission by arriving at a second check point.

The command and control configuration for our simulation environment and experimental study was as shown in Figure 3. The Battlemaster, who also played the platoon leader, was in charge of the experimental procedures, the progress of the scenario, and communication with the Unmanned Ground Vehicle controller, who was the actual experimental participant. The UAV controller was a virtual participant in the current scenario.

The command and control stations used by the Battlemaster and the UGV controller were very similar. This station as shown in Figure 4 is comprised of a tactical situation map, a 3D simulation of the environment, and a UGV status and communication display represented on three 19 inch monitors.

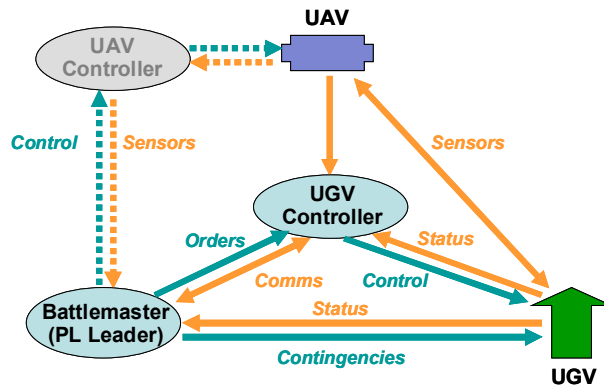


Figure 3 Command and Control Configuration

The operator used a keyboard, track-ball mouse, and joystick to control the system. The tactical situation map, generated by the military simulator OneSAF 2.5 on the center monitor, showed the UGV's geographic position and also gave an overview of the entire tactical situation.

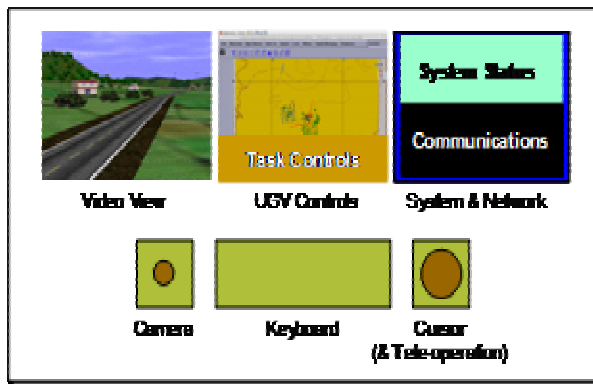


Figure 4 UGV Controller Station

The UGV control interface and the map control features enabled the operator to assign the UGV specific tasks within the mapped area. A 3D representation of this environment, simulated by MAK Stealth 5.4 display on the left screen, could show live action from the viewpoint of a camera mounted on the UGV. The operator could tele-operate the UGV using a joystick in combination with gas and brake pedals.

The right monitor showed a java-coded UGV status and communications display. The status display provided real-time information to the operator about the UGV, including its direction, current target, surrounding friendly and enemy units, supply levels, and any vehicle malfunctions. Communications between the operator and the Battlemaster were facilitated through instant-message capabilities.

Twelve young adults (4 females and 8 males) varying in age from 18 to 25 served as UGV Controllers. Most participants had several years of gaming experience. Participants first received about 1.5 hours of training on proper use of the UGV Controller Station by completing exercises using a training manual.

Participants then completed 5 trials of 3 levels of UGV Competency in Firing behavior (low, medium and high) totaling 15 trials. Fifteen unique scenarios were created varying the locations of the enemy to ensure that participants built up unique situation awareness during each mission. After completion of each trial, participants filled out the NASA Task Load Index (TLX) for subjective workload and an adapted version of a common subjective trust and self-confidence measure used in earlier automation research (Lee & Moray, 1992).

5.1 Initial Experimental Results

The experimental results supported our initial hypotheses based on the literature that: (1) overall mission time would increase as robot competency decreased, (2) the operators would intervene with more manual control as robot competency decreased, and (3) initially high workload in the low competency condition decreased as the trials progressed, reaching the levels of workload in the medium and high competency conditions as participants became more capable of handling the unreliable UV in manual control.

At the same time, a number of the results were less expected; these included:

1. *Operators Will Compensate For Lack Of UGV Competence, But Only If The UGV Competence Level Is Clearly Identifiable.* Our results for Mission Time show that the operator learns to compensate partially for low competence of the UGV, but there is little or no adjustment for medium competence; it seems that if the UGV is indeterminably competent, the operator has more difficulty adjusting his or her behavior to that of the semi-autonomous machine. Essentially the same results were seen for Kill Latency (the total time required for the human-robot team to destroy the target).
2. *First Impressions Matter.* The results indicated that for UGVs of either high or medium competence, the amount of manual control is less if the operator first experiences a UGV of high or medium competence rather than one of low competence. Similarly, for UGVs of low competence, the amount of manual control is less if the operator first experiences a UGV of low

competence. We conclude that it is important to train operators on a UGV with similar characteristics to the one he or she will be operating.

5.2 Follow-Up Study

We also performed a less formal study to test the utility of our newly defined objective measure of trust. This

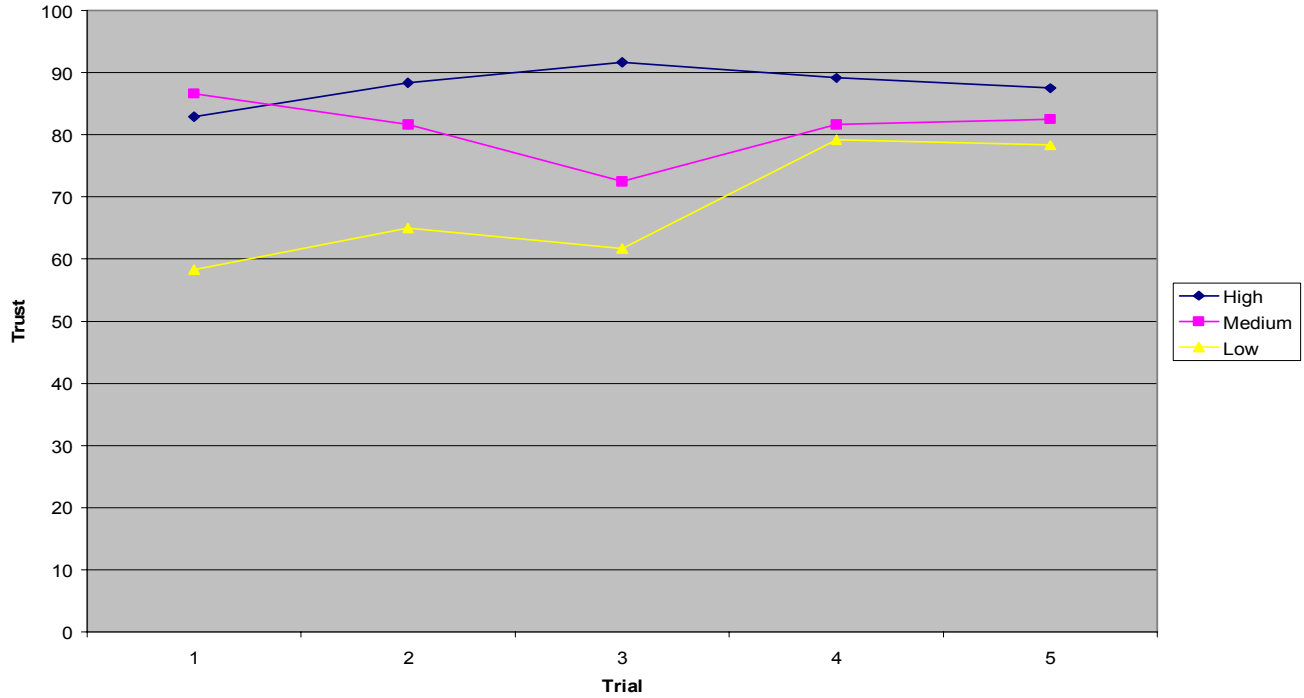


Figure 5 Subjective Trust as a Function of Trial Number for Three Robot Competency Levels

3. *Familiarity Breeds Understanding.* Our results show that overall, subjective trust correlates with the overall competency of the UGV as might be expected. But one result should make us reassess what “trust” actually means to UV operators. As shown in Figure 7, our results indicate that subjective trust increases about 35% from 58 to 78 over 5 trials for the low competence UGV, while for the medium and high competence robots trust remains relatively unchanged at levels of 85 and 90, respectively. It appears that trust concerns not just an expectation of *correct performance*, but also an expectation of *level of performance*. Once the operator knows that the UGV will make mistakes, he or she *trusts* it to make those mistakes. Because the operator of the medium competency level UGV cannot seem to get a good handle on what it will do, the subjective Trust level actually *decreases* slightly over 5 trials.

study used one subject controlling over a wide variety of robot conditions in approximately the same scenario as before. Figure 6 shows the expected loss for a trial computed by Equation (2) on the x axis and the override score for that trial on the y axis.

The graphed results for individual trials separate clearly into three behavioral clusters, which we have termed ‘over-trust,’ ‘under-trust’ and ‘proper-trust.’ These

1. An over-trusting operator overrides the robot a low number of times, even though the expected loss is high.
2. An under-trusting operator overrides the robot a large number of times, even though expected loss is low.
3. Proper-trusting of an operator is exhibited in the graph by the operator who overrides the robot a low number of times, when the expected loss is low and a large number of times when the expected loss is high.

The analytical methodology of Figure 6 can be used to diagnose specific operator trust behavior for the purpose of training and performance evaluation. The criteria for rational or “proper” behavior are represented by the oval region that contains the region in which expected loss should invoke an override. This region can be defined by exercising an expert with the system and obtaining the cluster of his behavior. This cluster can then be used as a standard for scoring operators.

$$(7) \quad TS_x = \sum_{i=1}^{i=N} |\phi_{ei} - \phi_{xi}|$$

Computation of $[\phi_{ei} - \phi_{xi}]$ can be performed by calculating the euclidean distance between ϕ_{ei} and ϕ_{xi} . Equation (7) can thus be written as:

$$(8) \quad TS_x = \sum_{i=1}^{i=N} \sqrt{\{ [eli_e - eli_x]^2 + [osi_e - osi_x]^2 \}}$$

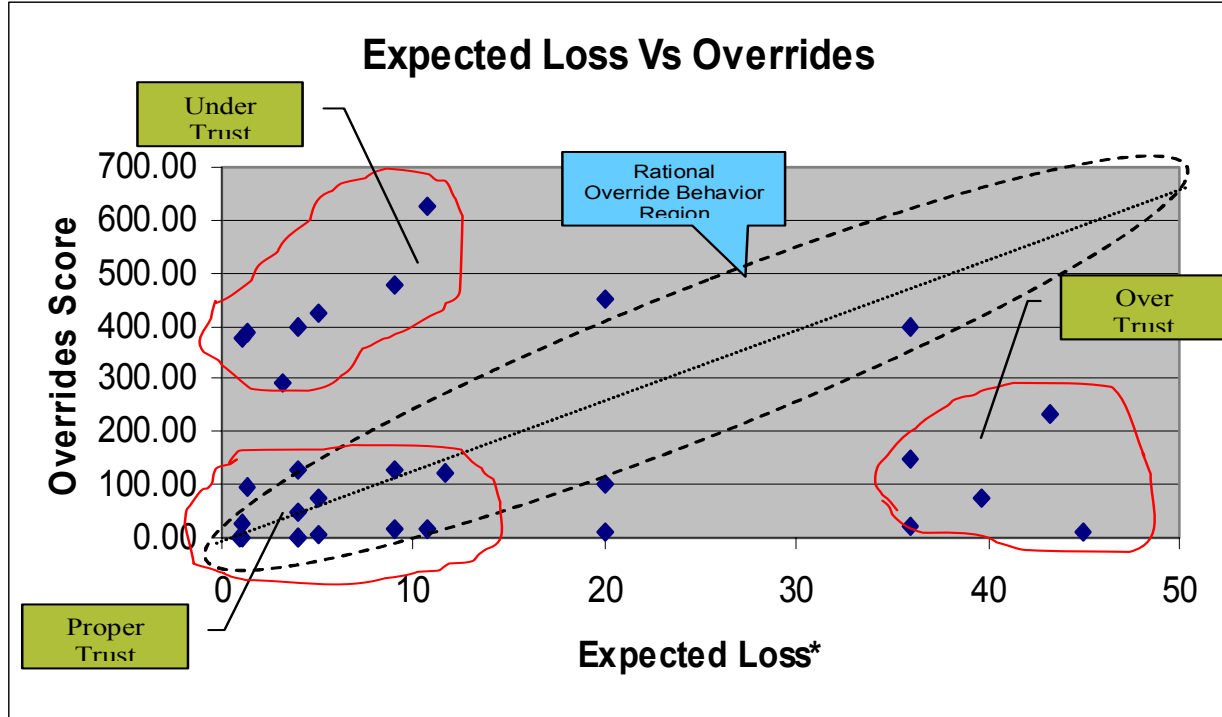


Figure 6 Analysis of Objective Trust Behavior

The cluster of specific operator location on the chart provides diagnostic value as to the characteristics of his behavior. As shown in figure 6 operators can be characterized as either over-trusting or under-trusting. To generate a measure of goodness for a specific operator, computational criteria can be created that measure his deviation from an expert.

This measure can be computed as follows: Let the Pattern Expected Loss vs Overrides in the Euclidean space in Figure 6 be:

$$(4) \quad \phi_e = [\phi_{e1}, \phi_{e2}, \dots, \phi_{en}]$$

and let the Pattern of Expected losses vs Overrides of operator x be

$$(5) \quad \phi_x = [\phi_{x1}, \phi_{x2}, \dots, \phi_{xn}]$$

across a set of events I, where

$$(6) \quad I = [i_1, i_2, \dots, i_N]$$

A Trust performance score TS_x can be computed as

$$i=1$$

Where osi_e is the experts override score and osi_x is the operator's override score

The score TS_x can be used to rate the operator skill in relation to an expert. By providing a quantitative measure of the “distance” between the operator relative expected loss and that of an expert, the graphic representation in figure 6 provides direct indication for the type of feedback the operator needs in training in order to modify his trust behavior.

6. CONCLUSIONS

We believe these early promising results are indicative of (1) the benefits that can be achieved from both subjective and objective measures of trust in human-robot collaboration, and (2) the type of new insights into human-robot team behavior that can be gained by combining the power of a measurement system like MITPAS with realistic simulations of tactical UV operations such as those represented in our OneSAF-

based experimental environment. Further empirical studies are needed to further validate and refine our objectives more completely. To encourage such studies, we plan to make the easily-customized MITPAS measurement system and simulation environment available to other researchers and developers for use with their human-robot simulations and scenarios.

ACKNOWLEDGEMENT

This research was supported by SBIR Phase II contract No. N61339-05-C-0003 funded by the U.S. Army RDECOM-STTC and administered in part by U.S. Army Research Institute, Orlando.

REFERENCES

- [1] Albus, J.S. "Metrics and performance: Measures for intelligent unmanned ground vehicles," In MIS Proceedings.
- [2] Barnes, M., Parasuraman, R. & Cosenzo, K. (In press). Adaptive automation for robotic military systems, Technical Report, NATO HFM, 2002
- [3] Clough, B.T. "Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway?" AFB in MIS Proceedings. Air Force Research Lab, Wright Patterson, 2002
- [4] DeVisser, E., Freedy, A., Freedy, E., Weltman, G., & Parasuraman, R. "A comprehensive methodology for assessing human-robot team" performance for use in training and simulation," In Proceedings of the 50th Human Factors and Ergonomics Society Conference, (In press).
- [5] Drewes, P. "Lessons learned in group robotic command and control," Unmanned Systems Conference, Orlando, FL, 2002
- [6] Endsley, M.R. "Direct measurement of Situation Awareness: Validity and use of SAGAT," In Endsley, M. R., Garland, D. J. (Eds.) Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum Associates, 2003
- [7] Fallon, C. K. "Improving user trust with a likelihood alarm display," Proceedings of the 1st Conference on Augmented Cognition, Las Vegas, NV, 2005
- [8] Freedy, A., McDonough, J.G., Freedy, E.T., Jacobs, R., Thayer, S.M., and Weltman, G. "A mixed initiative team performance assessment system (MITPAS) for use in training and operational environments," SBIR Phase I Final Report, Contract No. N61339-04-C-0020, Perceptronics Solutions, 2004
- [9] Lee, J. D., & Moray, N. "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, 35, 1992, pp 1243-1270.
- [10] Lee, J. D., & Moray, N. "Trust, self-confidence, and adaptation to automation," *International Journal of Human-Computer Studies*, 40, 1994, pp 153-184.
- [11] Lee, J. D., & See, K. A. "Trust in automation: Designing for appropriate reliance." *Human Factors*, 46(1), 2004, pp 50-80.
- [12] Parasuraman, R., & Riley, V. "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, 39(2), 1997, pp 230-253.
- [13] Parasuraman, R., Galster, S., Squire, P., Furukawa, H., & Miller, C. "A flexible delegation-type interface enhances system performance in human supervision of multiple robots: Empirical studies with RoboFlag," *IEEE Transactions on Systems, Man, and Cybernetics. Part A. Systems and Humans*, 35, 2005, pp 481-493.
- [14] Rovira, E., McGarry, K., Parasuraman, R. "Effects of imperfect automation on decision-making in a simulated command and control task," *Human Factors*, (In press)
- [15] Sheridan, T., & Parasuraman, R. "Human-automation interaction," *Reviews of Human Factors and Ergonomics*, 1, 2006, 89-129.
- [16] Smith- Jentsch, K.A, Johnson, J.H. and Payne, S.C. "Measuring team related expertise in complex environments," In J.A. Cannon-Bowers and E. Salas (Eds.), *Making Decisions Under Stress*. American Psychological Association, WDC, 1998.
- [17] Van Dongen, K., & Van Maanen, P. "Designing for dynamic task allocation," Paper presented at the 7th International NDM Conference, Amsterdam, The Netherlands, 2005