



**Aviation Human Factors Division
Institute of Aviation**

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

**UAV Automation: Influence of Task
Priorities and Automation Imperfection
in a Difficult Surveillance Task**

**Christopher D. Wickens,
Stephen Dixon & Nicholas R. Johnson**

**Technical Report
AHFD-05-20/MAAD-05-6**

December 2005

Prepared for

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

Contract ARMY MAD 6021.000-01

Abstract

This experiment replicated many aspects of a UAV simulation paradigm in which imperfect automation aids assisted an easy visual monitoring task, while pilots concurrently performed a difficult surveillance task for targets of opportunity, and a primary mission completion task. In that study, changes in the threshold of the automated alert affected both the automated task and the concurrent task in a way suggesting that false-alarm prone automation is more disruptive of both tasks, than is miss-prone automation. In the current study, automation was applied to the difficult surveillance task, and task priority was manipulated. The current results still revealed a greater cost of false-alarm prone automation, again affecting both the automated and concurrent task. Priority instruction tended to produce better performance on the favored task. However with false-alarm prone automation, favoring this task actually reduced its performance. The results are interpreted in terms of the dichotomy between reliance, fostered by miss free automation, and compliance, fostered by false-alarm free automation. These two constructs were found to be only partially independent of each other.

Introduction

Several previous studies within our laboratory have examined a UAV simulation supported by imperfect automation (Dixon & Wickens 2003; Dixon & Wickens 2004; Wickens, Dixon, Goh & Hammer, 2005; Levinthal & Wickens, in preparation; see also Dixon, McCarley & Wickens, 2005 for a related study). All of these studies have required pilots to perform a primary mission task, navigating to waypoints and doing surveillance at the waypoint, a secondary search task for camouflaged targets of opportunities (TOO task) and another secondary monitoring task, overseeing dynamic system parameters to determine if they are in a normal or “failed” state (system failure, or SF task). In all of these studies the SF task has sometimes been supported by an imperfect alert. Like many alerting tasks in the real world, in which imperfections result from a variety of sources, the **threshold**, can be adjusted to vary the relative frequency of two different classes of automation “errors”, making either automation misses or automation false alarms more likely. In most real world monitoring tasks designers adjust this alert threshold to minimize misses at the cost of producing more false alarms, on the basis of the assumption that a critical event that is missed, is more damaging to overall system performance, than a non event that is signaled by automation to occur.

Across these studies a few general trends have been noted. (1) Automation of high (i.e., 90%) but still imperfect reliability supports better performance than that observed in a manual baseline condition with no automation. (2) When automation reliability falls below about 75%, performance on the SF and TOO task tends to be no better than, and often worse than the baseline level (see also Wickens & Dixon, 2005, in press). (3) Generally, imperfect automation appears to disrupt performance of the automated SF task, more than that of the concurrent TOO task, as if pilots are protecting the latter from the problems associated with the former (see also Wickens & Dixon, 2005). That is, they are treating the concurrent task as “primary” and the automated task as secondary.

(4) Comparing the two types of automation imperfections created by adjustment of the threshold setting, the studies revealed that FA-prone automation tends to be more disruptive to overall performance of both the automated and the concurrent task, than does miss-prone automation (see also Bliss, 2003). (5) The two types of automation imperfection have qualitatively different effects on performance, as captured by the “reliance-compliance” distinction proposed by Meyer (2001, 2004; see also Maltz & Shinar, 2003). Automation with few misses breeds high **reliance** when the alert is silent, because the operator is confident that all critical events will be notified. This cognitive state of reliance avails ample spare capacity to perform concurrent tasks, although it will strongly disrupt event detection on those very rare occasions when the automation **does** miss an event. This latter phenomenon reflects a kind of “complacency”. Miss-prone automation thereby reduces reliance. In contrast, automation with few false alarms breeds high **compliance**, meaning a rapid switch of attention and response to the alert when the alert sounds. False alarm-prone automation then reduces compliance – the so-called “cry wolf” effect -- and leads to delay and sometimes absent responding to all alerts, whether true or false. However with a false-alarm prone system, concurrent tasks should be protected while the alarm is silent, as the operator can remain confident that true failures or events will always be alerted.

As originally proposed by Meyer, reliance and compliance were suggested to be independent functions of automation miss rate and false alarm rate respectively, and this indeed should be the behavior of an optimally functioning human operator as shown in Table 1. We refer to this as the “independence model” of reliance and compliance. However our research on imperfect UAV alerting systems suggests this to be only partially the case. On the one hand, miss rate does tend to influence only indices of reliance. On the other hand however, we have found that increasing alert false alarm rate appears to degrade **both** indices of compliance **and** reliance. In particular, high false alarm rates tend to disrupt concurrent task performance, even if such conditions signal that all true failures will be alerted (i.e., few or no automation misses), and therefore even if pilots would not need to divert attention to monitor the system status when the alert is silent. Because of the dual penalty of increasing false alarm rate, we refer to this as the “false alarms hurt” model, in contrast to the independence model. One purpose of the current research is to explore the independence vs. “false alarms hurt” question with a qualitatively different form of diagnostic automation (automated target recognition) than that used in prior studies.

Table 1.

Miss rate → reliance → concurrent task performance

FA rate → compliance → automated task

The fact that our prior research revealed that imperfections in the automation tend to disrupt that task that is automated, more than concurrent tasks (see (3) above), is of considerable theoretical and practical importance. There are at least two possible explanations or hypotheses

that can be offered to account for this finding. First, it is possible that humans implicitly assume that when a task is automated, that automation, by offering assistance, therefore is supposed to avail more resources to concurrent tasks, thereby implicitly signaling the latter as being “primary”, to be protected from other resource demands where possible, and signaling that the automated task is “secondary”, allowing its performance to bear the costs of imperfections. Second, in the paradigms employed in our UAV simulation, the SF task has always been an easy task to perform unaided. Indeed the only reason for automating this task in the first place is to allow foveal vision to more continuously monitor the other UAV displays, particularly for the target of opportunity (TOO). Hence it may be that the **easiness** of the SF task, rather than its automated properties, induced operators to treat it as secondary, allowing degradation in its reliability to impose on its own performance, rather than the more “primary” and difficult concurrent tasks. Importantly, in the paradigms above, task priority was never explicitly assigned nor manipulated.

Thus a second purpose of the current study was to examine the implications of both of these explanations, in the same UAV simulation as used in the previous studies, with two modifications. First, across different conditions, task priority was explicitly varied between the TOO task and the SF monitoring task. If the first explanation is correct, then whichever task is designated “secondary” should bear the brunt of the automation imperfections. Second, automation is now applied to the more difficult TOO task. If the second explanation is correct, then this more difficult task will always be treated as more primary, and the easier SF task will always bear the greater brunt of imperfections, independently of priority manipulations. In applying diagnostic automation to the more difficult perceptual task, we are actually replicating a fairly common form of automation related to “automated target recognition” (ATR; Maltz & Shinar, 2003; Goh, Wiegmann, Madhavan, & Wong, 2004; Wiegmann, McCarley, Kramer & Wickens, in preparation). This alteration from the previous UAV paradigm is important because it provides a greater degree of ecological validity for the current study, in that in most real world contexts, automation which is asked to detect difficult-to-see visual targets can be expected to make errors.

In the current study then, pilots performed the similar UAV simulation to that used in previous research, except that TOO surveillance was now assisted by imperfect diagnostic automation, while the system monitoring task was supported by perfect automation.

Because we varied the diagnostic threshold setting or the TOO automation (for the first purpose above), and TOO vs. SF task priorities (for the second purpose), it was possible for us to make predictions regarding the main effects and possible interaction between these two variables, in a way that highlights the two roles of attention as a modulator of performance: allocating visual attention between tasks when the alarm is silent, and switching attention to the automated task, when an alert sounds. These predictions or hypotheses are as follows:

1. Performance on the prioritized task will be improved: that is, a main effect of task priority on both SF and TOO dependent variables of speed and accuracy is anticipated.
2. Concurrent task performance will be degraded by miss-prone automation. This refers to the classic “reliance effect”, in that visual attention must be allocated to monitor the raw data (the 3D image window) of a miss-prone system, and hence away from other relevant

areas. The cost will show up in SF performance, and probably retention of command target information (use of repeat key).

3. Automated (TOO) task performance will be degraded by FA-prone automation. Here, because the threshold setting makes automation misses relatively rare, pilots will suffer “complacency” effects on the rare occasions when automation **does** miss the target. It should be noted that hypotheses 2 and 3 can sometimes be grouped together to describe different poles of the “reliance” syndrome. High miss rate pulls attention away from the concurrent task to the raw data: Low miss rate pulls attention away from the raw data of the automated task.
4. Automated (TOO) task performance will be degraded by FA-prone automation in a different way, reflecting the “cry wolf” syndrome of lost compliance; that is, the failure to switch attention to the TOO task even when a correct alert sounds.
5. Threshold setting and priorities will interact in their effects on both tasks. This hypothesis, offered with slightly less confidence than 1-4, is based on previous findings that “automation effects” tend to be amplified, as resources become more scarce (Wickens & Dixon, in press; Dixon & Wickens, in press). In previous research, resource scarcity was caused by an increase in concurrent task difficulty. Here we consider emphasizing the concurrent task to be a proxy for producing such resource scarcity. Thus:
 - The cry wolf ignorance of a true TOO alert, brought about by FA-prone automation will be amplified when the TOO task is neglected as a result of SF-emphasis instructions.
 - The concurrent performance cost of miss prone automation, brought about by needing to monitor the raw data in the TOO 3D image window, will be amplified as that concurrent task is neglected as a result of TOO-emphasis instructions.
6. Workload effects will mimic priority effects. That is, when heavy cognitive processing must be invested in one task or the other (e.g., system failure diagnosis), the imperfect automation costs for the other will be amplified.

Methods

Thirty-two undergraduate and graduate students at the University of Illinois received \$8 per hour, plus bonuses of \$20, \$10, and \$5, for 1st, 2nd, and 3rd place finishes, respectively, in their group of eight pilots. Figure 1 presents a sample display for a UAV simulation, with verbal explanations for each display window and task.

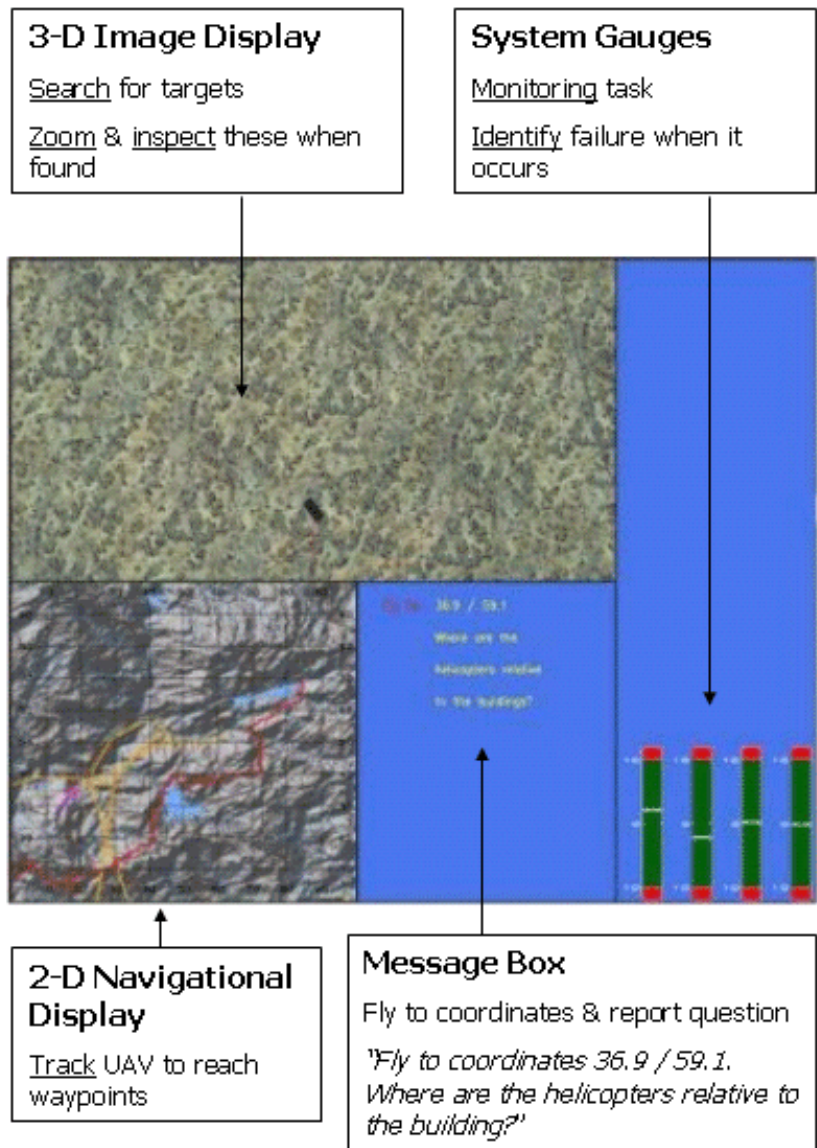


Figure 1. A UAV display with explanations for different visual areas.

As seen in Figure 1, the experimental environment was subdivided into four separate windows. The top left window contained a 3D egocentric image view of the terrain directly below the UAV. The sample figure shows a command target (CT) at normal viewing distance (i.e., 6000 feet altitude). During regular tracking periods, the operator could only view straight down to the ground at a 20-degree angle. During a loiter pattern, the operator was able to extend the viewing angle from 0 to 90 degrees along both the x- and y-axes. A zoom feature (up to 100x) was also available only in the loiter pattern.

The bottom left window contained a 2D top-down map of the 20 x 20 mile simulation world. Coordinates (which formed a grid) from 0-100 were placed along the x- and y-axes for

navigation purposes. The yellow and red lines denoted minor and major roads, respectively. The smaller blue lines denoted rivers, and the large blue shapes denoted lakes. The bottom center window contained the Message Box, with “fly to” coordinates and CT report questions. These flight instructions were present for 15 seconds, and could be refreshed for another 15 seconds by pressing a Repeat key. The bottom right window contained the four system failure (SF) gauges. Each gauge represented a different onboard system. The white bars oscillated up and down continuously, each driven by sine waves ranging in bandwidth from 0.01 Hz to 0.025 Hz. A SF occurred when one of the white bars moved into a red zone.

Participants used a Logitech Digital 3D joystick to manipulate the aircraft/camera and a X-Key 20-button keypad with which to indicate responses. The joystick had controls for turning the UAV, manipulating the camera on the x- and y-axes, zooming, detecting targets, loitering around targets (to the left or right), and detecting SFs. The keypad was used for indicating which system failure occurred, the ownship coordinates for that system failure, and for typing in mission coordinates during the Automation condition.

Each pilot flew one UAV through ten different mission legs, while completing three goal-oriented tasks commonly associated with UAV flight control: mission completion, target search, and systems monitoring. At the beginning of each mission leg, pilots obtained their flight instructions for that leg via the Message Box. Once pilots arrived at the CT location, they loitered around the target, manipulated a camera for closer target inspection, and reported back relevant information to mission command (e.g., *What weapons are located on the south side of the building?*). Around each CT were 1-3 tanks and/or helicopters, located within 10-30 feet of the building. These weapons were always located on the north, south, east, or west sides. Location was to be specified in cardinal directions, thereby forcing a relatively high level of spatial-cognitive activity (e.g., Gugerty & Brooks, 2001)

Along each mission leg, pilots were also responsible for detecting and reporting low-salience targets of opportunity (TOO), a task similar to the CT report, except that the TOOs were much smaller (1-2 degrees of visual angle) and were camouflaged. They were located randomly somewhere in the middle 60% of each leg (i.e., between 20% and 80% of distance traveled); however, participants were not told this. Similar to the CTs, each TOO contained 1-3 tanks and/or helicopters, located within 10-30 feet of the bunker, located on the north, south, east, or west sides. The question for TOOs was always the same: “*what weapons do you see and where are they located?*” As with the CTs, location was to be specified in cardinal directions, and these questions could only be answered once the operator had zoomed in close to the target. TOOs could occur during simple tracking (low workload) or during a pilot response to a system failure (high workload). These two types of TOOs occurred, respectively, with a ratio of roughly 3:1.

If the participant detected a CT or TOO, he or she was required to indicate detection by pulling the joystick trigger. The duration of time between when the target entered the 3D display and when the pilot pressed the detection button was recorded as target detection time. The participant then pressed the loiter button (loiter would be selected either left or right) on the joystick. This put the UAV into an automated oval pattern around the target. This oval pattern was 1.3 kilometers wide and 2.1 kilometers long, and took between 2.5 to 3 minutes to complete an entire 4.8-kilometer circuit. The UAV turned 3 degrees per second at the ends of the oval.

After making the report, the participant could then depress the loiter button again, which would unloiter the UAV and unzoom the camera, returning the egocentric view to 6000 feet altitude.

An automatic target recognizer (ATR) was available in all four conditions. This system would provide an auditory signal when it sensed that a TOO had appeared in the 3D image window, although it did not designate the location where it inferred the tank to be. In the miss-prone system, out of 40 possible TOOs, 12 of these were not signaled by the automation (3 high workload and 9 low workload), while there were 4 false alerts (1 high workload and 3 low workload). In the FA-prone system, the opposite occurred; that is, there were 12 FAs and 4 misses out of 40 possible events, with the same high-low workload ratio. False alerts were imposed. Pilots were not aware of the precise level of reliability provided by the automation aid; however, depending on their condition, they were told of the automation bias setting (i.e., more false alarms or more misses).

Concurrently, pilots were also required to monitor the system gauges for possible system failures (SF). When a gauge went “out of bounds” (i.e., the needle moved from the green zone to the red zone), they had to press a button to detect the SF, indicate which SF gauge had failed, and then report the current location of the UAV during the SF. SFs were designed to fail either during simple tracking (no TOO on the horizon) or just before a TOO entered the window. The SFs lasted only 30 seconds, after which the screen flashed bright red and a harsh auditory alarm announced that the pilot had failed to detect the SF (the UAV was considered to have “crashed” if pilots did not detect the failure quickly enough). There were a total of 40 SFs, with never more than three SFs occurring during any mission leg. SFs were temporally separated by 3-6 minutes.

Priority in the experiment was manipulated by instructing participants to focus 75% of their attention/priority on either the TOO task or the SF task, and 25% of their attention/priority on the other task, or vice versa. The participants were told that their performance rewards would be weighted according to the priority instructions; that is, if they were instructed to put 75% of their attention into the TOO task, then the TOO task performance measures would be weighted at 75% of the total.

Results

Using a trichotomy applied to previous studies, we distinguish those subtasks associated with the primary task of mission completion, from those associated with the two subsidiary tasks: the “automated task” of TOO surveillance, and the “concurrent task” of SF monitoring. Within the overall task hierarchy, mission completion is considered most critical, and therefore is truly “primary”. However within the two subsidiary tasks, our two instructional sets designated either the TOO task or the SF task to be “secondary” or of low priority relative to the other. Within the following presentation, there are occasions when we refer to certain statistical effects as not having sufficient power to reveal an apparent difference. The source of power reduction here is not the loss of degrees of freedom from fewer subjects. Rather it is that the conditions under scrutiny generated unstable data per participant because of a low number of trials. For example these would include the rare “automation miss”, or the “high workload” trial that only occurred in 1 out of 6 observations. As a consequence of the fact that each participant experiences this condition rarely, the sample mean for that participant is a less stable estimate of

the participant's true mean, and so that estimate can be expected to be more variable. Such variability is then reflected in increased between-subject variability, the source of low power.

Mission Completion

Three measures were used to assess mission completion performance. First, tracking (deviation from the flight trajectories) was found to be unaffected by any of the independent variables. Second, command target reporting speed was found to be 5 seconds faster under TOO-priority instructions ($F(1, 27) = 28.04, p < .01$). This effect can be well explained by the fact that such instructions moved visual attention to the 3D image window (where the TOOs appeared) and thereby inherently supported earlier detection of the command targets when they appeared within that window. Command target reports were also 2 seconds slower with the FA-prone automation than with miss-prone automation ($F(1, 27) = 3.58, p = .07$), an effect that can be explained by the allocation of attention **away** from the 3D image window under those circumstances (FA-prone threshold) when automation misses were assumed to be rare.

As a third index of mission completion, we measured the number of times that pilots requested "repeat" instructions of the command target identity and location, as an implicit measure of competition from concurrent tasks; assuming that more such repeats reflected more competition for memory of this mission-critical information. Repeats were greatly increased with False-Alarm prone automation, a 3-fold increase from their level in the miss-prone condition ($F(1, 27) = 54.95, p < .001$). This effect was opposite from the predictions (H2) that concurrent task performance would be more degraded by a **miss-prone** threshold setting.

System Failure Detection (the "concurrent task")

There were no effects of condition on system failure detection accuracy, which was essentially perfect, even when the TOO task was emphasized. This perfection can be attributed to the effectiveness of a totally reliable auditory alert in eventually capturing attention. Thus all experimental variance was cast into SF RT.

SF RT revealed a strong effect of priority ($F(1, 27) = 15.35, p < .01$), as shown in Figure 2. Not surprisingly, when the SF task was emphasized, SF RT was faster. There was no significant effect of TOO alert threshold (miss vs. FA prone) on SF RT. However a significant threshold X priority interaction ($F(1, 27) = 9.6, p < .05$) revealed that when the SF (concurrent) task was emphasized, its own performance suffered more (longer RT) under miss-prone automation than FA-prone automation ($t(14) = 1.83, p < .05$), a trend consistent with the "reliance" construct. However when the TOO (concurrent) task was emphasized, this pattern reversed, and FA-prone automation disrupted the concurrent (SF) task more than did miss-prone automation ($t(13) = 2.46, p < .05$). This latter pattern of results is not predicted under the reliance-compliance independence model, which would predict consistently (across both priority conditions) greater disruption to concurrent tasks caused by miss-prone automation. The effect supports the "false alarms hurt" model. An alternative way of viewing the interaction is that with a FA-prone system, priorities had their expected effect on the concurrent task. But with a miss-prone system, they did not.

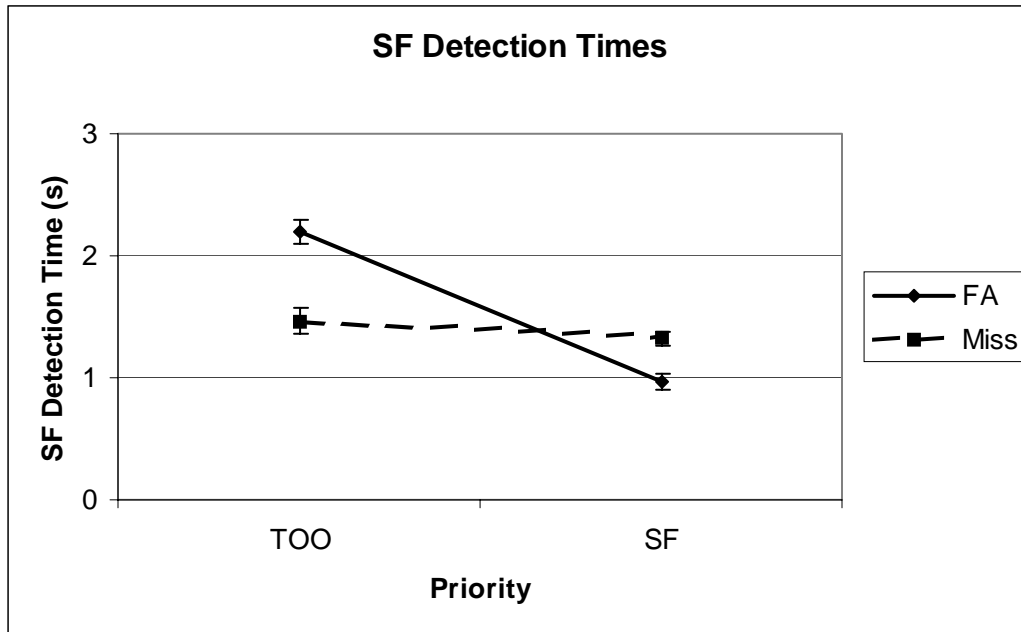


Figure 2. System failure response time.

SF response time was also examined as a function of concurrent task workload, defined as whether a system failure occurred while pilots were simply monitoring for TOOs, or were actually engaged in TOO inspection. Somewhat surprisingly, no effect of workload was found in any of the conditions.

It will be noted that SF detection time in the baseline condition of the previous study (Wickens et al, 2005), in which no TOO automation was present, was 2.2 seconds in low workload, and 4.8 seconds in high workload. The fact that all entries in Table 2 are equal to or less than even the low workload value clearly establishes the advantage of automation in the current simulation, even at this fairly low reliability level.

TOO Detection

Unlike the SF task, performance differences in the surveillance task of monitoring for targets of opportunity (TOOs) were reflected in both RT and accuracy. The two variables will be discussed separately here, although they could readily be combined in a single measure of performance quality.

TOO RT. Shown in Figure 3, there was a main effect of task emphasis on TOO RT, in the expected direction revealing that response times shortened when the TOO task was emphasized ($F(1, 27) = 32.44, p < .001$). This effect was equivalent for both automation threshold conditions, as there was no significant interaction between threshold and priorities. While threshold condition did not significantly effect TOO RT, closer examination of the data points reveals that when the surveillance (TOO) task was emphasized, FA-prone automation significantly lengthened TOO RT compared to miss-prone automation ($t(13) = 1.59, p = .07$). When the SF

task was emphasized, there was no significant difference between the two alert threshold conditions. The source of this difference in threshold effects between the two priority levels is attributed to the substantially smaller standard error measure (and greater statistical power resulting) when the TOO task was emphasized.

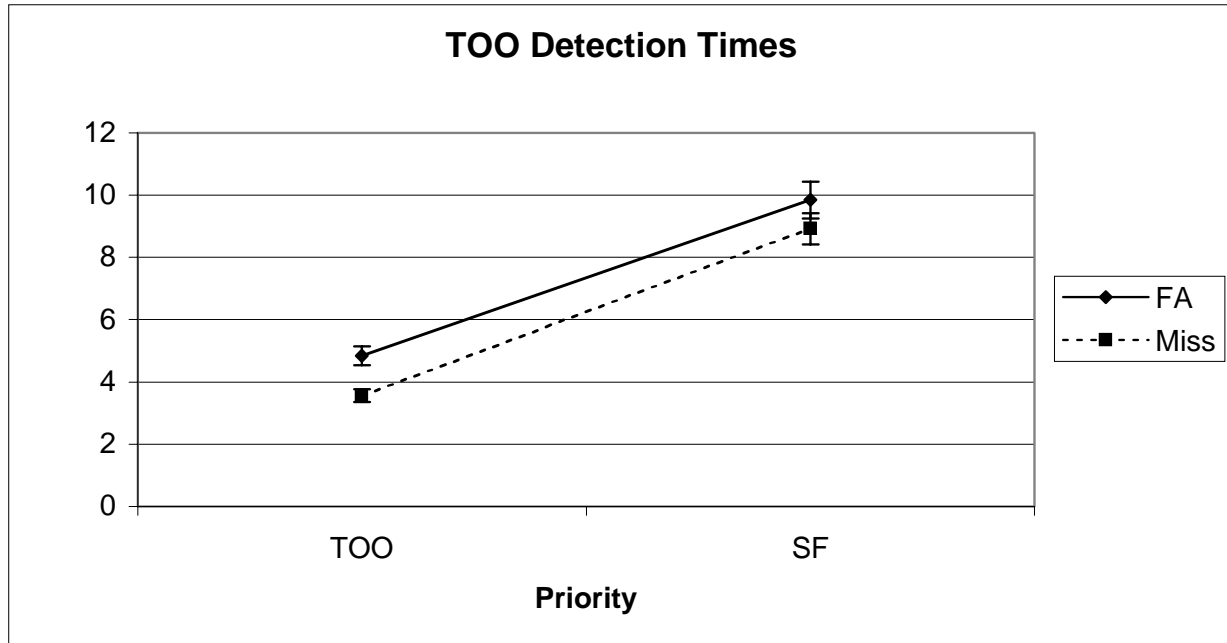


Figure 3. TOO detection time.

TOO detection times were also lengthened at high workload ($F = 7.9, p < .01$). Although workload did not interact significantly with other factors, its interaction with threshold level approached significance ($F = 2.87, p = .10$). This leads us to examine more carefully the threshold effects specifically at low and high workload. Not surprisingly, the low power, caused by the few trials/subject at high workload, precluded any significant effects. The higher statistical power at lower workload however revealed a significant delay of FA-prone automation, relative to miss-prone automation ($F = 4.65, p = .04$). Finally, when TOO detection times were then analyzed as a function of whether automation detected or missed the TOO event, as expected, there was a 2 second lengthening of RT on the automation-missed trials ($F = 19.6, p < .01$). While this lengthening was itself longer on FA-prone automation (3 seconds) than miss-prone automation (1 second), this difference was not significant ($p > .10$). However, the difference was in the direction predicted by the “complacency effect”. That is, in FA-prone automation, automation misses were relatively rare, therefore unexpected, and therefore should yield a slower response, compared to the response in miss-prone automation, where more attention should be given to the raw data. The absence of statistical significance here, can probably be attributed to the high variance and low statistical power of the rare automation miss trials.

It should be noted that the TOO RT for a corresponding baseline condition, as measured in the previous experiment (Wickens et al., 2005) was approximately 7 seconds, reflecting equivalent performance to the current study.

TOO error rate. Figure 4 depicts the TOO miss rate as a function of priority and threshold setting. The analysis revealed that neither priority nor threshold setting influenced error rate, but the two variables imposed a significant interaction ($F(1, 27) = 5.16, p < .05$), which can be clearly seen in the figure. This interaction can be described in either of two ways: (1) When the SF task was emphasized there was no effect of threshold setting, but when the TOO task was emphasized, FA-prone automation created more errors (misses) than did miss prone automation ($p < .05$). (2) TOO task emphasis improved TOO accuracy in the miss prone condition (an expected effect; $p = .07$), but actually **disrupted** TOO performance in the FA-prone condition ($p < .06$). The analysis also revealed that load (not plotted in Figure 4), imposed a mean 10% cost to detection rate. This workload effect was constant across priorities and threshold setting (all interactions with load, $p > .10$).

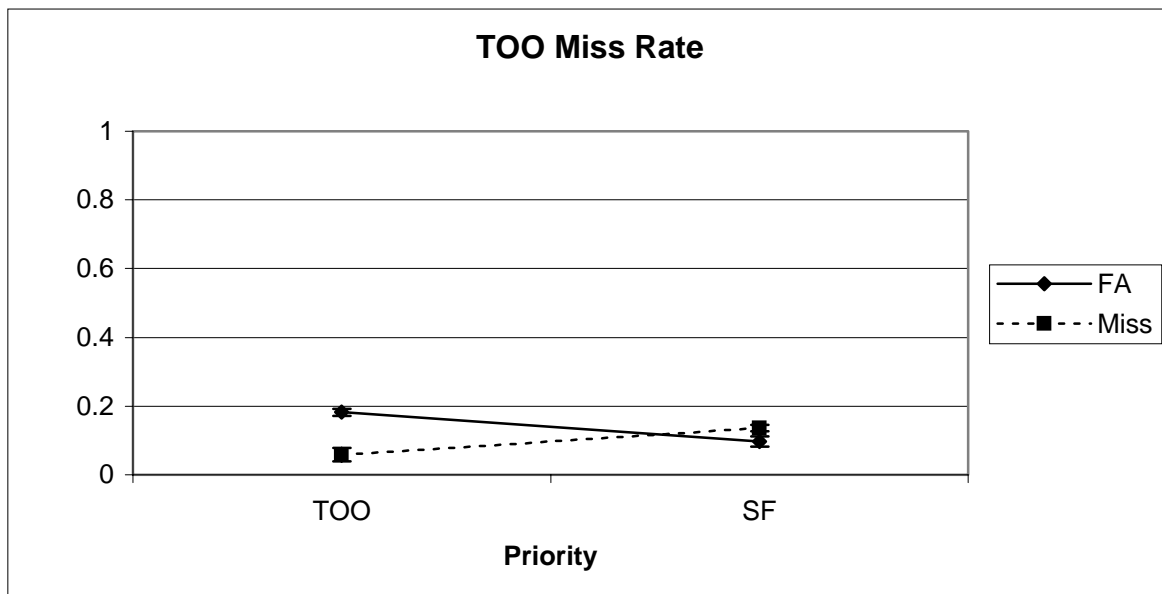


Figure 4 TOO Miss rate.

When the four data points depicted in Figure 5 are examined in more detail, the high miss rate of FA-prone automation with TOO emphasis can be understood to result from two factors: the “cry wolf” effect, leading to pilots ignoring some true alerts, and the “complacency” effect, imposing a particularly high miss rate on that minority of failures in which the automation also fails (i.e., the “miss” effect reported on TOO RT above).

It is puzzling however why performance with FA-prone automation deteriorates, when the automated task is emphasized. In order to examine in more detail the nature of behavior in this automation emphasis condition, the two data points on the left side of Figure 4 were broken

TOO Priority Conditions:

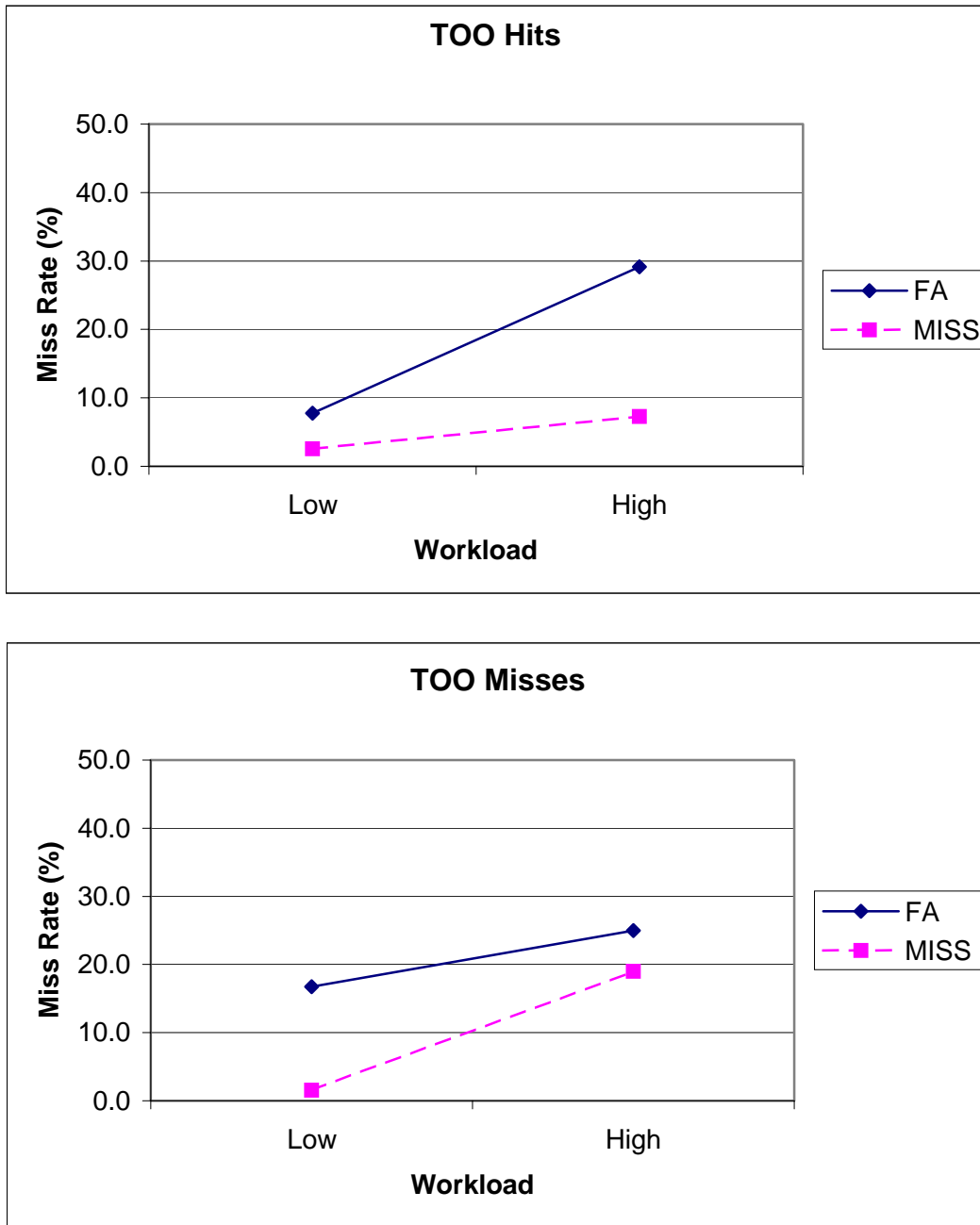


Figure 5. TOO miss rate as a function of workload (X axis) and alert threshold (the two lines), for automation hits (top graph) and misses (bottom graph). TOO emphasis condition only.

down further by whether automation detected or missed the system failure, and by whether concurrent task workload was low or high (the latter, when a system failure diagnosis was ongoing). These data are shown in Figure 5. A main effect of workload, echoing the findings reported above, indicated that in all conditions, detection accuracy was degraded in high workload. The lower panel in Figure 5 (when automation missed) revealed no other significant effects, an absence attributable to the small sample size of the data (automation misses were rare, particularly in the FA-prone condition). However the non-significant trend for higher miss rate with FA-prone automation can certainly be attributable to the abovementioned “complacency” effect. Automation misses were rare and therefore unexpected here, leading to less vigilant processing of the raw data within the 3D image window.

In contrast, the data for TOO hits depicted in the top panel, more powerful because of the greater stability of these data (more observations per participant), showed significant effects: a significant cost of FA-prone automation ($F(1,13) = 7.54, p < .02$), and a marginally significant amplification of this cost at high workload ($F(1,13) = 3.54, p = .08$). This pattern can be directly attributed to the “cry wolf” effect and its amplification at high workload. When a true alarm (but with the high potential to be false), is issued during low workload periods, the participant will switch to the alarm domain and detect the target. But if such an occurrence occurs during high workload, the participant will be more reluctant to disengage from the ongoing system failure diagnosis, and instead, will accept the reasonable likelihood that the alarm is a false one, and fail to check the raw data 30% of the time.

Left unanswered by this analysis, is why this plausible accounting of effects shown in Figure 5 only emerges when the automated task is emphasized. We infer that such emphasis does not bring more attention to the raw data itself, but rather to the TOO task set as a whole. When it does so, it calls as much attention to the imperfections of the automation (and its associated “cry wolf” effect) as it does to examining the raw data underlying the automation. Hence the cry wolf effect is amplified, relative to the circumstances when the TOO task is de-emphasized.

Finally, it should be noted that TOO accuracy in the baseline (non-automated) condition from the previous study (Wickens et al., 2005) was 0.18 at low workload, and 0.42 at high workload indicating that, overall, the automated alert improved accuracy in this difficult task, in spite of the relatively low level of automation reliability.

Subjective Trust

Ratings of the subjective reliability of the automation, estimated by the operator’s after the experiment were collected as an objective metric of “trust”. These data revealed a mean rating of 0.69, which did not differ significantly between conditions ($p > .10$).

DISCUSSION

In the current experiment, the collective pattern of data were somewhat, but not entirely consistent with the sets of hypotheses offered at the end of the Introduction. We consider each in turn.

H1: Task Emphasis: Generally it was found that emphasizing a task improved its performance (Norman & Bobrow, 1975). This was true for response time to both of the tasks,

and accuracy for the TOO task during miss-prone automation. The trend was not contradicted by SF accuracy (which was at a ceiling level) and was also consistent with the more rapid detection of command targets when the TOO task was emphasized, since those command targets appeared within the same 3D image window as did the TOO. Hence task emphasis, at least in part, drives the allocation of visual attention. The only exception to the predicted finding occurs with False-alarm prone automation, where emphasis produced reduced, not improved detection accuracy (Figure 4). We explain this by assuming that one consequence of emphasizing the TOO task is to draw attention to the alerts themselves, and therefore their false-alarm proneness, as much as to the raw data underlying these. As the alerts gain more attention, their high false rate gains more salience, and the “cry wolf” syndrome itself becomes more manifest, leading to an increased miss rate.

H2: The “reliance” effect postulated by Meyer (2001, 2004) and Dixon and Wickens (in press) is hypothesized to be manifest in concurrent task performance. As automation becomes more miss-prone, more attention must be allocated to the raw data, and concurrent tasks will suffer. Here we found only partial agreement of the data. For the system failure task, this pattern was observed only when the SF task was itself emphasized. Here we assume that a large portion of visual attention is allocated to the SF display, but is then withdrawn as the 3D image must be scanned with more vigilance, in miss prone automation (right side of Figure 2).

In contrast, when the TOO task is emphasized, we find the opposite pattern. The concurrent task of SF detection is delayed more under FA-prone automation than miss prone automation. Furthermore, the most sensitive index of the primary mission completion task (repeats) does not demonstrate the anticipated decline in performance that would be predicted for this concurrent task with an increased monitoring of raw data in the 3D image window. Thus the current data suggest that the disruptive effects of false alarm prone automation on other tasks are at least as strong as are those of miss-prone automation, at least when the automation task receives emphasis, a finding consistent with our earlier work (Dixon & Wickens, in press; Wickens, Dixon, Goh & Hammer, 2005).

H3: A second aspect of the “reliance” state posits that high reliance (on automation that detects nearly all events), will lead to a decreased monitoring of the raw data, and hence a degraded human detection of those (now rare) events that automation also misses. In the current data, this effect was, in part supported by an increase in RT to TOO events that the automation misses (relative to automation hits), the so-called “complacency” effect. This effect in turn was 2 seconds greater when those misses are rare (FA-prone condition) than when they are more frequent (Miss-prone condition), although low power (low N per subject) in the former condition, precluded a difference of statistical significance. The effect is also shown by the increase in TOO miss rate of FA-prone relative to miss-prone automation when the automation misses (Figure 5), but this trend was also not significant ($p > .10$).

H4: the “Cry wolf” effect: loss of compliance with FA-prone automation. This effect – delaying a response to or ignoring altogether a true alarm – was manifest in different forms, in both the TOO RT and detection accuracy data. In the former measure (RT), a delay in responding was found (Figure 2) to be significant when the TOO task was emphasized, and present, but not significant when the system failure task was emphasized. In the latter measure

(accuracy), degraded accuracy was found (Figure 5, top panel), although significantly so only in high workload.

H5: Amplification of automation costs as resources are withdrawn. As suggested above, this hypothesized generic interaction effect is inconsistent across the current data set. In particular, both the reliance and compliance problems associated with false-alarm prone automation for the automated task appeared to be manifest as *more*, rather than fewer resources were allocated to this task (i.e., amplified under TOO emphasis conditions). Such an effect is better explained by the cognitive effects of allocation instructions, to pay as much attention to the automation **governing** a task (i.e., the alert system including its imperfections) as to the raw data underlying the task product itself.

H6: Amplification of automation costs at high workload. This effect, paralleling that of resource withdrawal described in H5, was certainly found with respect to the “cry wolf” effect (H4), at least when the TOO task was emphasized (Figure 4). Here the cry-wolf compliance degrading of detection accuracy only appeared at high workload. Nowhere however was this hypothesis contraindicated by the data.

Another way of summarizing the current set of effects is represented in Table 2, which separates the “compliance effects” associated with varying FA proneness (cry wolf) at the top of the table, from the “reliance effects” associated with varying miss proneness (both increased concurrent task performance, and “complacency” on the rare automation misses or “automiss”), shown at the bottom. Within the table, the influences on reliance and compliance, represented by a change in miss rate or false alarm rate, are represented on the left, and themselves are separated by whether they were observed in the TOO emphasis or SF emphasis condition (**BOLD UNDERLINE**). These are connected, by arrows, to their observed effects on the right, either on the automated TOO task itself (far right column) or on the concurrent task. Finally, the heavy solid arrows represented effects in the direction predicted by the reliance-compliance independence model (Table 1). The single dashed arrow represents an effect in the **opposite** direction predicted by the independence model, and therefore consistent with the “false alarms hurt” model

Table 2. Summary of reliance-compliance effects.

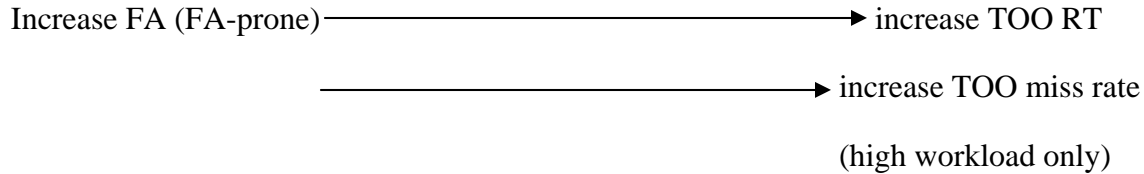
COMPLIANCE EFFECTS

EMPHASIS:

ON TOO TASK

On concurrent task

On automated task



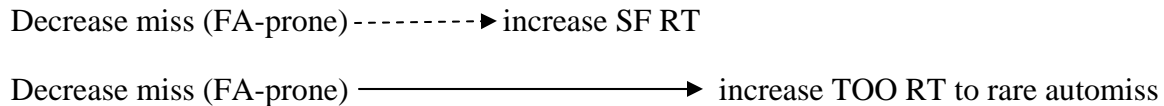
ON SF TASK

No Effects

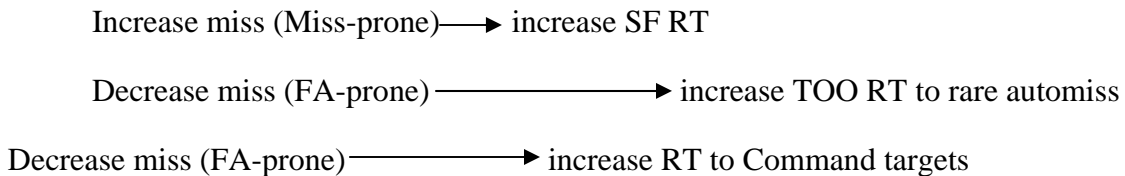
RELIANCE EFFECTS

EMPHASIS:

ON TOO TASK



ON SF TASK



Note that in the last line of the table is presented the one effect of alert threshold setting on a third task, command target (CT) detection. This influence is positioned mid way between the two emphasis conditions, as it was manifest in both conditions. Its effect is also considered to be both on an automated task and a concurrent task because, although command target detection was not automated, its information did lie within the same 3D image window as that where the TOOs appeared, and so its performance would be linked to reliance-driven attention strategies relative to that window.

The figure reveals the general dominance of solid (6) over dashed (1) arrows, indicating that the independence model of Table 1 is relatively well confirmed. In fact, the model is completely confirmed (not counting null effects) when the concurrent SF task is emphasized, and it is mostly confirmed when the TOO task is emphasized. However in this latter emphasis is observed the false-alert prone cost imposed on the concurrent task, not predicted by the independence model, but nevertheless consistent with the disruptive effects of false alarms on concurrent tasks that we have observed previously (Dixon & Wickens, in press; Wickens et al., 2005).

Using the data representation in Table 3 as a framework, we can then address two of the meta-issues that provided the rationale for this study. Recall that one purpose of the study was to examine whether the “independence model” deriving from the original formulations of reliance-compliance, or the “FA-hurts” model, observed more recently in empirical tests, better described the data both when there are increased demands of the automated task (to a perceptually challenging visual detection task), and when priorities are manipulated toward the automated task and away from the concurrent one.

First, in comparing across experiments, we note that here, with the more difficult automated task (ATR), the “FA-hurts” model still appears to represent the data better than the total independence model just as it did with the easier automation task in the previous studies.. We do note however, that the FA-hurts model is only directly supported when attention is focused on the difficult automated (TOO) task, just as attention was inferred to be more focused on the difficult (but then **not**-automated) TOO task in the previous research.

Second, when attention is directed away from the automated task (toward SF), as inferred to be the case in the prior research, here the FA-hurts model does not seem to account for the data. It is only when attention is directed toward the difficult (here TOO) task that the costs of false alarms to the concurrent (and neglected) SF task becomes apparent. Thus we can conclude that the “False alarm hurts” model, and its cost of false-alarm prone automation to the concurrent task, does not appear to be specific to paradigms using an easy automated task, nor to those in which the automated task is neglected, as true in the prior research.

It is of interest then to explain why a false alarm prone system should hurt the concurrent task. One explanation, offered by Dixon and Wickens (in press) which we have no reason to contradict here, is that the presence of false alarms represents salient evidence that the automation is faulty. The operators’ mental model then may signal that “faulty automation is bad” (e.g., reduced sensitivity, not just a shifted threshold) and can therefore will produce misses as well as false alerts, thus requiring greater supervision of the automated task, and reduced resources allocated to the concurrent tasks. In the current data, this shift of resources to the raw data of the automated task was not apparently great enough to fully offset the longer response time and greater miss rate of the rare automation misses.

The second meta-issue addressed by the current experiment is whether the general trend observed in prior literature (Wickens & Dixon, in press) for the automated task to bear greater costs of automation imperfections than concurrent tasks, is because the latter is usually assumed to be “secondary”, and the non-automated concurrent tasks to be “primary”. If this causal explanation were true, then in the current study, as we direct attention away from the non-

automated concurrent task, toward the TOO task, costs of automation imperfections should grow. We can answer this question within Table 2 by comparing imperfection costs on the concurrent task (middle column) when it is emphasized (SF emphasis) and when it is not (TOO emphasis). Here the data indicate that the costs of decreasing reliance to the concurrent task (increase in SF RT) are equally distributed across both emphasis conditions, thereby providing no evidence to support the “intrinsic priority” explanation of the asymmetry of automation imperfection effects.

Finally, it is important to note that, while all conditions here had imperfect automation, compared with baseline performance collected in an earlier experiment. Performance was generally improved. This suggests a general pattern of results: automation was always **depended upon** (high automation dependence). This dependence was differentially manifest in reliance and compliance, contingent upon the alert threshold setting. The costs of high reliance on imperfect automation were generally more than offset by its benefits.

References

- Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology*, 13(3), 249-268.
- 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. R., & Wickens, C. D. (in press). Automation reliability in unmanned aerial vehicle flight control: Evaluating a reliance-compliance model of automation dependence in high workload. *Human Factors*.
- Goh, J., Wiegmann, D. A., Madhavan, P., & Wong, J. (2004). Effects of spatial cueing errors on trust and reliance. Paper to be presented at the 112th Annual Meeting of the American Psychological Association. Honolulu, HI.
- Gugerty, L., & Brooks, J. (2001). Seeing where you are heading: Integrating environmental and egocentric reference frames in cardinal direction judgments. *Journal of Experimental Psychology: Applied*, 7(3), 251-266.
- Levinthal, B., & Wickens, C. D. (in preparation).
- Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, 45(2), 281-295.

- 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*.
- Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44-64.
- 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. (in press). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Sciences*.
- 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. *Proceedings of the 13th International Symposium on Aviation Psychology*. Dayton, OH.
- Wiegmann, D., McCarley, J., Kramer, A., & Wickens, C. D. (in preparation).