



**Human Factors Division
Institute of Aviation**

**University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874**

**Human Machine Interface Analysis of
Unmanned Vehicle Systems**

**Christopher D. Wickens
& Stephen R. Dixon**

**Final Technical Report
HFD-06-1/MAAD-06-1**

January 2006

**Micro Analysis and Design, Inc.
Boulder CO**

Contract ARMY MAD 6021.000-01

Human Machine Interface Analysis of Unmanned Vehicle Systems

Final Report

Christopher D. Wickens and Stephen R. Dixon

Abstract

The results of eight experiments and a meta-analysis are summarized that address three general areas of human-system interaction within the context of a UAV simulation: (1) In examining and modeling the effects of different workload mitigation techniques, regarding automation and multiple resources, we conclude that single channel theory is inadequate to fully model these effects, and that one pilot flying two independent UAVs will sacrifice performance on secondary surveillance tasks, even with perfect automation. (2) In examining the consequences of **imperfect** automation, we find that such imperfection (unreliability) effects have differential impact on performance depending on the nature of the automation, being less severe for an imperfect autopilot than for imperfect diagnosis. We reach a general conclusion that imperfect automation is acceptable, but that reliabilities less than 0.75 are worse than no automation at all. (3) We examine the consequences of adjusting the alert threshold in diagnostic automation, trading off automation misses for false alarms, and interpret our data in terms of constructs of automation reliance and compliance. Based on the data from these experiments, the preliminary foundation of a computational model of UAV workload is provided.

1. Introduction

Unmanned air vehicles such as the Army's Hunter and Shadow have contributed substantially to supporting mission effectiveness in recent operations with their surveillance capabilities. Future operations will undoubtedly require increased use of these and other UAV assets. However such needs may encounter the constraints of human personnel to supervise the UAV, given currently the requirement for two soldiers to coordinate the in-flight operation of a single UAV.

Our research effort at Illinois has focused on the strategies for reducing the manpower requirements of UAV supervision in a Hunter/Shadow type simulation, from a 2:1 ratio of soldiers to assets, to a 1:1, and 1:2 ratio. Such strategies require the consideration of two important human performance concepts, related to workload and automation dependence. Regarding workload, the effort to assign tasks normally associated with multiple-operators (2:1), to a single operator (1:1), and then to double the number of tasks (1:2), can potentially overload the operators limited processing resources, leaving performance on certain tasks vulnerable (Wickens & Hollands, 2000). In other systems, such as the commercial airliner cockpit, such potential workload increases associated with downsizing (from 3 to 2) have been offset by automation to replace the activities of one of the missing participants (in this case, the flight engineer on the Boeing 737). However a long history of research in human-automation interaction (e.g., Sheridan, 2002, Parasuraman & Riley, 1997, Parasuraman, Sheridan & Wickens, 2000), has revealed that automation may not entirely eliminate human cognitive demands, to the extent that such automation will require both set up and supervision. The latter

task of automation supervision is particularly relevant to the extent that the automation may be **imperfect** (e.g., less than 100% reliable).

In typical UAV operations, the sources of such imperfections are manifold. As two simple examples, an automatic target recognition (ATR) device to aid surveillance will undoubtedly make some misclassifications, if it is only provided low resolution imagery to work with; or an autopilot may become challenged to hold a precise course, if icing or severe turbulence disrupts the handling qualities. As we discuss below, the impact on soldier workload of supervising UAV automation depends critically upon the level of reliability of the automation, as well as the qualitative kinds of failures that may occur.

In seven experiments on the UAV Hunter/Shadow simulation at Illinois, along with another experiment based on the SIL simulation, we have examined the workload effects of imperfect automation, and have also attempted to develop a computational model of these effects. Such models, if valid, are of considerable importance as they may be used to make predictions of soldier capabilities in the absence of time-consuming human-in-the-loop simulation data.

2. The General UAV Simulation

Figure 1 presents the interface used by our pilots to fly the UAV Hunter/shadow simulation.

In the simulation, pilots were responsible for

- (1) a primary mission task, in which they tracked the UAV to waypoints and reported on “command targets” (CT) at those known coordinates by reference to the navigational display in the lower left. They could refresh their memory for command target information as required by depressing a “recall key”. Tracking was normally accomplished with a rotational heading control necessary to compensate for periodic disturbances. (Altitude and airspeed were fixed).
- (2) a secondary surveillance task in which they searched for “targets of opportunity” (TOO) at unknown locations while en route. Unlike the CTs, these TOOs were camouflaged, and very difficult to see as they passed through the 3D image window (upper left) while the UAV flew overhead.
- (3) a secondary systems monitoring task, in which they were required to detect and respond to on board system failures (SF). This required monitoring of four gauges that slowly oscillated, with one occasionally crossing into a danger zone. When this occurred, the pilot was required to detect it (with a button press), and enter certain critical digital information related to diagnosis and current location.

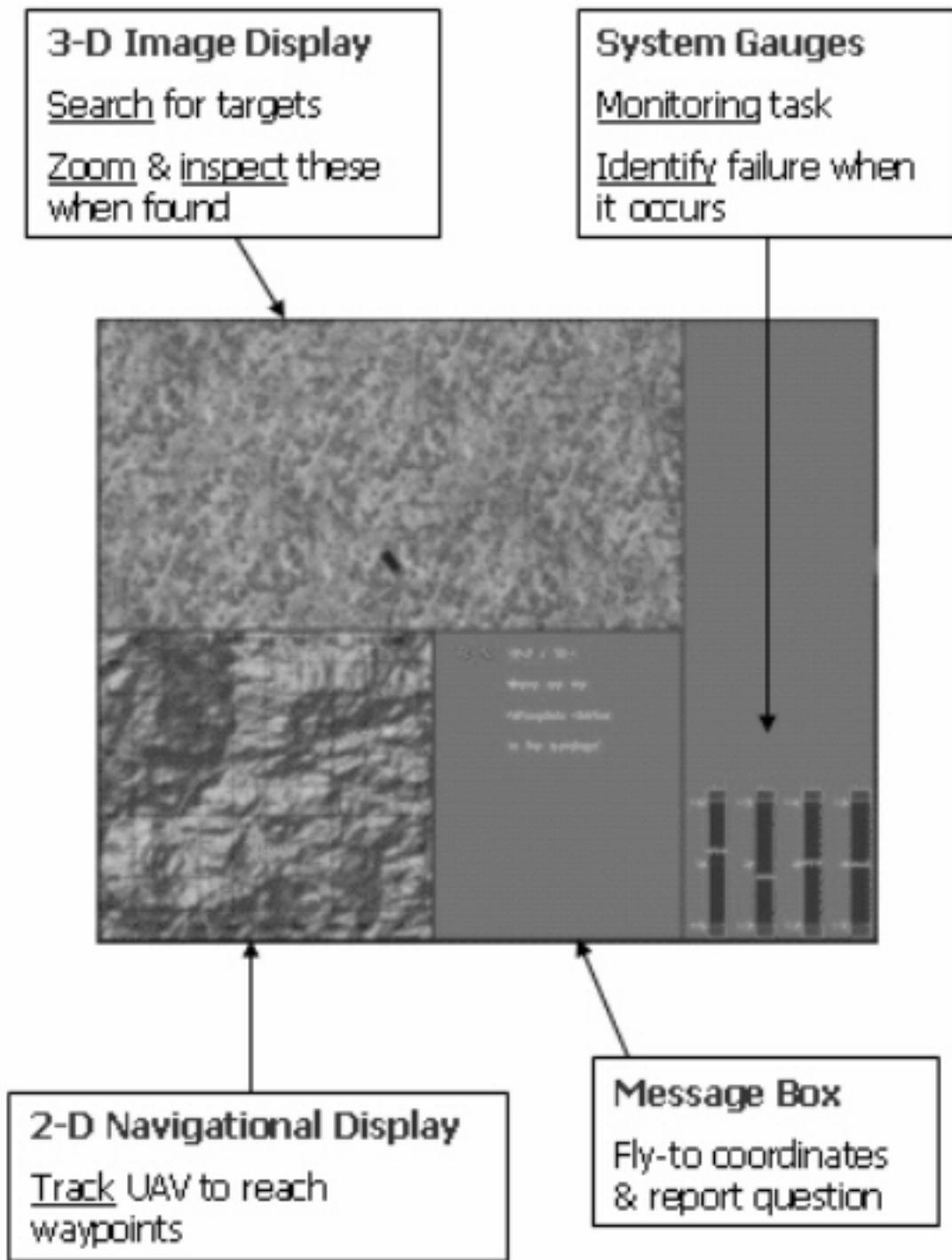


Figure 1. Illinois UAV simulation display.

Once either a TOO or a CT was found, the pilot was required to enter a ‘loiter mode’ by depressing a key which brought the UAV to a race-track pattern around the target. They were then required to zoom in their camera image, and orient it so as to keep the target in view, and report properties of the target (e.g., what side of the bunker contained tanks, how many helicopters were present at the target). The combined operations in these tasks imposed very high workload on the pilots. In the following, we summarize the results of the five experiments and one meta-analysis of the literature addressing three general categories of effects: pilot workload mitigation and modeling, effects of imperfect automation.

The majority of participants in our simulation were student pilots nearing completion of their private pilots’ certificate course. Each participant typically flew some practice legs, followed by three scenarios each consisting of ten legs (each leg defined by a trajectory to a single CT, overflying a TOO, and possibly encountering one or more system failures).

3. Workload Mitigation and Modeling: Experiments 1 and 2

In experiment 1 (Wickens & Dixon, 2002), employing the general UAV paradigm described above, we examined two techniques for mitigating the high “triple task” workload of controlling a single UAV. An *autopilot* mitigation essentially replaced the entire task of flying the UAV (monitoring and controlling the heading trajectory) with a perfectly reliable autopilot, requiring only keyboard entry of the coordinates of the next CT. An *auto alert* mitigation provided an auditory warning whenever one of the system gauges passed its critical level. Because we were also particularly interested in the role of the auditory modality in offloading the heavy visual monitoring load of the UAV task, we also switched a second task to auditory presentation: the display of the command target information. Distributing information delivery between auditory and visual channels is a mitigation that is predicted by multiple resource theory (MRT) to improve time sharing ability (Wickens, 2002). However it is sometimes asserted that such benefits, while found in the laboratory, may not be apparent in more real world environments in which people tend to become single channel processors of information. That is, it is a central processing bottleneck, rather than limits of sensory-perceptual mechanisms that constrain multiple task performance (Liao & Moray, 1993). Thus the MRT-predicted benefits of an auditory offload were contrasted with the single channel theory prediction of no benefits.

The results of experiment 1 (Wickens & Dixon, 2002) clearly indicated significant benefits of both types of mitigation. Autopilot automation removed a task and improved time sharing by reducing overall demands. Auditory offload re-distributed resource demands across modalities, and enabled better time sharing by capitalizing on multiple resources. However because of the categorical differences between the two (the autopilot involved total elimination of one task, the auto-alert involved elimination of the visual monitoring component of one task, and the visual reading of another), it was difficult to draw any scientifically meaningful conclusion regarding the relative benefits of the two forms.

In experiment 2 (Wickens, Dixon & Chang, 2003; Dixon, Wickens, & Chang, 2005) we extended the two mitigation techniques to a two UAV simulation, in which the pilot sat in front of two workstations. The two UAVs were independent of each other. In different conditions, pilots could either control one or two UAVs, and in each of these conditions, the UAVs could be in a baseline configuration, or in either of the two workload mitigation configurations (auditory

or autopilot). In all dual UAV conditions, the mitigation configuration of the two UAVs was the same (both autopilot or both auditory).

The results were evaluated from both an overall performance and a fine grained attentional modeling perspective. Regarding overall performance the results suggested that the two mitigation techniques were somewhat successful in buffering the dual UAV workload costs on primary mission completion and system monitoring, but that they failed to provide any protection for the TOO surveillance task. That is, flying two UAVs, even with mitigation, will allow only mission critical tasks to be performed, but would sacrifice the capability of the single pilot to perform effective en route surveillance. And even this assumes perfect automation, the issue of imperfection to be considered in the following section.

Our analysis of attention models focused on the viability of three classic models of multiple task performance to account for variance in performance between the different conditions, defined by the three different UAV configurations (baseline, auditory, autopilot) invoked in single and dual UAV control, during low and high workload periods of flight. The **single channel model** (Welford, 1968; Liao & Moray, 1993), briefly considered in the context of Experiment 1, assumes the pilot capable of only one task at a time. If a second task is imposed while a first is ongoing, the latter must be fully delayed till the former is completed. The **single resource model** predicts that concurrent performance (parallel processing) is possible to the extent that the demand level of each time shared task is reduced, as if all are drawing from a single pool of mental resources of limited capacity (Kahneman, 1973; Sarno & Wickens, 1995). The **multiple resource model**, as described earlier, assumes that demands may also be offloaded by distributing tasks across resources. Each of these models may be invoked as part of more complex performance models such as those embodied in IMPRINT (Laughery & Corker, 1997), so that testing their relative viability in accounting for data in a realistic simulation is important.

We examined the viability of the **single channel model** to account for the data through two tests (see Wickens, Dixon & Chang, 2003 for details). In the “**summing test**” we predicted the time it would take to perform two tasks on different UAV workstations by summing the single task times. If the actual time to complete both, when one task arrives before the other task is completed, is equal to the sum of the single task times, then the single channel assumptions are upheld. If the sum is greater, then not only is single channel theory upheld, but an added penalty for attention switching between workstations is manifest. If, on the other hand, the actual completion time is less than the sum of single task times, it implies some degree of parallel processing, consistent with (single or multiple) resource sharing. Our analysis revealed that dual UAV control in the baseline condition was well modeled by the single channel model plus a substantial cost for switching between tasks. In the autopilot condition, the data were well modeled simply by single channel theory, without switching cost. (Note that this is also consistent with a resource sharing plus switching cost model, where the sharing benefits and switching costs offset each other). Most importantly, the auditory mitigation condition provided evidence for resource sharing. The actual completion time was substantially (4-5 sec) less than that predicted by a single channel model.

In the second test of single channel theory, the **arrival time** test applied the classic procedures developed from single channel models (Welford, 1968; Keele, 1973), whereby the completion time of a second arriving task is predicted to be a linearly increasing function of how

soon it arrives following the initiation of the first arriving task. That is, every second earlier that the second task arrives, adds one second longer that it has to wait before it has access to the pilots' single channel information processor, and hence adds one second for its total completion. We selected all instances when a task (TO or CT inspection, system failure report) on one UAV arrived while a task on the other UAV was ongoing; we computed the inter-arrival time, and then plotted this time against the completion time of the second arriving task. Importantly, this analysis revealed no linear relationship between arrival time and completion time, again casting doubts on the viability of a pure single channel theory to account for all aspects of the data.

In experiment 2, we examined the viability of a **single resource model** to account for the data by the following procedure. First, we obtained independent estimates of the demand level of each of the single task components, in each of their different manifestations (e.g., demand level of system monitoring, or TOO monitoring, demand level of diagnosing system failures, of zooming and reporting TOOs, etc.). Using a time line analysis, we then identified the different intervals of time during which all possible combinations of tasks occurred. During each of these intervals we then predicted the amount of workload the pilot would experience under a single resource model by summing the single task demand values (Sarno & Wickens, 1995; Wickens, 2002). These demand values were then correlated with the actual performance level of the various subtasks, measured within the relevant time interval. These correlations revealed that the single resource model did a modest job in predicting performance for TOO surveillance performance $r = 0.25$ (detection time) and $r = 0.57$ (detection miss rate); for system failure monitoring, $r = 0.37$ (detection time) and $r = 0.41$ (detection accuracy). However for detecting target trajectory tracking performance (in the baseline and auditory conditions when this was not automated) the correlation was actually in the opposite direction ($r = -0.44$) such that higher predicted workload was associated with reduced error.

We then considered the extent to which some assumptions of **multiple resource model** might improve the fit. Rather than invoking a full fledged multiple resource model (see Sarno & Wickens, 1995; Wickens, 2002; Horrey & Wickens, 2003), we instead provided one very plausible augmentation to the single resource model predictions described in the preceding paragraph. Whenever the auditory supported system monitoring task competed with another task in single UAV performance, the workload penalty (computed by the sum of demands) was reduced by 2.0 (relative to visual system monitoring). Whenever they competed within a dual UAV combination, the penalty was reduced by 3.0. Recomputing predicted values in this way substantially improved the prediction of system failure monitoring performance, increasing the correlation for detection time from 0.37 to 0.76, and the correlation for SF miss rate from 0.41 to 0.74. The revised computations had little effect on prediction of performance for the other tasks (tracking and TOO surveillance), suggesting that the multiple resource benefit (or cost for shared visual resources in the baseline) was realized only by the task whose modality was changed.

Thus collectively the modeling data suggest that single channel theory may be adequate under high load dual UAV conditions and when only a single perceptual resource is used, but become less sufficient when load is reduced, and when separate perceptual resources are employed (the former reflecting some of the findings by Liao & Moray, 1993). Under these circumstances assumptions of both single resource theory (demand values) and multiple resource theory (reduced penalties for separate resources) should be incorporated into performance models (see also Sarno & Wickens, 1995; Wickens, Goh, Helleberg, Horrey, & Talleur, 2003).

4. Imperfect Automation: Experiments 3 and 4 and a Meta Analysis

The evaluations of experiments 1 and 2 were, in a sense, best case scenarios in which automation functioned perfectly. In the reality of UAVs, the assumption of perfect automation is problematic. Interviews with subject matter experts (Hunter/Shadow pilots), revealed the numerous occasions of encounters with aspects of UAV supervision where the automation did not work as expected, or where on board system failures thwarted smooth mission functioning. Furthermore, UAVs are often called upon to carry out automated surveillance functions that challenge computer vision, just as they might challenge human vision. In experiment 3, we selected two qualitatively different aspects of automation that might “fail”, each linked with the two automation-based workload mitigations that were examined in experiments 1 and 2. We failed the autopilot, leading to unpredictable, but subtle “drifts” of the trajectory off the pre-selected course, and we failed the auto-alert system, in a way that parallels a long history of research on human complacency with automated diagnostic systems (e.g., Parasuraman, Molloy & Singh, 1993).

In two subsequent experiments (3& 4; Dixon & Wickens, 2003, 2004, in press) several different conditions were created; baseline (no automation), perfect automation of each aspect (with baseline in the other aspect), perfect automation of both aspects, and six conditions of imperfect automation. Of these six, one had a 70% reliable autopilot, the other five varied reliability of the system failure monitoring automation from 80% to 70% (two versions) to 60% (two versions). (The two versions varied in terms of whether the alert threshold was set to generate more misses or more false alarms, a distinction not addressed in the current chapter; see Dixon & Wickens, in press). Pilots only flew a single UAV, and the different kinds, availabilities and reliabilities of automation were varied between pilots. Prior to the experiment, each pilot was informed generally as to the level of reliability of any automated system component.

One important distinction we make here is between low and high workload. Thus system failure monitoring can be discriminated between low workload periods, when pilots are simply monitoring the flight path and the TOO image window, and high workload periods, when system failures occur while the pilot is engaged in a TOO or command target inspection (zooming and panning). Correspondingly, TOO monitoring can be discriminated between low workload periods, and those high workload periods when a TOO appears in the image window while the pilot is diagnosing and responding to a system failure.

The collective results of experiments 3 and 4 (see details in Dixon & Wickens, 2003, in press), revealed that both forms of automation were beneficial when perfect (replicating effects in experiments 1 and 2). The benefits were most realized in high workload periods. The results also revealed that unreliability of both forms degraded performance, but such degradation was significantly less for the imperfect autopilot than for the imperfect diagnostic automation, at the equivalent 70% reliability value (i.e., both were designed to “fail” on 30% of the encounters). We believe that this difference may be due to the fact that navigation performance of the UAV trajectory was viewed by our pilots as more critical to mission success (reaching the command targets), than health monitoring, and therefore pilots treated path monitoring as the “more primary” of the tasks.

Then, examining only imperfect diagnostic automation, we compared the 80%, 70%, and 60% levels with the baseline level of performance, in essence asking the question: “how poor can this automation be, before it is worse than no automation at all?” The results were clear cut: 80% was better, but 70% and 60% was worse (a finding replicated by a fifth experiment in which visual scanning was measured: Wickens et al., 2005). Generally, false-alarm prone automation was more disruptive than miss-prone automation.

Two additional conclusions were drawn from the data: (1) imperfect automation costs primarily emerged at higher workload. (2) These costs (and increasing costs with lower reliability) were borne more by the automated system monitoring task than by the two concurrent tasks (TOO monitoring and trajectory guidance). Such effects can be modeled by a resource model in which resources are allocated away from the automated task toward the mission critical “primary” task (trajectory monitoring). The former suffers the decrement of imperfection, since the pilot was not effectively monitoring the raw data (system gauges), and the costs were more amplified when resources were more scarce under high workload. This resource allocation effect was confirmed when visual scanning was measured in four of the experimental conditions (Wickens et al., 2005).

The potential implications of an approximate “reliability threshold” below which automated reliability should not fall, was investigated in the final work described here. We conducted a quasi-meta analysis (Wickens & Dixon, 2006, in press), in which the human performance data were extracted from all available studies that we could find of imperfect diagnostic automation. We only considered studies in which the humans also had perceptual access to the raw data upon which the automation made its diagnosis (e.g., in the current UAV simulations, these “raw data” are reflected by the system failure gauges). For each study we assessed the degree to which performance with the diagnostic aid was better than, equivalent to, or worse than the baseline performance, and regressed this trichotomous measure onto the actual reliability of the aid.

This regression is shown in Figure 2, and reveals the relatively stable linear relationship (correlation $r = + 0.64$) suggesting that, not surprisingly, combined human-automation diagnostic performance increases linearly with aid reliability. Three characteristics of the study however are less intuitively obvious. First, confirming the trend revealed in experiments 3 and 4, there is a point below which the availability of the aid leads to worse performance than having no aid at all. This point is at reliability $r = 0.71$ (95% confidence interval = $\pm .07$). Below this cutoff, we use the metaphor of the “concrete life preserver” to characterize the diagnostic aid. That is, the user appears to depend upon it even when better performance would be obtained if it were ignored. Second, we performed a separate regression on those studies that were carried out (like ours) within a dual task context, and found that for this subset, the linear model fit even more strongly (correlation $r = +.78$), a finding that is reasonable if we infer that automation dependence would be greater when human processing resources are made scarce by the dual task requirements. Third, when we examined performance on those **concurrent tasks** (such as the TOO monitoring in the current paradigm) rather than the automated diagnostic task, as a function of diagnostic automation reliability, the regression line was essentially flat. This finding indicates that people generally tend to treat diagnostic automated tasks as “secondary”, buffering the primary concurrent tasks from whatever resource demands are imposed by decreasing reliability.

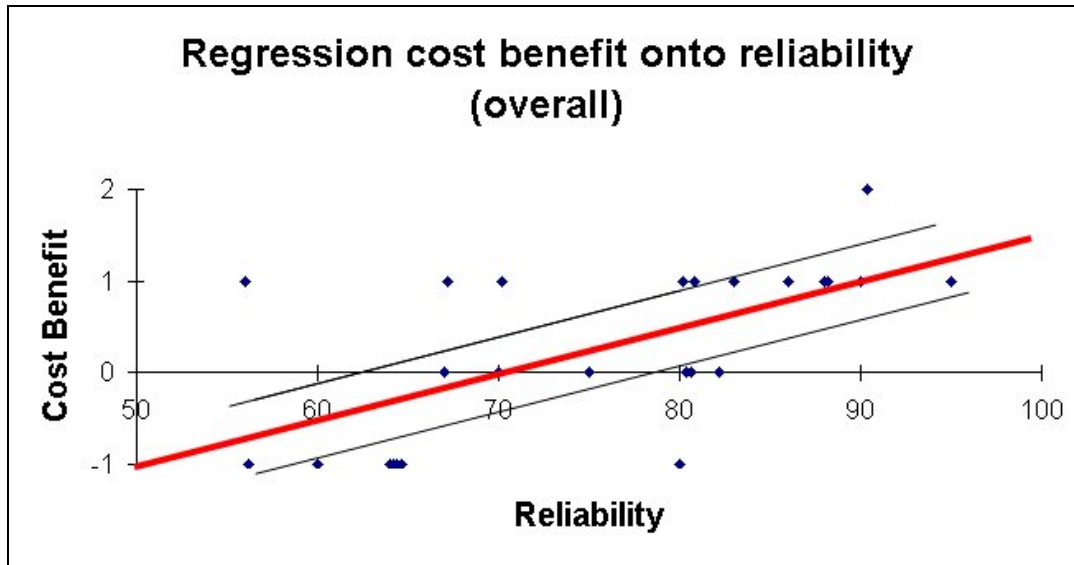


Figure 2. Regression of benefits/costs relative to baseline (heavy line) on automation reliability, with a 95% confidence interval (light lines). A single point may sometimes represent more than one study, which may have identical data X-Y values.

Importantly, the collective results described above suggest that pilots do have a tolerance for imperfect automation, as long as this reliability does not drop below perhaps an 80% level and, we note, as long as pilots are pre-warned of the nature and source of the imperfections. When reliability drops below, problems may occur. As we saw with regard to autopilot guidance, these problems may not always be severe.

5. Setting the Diagnostic Threshold: Reliance vs. Compliance

Across six of our experiments (experiments 3 and 4 discussed above, along with experiments 5, 6, 7 and 8, described briefly below), we have carried out an in depth analysis of the consequence, to human-system diagnostic performance, of a designer’s decision to adjust the **alerting threshold** of an imperfect automated diagnostic or detection system. A low threshold will produce many automation false alarms, but reduce misses. A high threshold will reduce the false alarm rate, but increase automation misses, requiring that the human be vigilant to the “raw data”, examined by the automation. The two conditions of miss-prone and false-alarm prone automation have been modeled by Meyer (2001, 2004) to create two qualitatively different changes in human **dependence** on automation. Miss prone automation reduces the **reliance** on the automation, to alert the operator should a failure occur. These differences in reliance are best seen in behavior and cognition when the alert is silent: a reliant operator assumes that there is no event when the alarm is silent. False alarm prone automation reduces the **compliance** to the automated alert. Thus differences in compliance are reflected when the alert sounds.

When the human and automation can both access raw data in parallel for a diagnostic decision, there is a straightforward interpretation as to how the setting of the automation threshold (beta) **should** effect performance in a multi-task environment. It is based on three

potentially independent phenomena: the “cry wolf” effect (Breznitz, 1983), “complacency” (Parasuraman et al., 1993) and spare capacity (Dixon & Wickens, in press; Wickens, Dixon, Goh & Hammer, 2005). These effects contribute as follows:

1. False alarm prone automation (FAP), created by lowering the alert threshold and reducing compliance is well known to produce the “**cry wolf**” effect, as the human is reluctant to leave a concurrent task when an alert occurs. This effect is most likely shown in a delayed RT to the alerting task (when and **ONLY** when an alert sounds, whether true or false), and may produce a loss of detection accuracy, if the person chooses not to leave at all. The cry wolf effect characterizes the lack of **compliance**.
2. If the operator *does* elect to leave the concurrent task to address an alarm (whether true or false), this will produce a substantial disruption of the concurrent task at the time of the alert, either to “deal with” the event if the alert is true, or establish that it is false (if it is false). Such disruption will occur when, but only when, the alert occurs. Thus for example if alerts only occur every two minutes (120 sec) and require 12 seconds to process, then disruption due to alerts is only occurring 10% of the time; this amount of disruption will be reduced if most alerts are false (compared to when they are true) and therefore the added processing is required only to double check the raw data to assure that the alert is false (rather than to deal with the alerted event).
3. During the (typically) longer period when the alert is “silent” (in the above example, 90% of the time), if the user knows that the system is FA-proneness is created by adjusting the alert threshold, and therefore *will sound* if there is an event, then this high reliance should allow a great deal of **spare capacity** to be allocated to concurrent tasks, and these tasks should be performed well, *during that 90% of the time when the alert is “silent”*. In contrast:
4. Miss-prone automation (MP) breeds low reliance, and a relatively enduring need to “check the raw data”, to pick up events that the automation likely has missed. **Spare capacity** to the concurrent task is thereby diminished during this 90% of the time, particularly if the events occur randomly in time, so the observer must continuously check the raw data to assure that one has not occurred. Still, this visual monitoring may be relatively low in terms of its cognitive demands; so while it will surely disrupt other visual tasks (that rely upon focal vision), it may not impose extensive costs on non-visual concurrent tasks (or those visual tasks that can rely upon peripheral vision).
5. The lack of reliance produced by miss prone automation will actually **improve** the response to the rare automation miss, as the user will have been more vigilant of the raw data because of a lack of **complacency**.
6. Correspondingly, if the high false alarm rate (see 1) results from a lowering alert threshold adjustment that will simultaneously **reduce** the miss rate, this will mean that FAP automation will engender long responses on the occasions when the automation misses, because of human complacency. Thus FAP automation is penalized both when the alert sounds (reduced compliance of the cry-wolf effect) and when the alert is silent (lost reliance of complacency). So response time in the FAP automation must be addressed separately for both kinds of events (alarm sounds, alarm misses), since the delay in each is due to a qualitatively different cognitive mechanism.

We employed this reliance-compliance distinction as a framework for examining the threshold shift effects across our experiments. In the two experiments (3 & 4) reported in Dixon and Wickens (in press; see also section 4 above), the threshold shift imposed was upon that of automation monitoring for system failures (see Figure 1). Here we observed the general tendencies described above, as miss rate and false alarm rate were independently varied; but noted that FAP automation produced a greater overall reduction in performance than did MP automation.

Experiment 5 was essentially a replication of three of the conditions of experiment 4 (miss prone, false alarm prone, and high reliability automation), with eye movements recorded (Wickens, Dixon, Goh & Hammer, 2005). Here we did find all three classes of effects (cry wolf, complacency and concurrent spare capacity), with the first (cry wolf) effect more likely to emerge under high workload. Importantly, we found these effects to be manifest in visual scanning measures. That is, on the one hand, miss prone automation shifted the amount of time spent examining the raw data of the system failure gauges at the expense of visual attention to the 3D image window which supported the concurrent TOO monitoring task (see Figure 1). Detection of targets of opportunity visible at the latter location, suffered accordingly. On the other hand, FAP automation delayed the time it took for visual attention to shift from the 3D image window (the dominant “sink” for visual attention) to look at the SF gauges when an alarm occurred.

In experiment 6 (Wickens, Dixon & Johnson, 2005), we changed the task to which we applied diagnostic automation, now using the TOO surveillance task, assisted by a fallible automatic target recognition (ATR) system, again with MP or FAP automation at a 70% reliability level. The priority of the two tasks was also varied between participants. In this experiment, both threshold settings of automation still improved TOO detection performance relative to a baseline (non automated) condition, assessed in experiment 4. Also, we did not find the expected “complacency” effect (longer RT to the rare automation misses in the FAP condition). We did however find evidence for the “cry wolf” phenomenon, but only when the automation task itself was emphasized. Finally, concurrent tasks suffered more under miss prone than under FAP automation (as predicted), but only when the concurrent task (detecting system failures) was emphasized. When the automated task was emphasized, then FAP automation actually disrupted the concurrent task *more*, an unexpected effect, but one reflecting the general disruptive effect of automation false alarms (Bliss, 2003).

In experiment 7 (Levinthal & Wickens, 2005), we again automated a perceptually difficult target surveillance task through a 3D image window, as in experiment 6. However here reliability was at 60% and we used the SIL (Systems Integration Laboratory), provided by MAAD and General Dynamics, as the realistic platform within which to examine threshold setting effects. We did not observe any effect of alert threshold setting on the concurrent task of supervising the UAV trajectories (i.e., no spare capacity effect), a finding that reflected our pilot participants placing highest priority on mission management. We did however observe a strong “cry wolf” effect in FAP (vs. miss-prone) automation, and also observed two different performance manifestations of complacency, reflected in performance with the low-miss (FAP) automation. Thus here again, FAP automation appears to disrupt performance more than miss-prone automation.

In experiment 7, we also manipulated the number of UAVs supervised, between 2 and 4. Our results indicated that a single-channel plus switching model (see section 3) was adequate to describe performance, since the penalty suffered from increasing load from 2 to 4 was greater than the pure time penalty of doubling the number of items to be supervised.

Finally, in experiment 8 (Dixon, McCarley & Wickens, 2005), we departed from the realistic UAV simulations, to use a more generic information processing suite of an automated diagnostic task and a highly demanding tracking concurrent task designed to amplify automation effects. Our threshold settings here were more extreme, with a high rate of one type of automation error (miss or FA) coupled now with only a 1% chance of the other type, and no pre-warning to the participants that the rare error would occur at all. Here the results again suggested a strong cry wolf effect created by FAP automation, as well as the strong complacency effect revealed by the single automation miss in the FAP condition. However in contrast to most previous studies (but replicating the automation-emphasis condition of experiment 6), we observed that FAP automation disrupted the concurrent task *more*, rather than less, than MP automation.

Collectively then, these studies reveal the general success of the reliance-compliance model, certainly in accounting for performance with the automated task: FAP automation producing a “cry wolf” effect, and, as well, by reducing the likelihood of automation misses, FAP automation producing a “complacency effect” on those rare automation misses. The less consistent aspect however is the predicted effect of miss prone automation in reducing concurrent task performance. While sometimes found (experiments 3 & 4, high workload, experiment 5, experiment 6, concurrent task emphasis), this pattern was not found when the automated task itself was perceptually or cognitively demanding (experiment 6, automated task emphasized, experiment 7 and experiment 8). Further explanations for the malleability of this effect remain obscure.

6. Conclusion: Challenges for a UAV performance Model

An effective human performance model of UAV supervision must include both effects of automation and of workload. By “workload” we refer here to the capacity to perform multiple tasks. An approach we present below is to begin at a very coarse level, and then identify necessary refinements and elaborations or qualifications.

The simplest workload model is:

$$(1) \text{ WL} = \text{N}$$

where N is the number of assets (i.e., UAVs). This is consistent with single channel theory, which, as we saw from experiment 2, was a reasonable approximation for all-visual interfaces. However we also note that a more appropriate model associates N with the number of visually rendered tasks which, in the current simulations was either 3 or 2 (autopilot) for single UAV control, or 6 or 4 (autopilot) for dual UAV control. This provides greater resolution. We also note from the results of experiment 7, when N was raised from 2 to 4 (UAVs), that the added switching penalty was imposed, suggesting that:

$$(2) WL = KN \quad (K > 1.0).$$

To elaborate this model, as we have noted above, workload costs may also be manifest in the costs of imperfect automation, although the data suggest that these may reflect a drop in performance of the automated task, more than they increase the resources allocated to that task (and degrade concurrent tasks). A simple, and computationally elegant representation here is:

$$(3) WL = KN/r$$

where r is the reliability of the automated component. Thus, for a single automated component, this relationship captures the rough linearity of performance with reliability shown in Figure 2.

As revealed in experiments 1 and 2, auditory interfaces can substantially decrease the workload via multiple resources. A computational approach to this would be to decrease the total task workload by a “modality mix” factor ranging from 1.0 (no auditory) to 0.75 (maximum reduction), a factor proportional to the number of channels off-loaded from vision to audition.

Also as revealed by both the modeling efforts of experiment 2, and by the emergence of automation unreliability effects primarily at high workload, some accommodation should be allowed for differential resource demands of component tasks. To accomplish this, the model would replace N with $\text{SUM } D$, where D = the demand level, ranging from 0 (fully and reliably automated) to 1 (a cognitively demanding manual task).

The final challenge of such a model, and its greatest practical implications and value, must be to establish (or recommend) a limitation as to how many UAVs is “too many”, exceeding a workload “red line” where critical performance will fail. With the simplest form of the model, our data from experiment 2, suggest that 1 UAV (or two non-automated tasks) defines this limit. Extrapolation of these predictions with the complete model remains a work in progress.

Acknowledgements

The authors wish to acknowledge the invaluable software support of Ronald Carbonari, and contributions to data collection made by Dervon Chang, Juliana Goh, Brian Levinthal, Nicholas Johnson and Benjamin Hammer. Dr. Michael Barnes and personnel of the E CO 305th Military intelligence battalion at Ft Huachuca provided subject matter expertise in assisting us to develop the simulation. The opinions expressed in this report are those of the authors and do not necessarily reflect those of the US Army.

References

- Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology*, 13(3), 249-268.
- Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, NJ: Lawrence Erlbaum.

- Dixon, S., McCarley, J.S., & Wickens, C.D. (2005). *Miss-prone vs. false-alarm-prone automation* (AHFD-05-16/MAAD-05-4). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dixon, S., & Wickens, C. D. (2003). *Imperfect automation in unmanned aerial vehicle flight control* (AHFD-03-17/ MAAD-03-2). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dixon, S. R., & Wickens, C. D. (2004). *Reliability in automated aids for unmanned aerial vehicle flight control: Evaluating a model of automation dependence in high workload* (AHFD-04-5/MAAD-04-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dixon, S., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle flight control: Evaluating a reliance-compliance model of automation dependence in high workload. *Human Factors*.
- Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of unmanned air vehicles: A workload analysis. *Human Factors*, 47.
- Horrey, W.J. & Wickens, C.D. (2003). Multiple resource modeling of task interference in vehicle control, hazard awareness and in-vehicle task performance. *Proceedings of Driving Assessment 2003: 2nd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*. Park City, UT.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
- Keele, S. W. (1973). *Attention and human performance*. Pacific Palisades, CA: Goodyear Publishing Company.
- Laughery, K. R., & Corker, K. (1997). Computer modeling and simulation. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed.) (pp. 1375-1408). New York: Wiley.
- Levinthal, B., & Wickens, C. D. (2005). *Supervising two versus four UAVs with imperfect automation: A simulation experiment* (AHFD-05-24/MAAD-05-7). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Liao, J., & Moray, N. (1993). A simulation study of human performance deterioration and mental workload. *Le Travail humain*, 56(4), 321-344.
- Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, 43, 563-572.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*.
- Parasuraman, R. M., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced “complacency”. *International Journal of Aviation Psychology*, 3, 1-23.

- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000, May). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics*, 30(3), 286-297.
- Rantanen, E. M., Levinthal, B. R., & Yeakel, S. J. (2005). *En route controller task prioritization research to support CE-6 human performance modeling* (Final Technical Report AHFD-05-3/MAAD-05-3). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Sarno, K. J., & Wickens, C. D. (1995). The role of multiple resources in predicting time-sharing efficiency. *International Journal of Aviation Psychology*, 5(1), 107-130.
- Sheridan, T. (2002). *Humans and automation: System design and research issues*. N.Y.: Wiley Interscience.
- Welford, A. T. (1968). *Fundamentals of skill*. London: Methuen.
- Wickens, C. D., (2002) Multiple Resource and Performance Prediction. *Theoretical Issues in Ergonomic Sciences*. 3(2), 159-177.
- Wickens, C. D., & Dixon, S. (2002). *Workload demands of remotely piloted vehicle supervision and control: (I) Single vehicle performance* (AHFD-02-10/MAAD-02-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., & Dixon, S. (2005). *Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature* (AHFD-05-1/MAAD-05-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., & Dixon, S. (2006, in press). *The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Sciences*. Also available as (AHFD-05-1/MAAD-05-1 Technical report). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C.D., & Dixon, S. (in progress). *Task priorities and imperfect automation* (AHFD-05-17/MAAD-05-5). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., Dixon, S., & Chang, D. (2003). *Using interference models to predict performance in a multiple-task UAV environment – 2 UAVs* (AHFD-03-9/MAAD-03-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., Dixon, S., Goh, J., & Hammer, B. (2005). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis* (AHFD-05-2/MAAD-05-2). Savoy, IL: University of Illinois, Aviation Human Factors Division.

- Wickens, C. D., Dixon, S., & Johnson, N. R. (2005). *UAV automation: Influence of task priorities and automation imperfection in a difficult surveillance task* (AFHD-05-20/MAAD-05-6). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., Goh, J., Helleberg, J., Horrey, W., & Talleur, D. A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors*, 45(3), 360-380.
- Wickens, C. D., & Hollands, J. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- 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.