



**Aviation Human Factors Division
Institute of Aviation**

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

**Pilot Dependence on Imperfect
Diagnostic Automation in Simulated
UAV Flights: An Attentional Visual
Scanning Analysis**

**Christopher D. Wickens, Stephen R. Dixon,
Juliana Goh, & Ben Hammer**

**Technical Report
AHFD-05-02/MAAD-05-02**

March 2005

Prepared for

**Micro Analysis and Design
Boulder CO**

Contract ARMY MAD 6021.000-01

Pilot Dependence on Imperfect Diagnostic automation In Simulated UAV Flights:
An Attentional Visual Scanning Analysis

Christopher D. Wickens, Stephen R. Dixon, Juliana Goh, and Ben Hammer

Abstract

An unmanned air vehicle (UAV) simulation was designed to reveal the effects of imperfectly reliable diagnostic automation – a monitor of system health parameters – on pilot attention, as the latter was assessed via visual scanning. Four groups of participants flew a series of legs under different automation conditions: a baseline (no automation) control, and automation which was either 100% reliable, 60% reliable with a low-threshold bias to produce false alerts, and 60% reliable with a high threshold to produce misses. Visual scanning was recorded to assess the effects of reliance and compliance on the allocation of visual attention. A high workload mission completion task and ground surveillance task were simultaneously imposed. Consistent with the reliance-compliance model of imperfect automation developed by Meyer (2001), miss-prone automation removed visual attention from the surveillance task, while FA-prone automation delayed the alert-driven attention shift to the system monitoring task.

Introduction

Unmanned air vehicles (UAV) have realized a recent successful history in military aviation, and presently are forecast to play an important role in civil aviation. UAVs, almost by definition, will require high levels of automation, and hence bring into play issues of pilot monitoring of that automation. Whether the pilot is called on to supervise a single UAV, or two or more UAVs, as envisioned in many military applications (Dixon, Wickens & Chang, 2005), there are two major factors that mitigate the effectiveness of automation in UAV control.

The first factor is the level of “**workload**” experienced by the human operator. Here we define workload as the load imposed on the limited information processing resources of the unaided (without automation) human operator, in what we describe as the “baseline” or “manual” condition. This load can be imposed from two qualitatively distinct sources: the single task **difficulty** of the task that might otherwise be automated, and the **multi-task load** in which the baseline (vs. automated) task is performed. In these two cases, the automation benefits are likely to increase, to the extent that the single task to be automated is more difficult (Maltz & Shinar, 2003; Dixon & Wickens, 2004, in press), or that concurrent or multi-task load is imposed (Parasuraman et al., 1993).

The second factor is automation **reliability**. There is little doubt that total human-system performance will be quite good if automation is perfect. Conversely, when performing a difficult task, performance will be poor when automation is so unreliable as to be useless. However in between these extremes, lies a range of reliability levels where the benefits of automation over the baseline may be uncertain (Wickens & Dixon, 2005).

Of course there are a wide array of types of automation that can be employed to assist the UAV pilot, as well as a wide variety of ways in which automation can fail. In the current research we focus on automated alerts, that are of particular value under high levels of pilot workload, because the attention-grabbing properties of such alerts typically relieve the pilot of continuous visual monitoring of the “raw data” in the “alerted domain” (Pritchett, 2001). In our particular domain, the raw data represent indicators of the health of various systems on board the aircraft.

Three reasons lay behind our selection of this automated task for our research. First, because system monitoring is generally lower on the pilot’s task hierarchy (Schutte & Trujillo, 1996), it is logical to relegate this to an automated alert system. Secondly, interviews with subject matter experts of the Army’s Hunter-Shadow UAV (Wickens & Dixon, 2002), revealed the plausibility of rendering such system failures as relatively frequent events, and therefore legitimate subjects of an experimental inquiry of imperfect automation. Finally, the nature of potential automated failures in monitoring system events generalizes to a much wider class of automated diagnostic systems, such as conflict and collision alerts (Bliss, 2003; Pritchett, 2001), or automatic target recognition (ATR) systems (Maltz & Shinar, 2003), so that lessons learned regarding the implications of this imperfect automation for pilot attention and decision, can be widely applied.

Underlying our current modeling approach is the fact that automated diagnostic systems must discriminate two kinds of events: a “failure” (or dangerous event) and a “normal operating

condition”. When asked to make such a discrimination in a probabilistic imperfect world, with potentially unreliable sensors, automation will make occasional errors. It is then the responsibility of the alert designer to “set the threshold” of the alerting system to achieve the appropriate balance of alert misses, and alert false alarms. Generally, designers have chosen to bias this setting in favor of a low threshold, which generates many more false alerts, than it does missed events (Pritchett, 2001); however, neither type of automation error is immune from human performance costs, imposed on the pilot who must (a) respond to the alert output (if it is true, but not if it is false), (b) provide some attention to the “raw data” in the alert domain (to the extent that the alerting system may be miss-prone) and (c) perform a host of attention demanding concurrent tasks (see Figure 1).

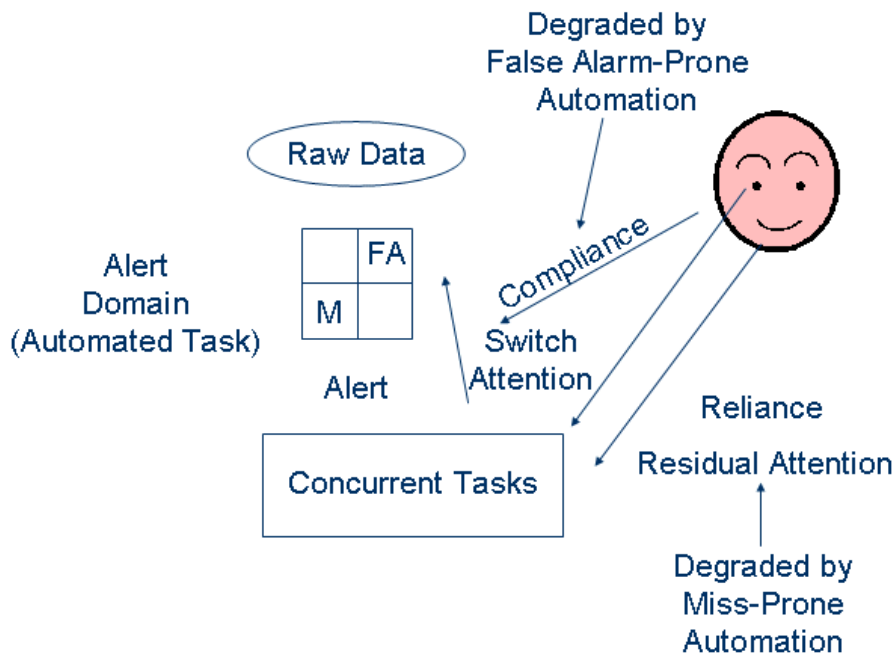


Figure 1. Attentional consequences of reduced reliance-compliance on primary and secondary tasks.

As we can see from Figure 1, visual and cognitive attention switches from the concurrent task to the automated task during alarms are discrete events, and should only occur when there is actually an alarm. When there is no alarm, the operator can assume that all is well, because he or she knows that there will be no events that are not accompanied by an alarm. Therefore, the operator can ignore the raw data behind the automation at all periods of time when there is no alarm. Only when there *is* an alarm does the operator have to switch attention to the raw data to verify the veracity of the alarm. On the other hand, to the extent that the automation is miss-prone, the effects on visual and cognitive attention are continuous because if the operator does not trust the automation to catch all the system failures, then he or she must do it themselves; that is, the operator must maintain a greater allocation of attention to the raw data behind the automation in order to catch all the failures that are not noticed by the automation. So with false

alarms, the operator only needs to switch attention to the raw data at each discrete alert event (whether true or false), while with misses, the operator must continuously put attentional resources into the raw data.

Some more specific description of what these costs are emerges from a treatment of alert systems developed by Meyer (2001, 2004) and Maltz and Shinar (2003) (also see Dixon & Wickens, in press), who distinguish between two cognitive states of human dependence on alerting automation: (a) **Reliance**, characterizes human cognition when the alert is silent. A reliant operator will assume that the alert will unfailingly sound when the raw data go out of tolerance, or a dangerous event occurs, and hence will have no need to examine those data while the alert is silent. Full residual attention will be available for concurrent tasks. However an imperfect alerting system that generates automation misses will reduce reliance, at the expense of visual attention to concurrent tasks. But because this reduced reliance avails more attention to the raw data, the automation misses will now be *more effectively* detected by the human operator/supervisor. (b) **Compliance**, in contrast, characterizes the operator response when the alert sounds. A highly compliant operator will rapidly abandon concurrent tasks and switch attention to the alerting domain once the alert sounds. However an imperfect alerting setting that generates many false alarms (the more frequently adopted type of threshold setting in most systems) will reduce compliance, even if this setting has minimal effect on reliance.

In a pair of UAV simulation experiments, Dixon and Wickens (2004, in press) varied the auditory alerting threshold as well as the overall reliability of system monitor gauges in their simulated UAV based on characteristics of the Army's Hunter/Shadow. Examining performance on the system monitoring task itself, along with performance of a concurrent image surveillance task, and a primary mission task, they were able to demonstrate performance effects that appeared to mirror some of the expected changes in reliance and compliance: increasing automation miss rate reduced concurrent monitoring; increasing automation false alert rate reduced pilot response to system failures. Both of these effects reflect the inferred influence of automation reliability on **pilot attention**, either to monitor concurrent tasks rather than the raw data (indexing high reliance), or to be immediately switched when an alert occurs (for a compliant pilot). While these inferences about attention are plausible, in the two studies described above, we could only infer attention from changes in performance, since we had no direct measures of attention.

Because of the critical role played by visual attention in aviation (Talleur & Wickens, 2003; Wickens, Goh, Helleberg, Horrey & Talleur, 2003), in the current study, we measured direct visual scanning indices of visual attention, as four groups of pilots monitored UAV simulations that varied in the reliability of the automated system status monitor: a 100% reliable system, an unreliable system ($r = 0.60$) with a bias to false alerts, an equally unreliable system ($r = 0.60$) with a bias to misses, and a baseline system with no auditory alerting whatsoever. In each system we measured performance, as well as the balance of visual attention between the system gauges and concurrent tasks (measuring miss-influenced reliance), and the visual attention switching time following an alert (measuring false-alert influenced compliance).

Methods

Forty 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 2 presents a sample display for a UAV simulation, with verbal explanations for each display window and task.

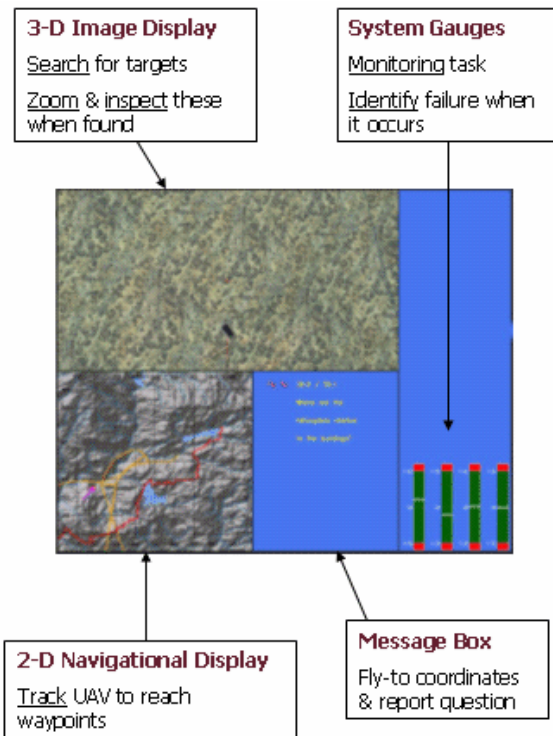


Figure 2. A UAV display with explanations for different visual areas.

As seen in Figure 2, 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. The experimenter used a separate keypad to record correct or incorrect responses and to indicate when the operator detected a target of opportunity (TOO) or a command target (CT).

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 4:1.

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.

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 (i.e., easy concurrent task) or during TOO and CT inspection (i.e., difficult concurrent task). 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 10 SFs, with never more than two SFs occurring during any mission leg. SFs were temporally separated by 4-15 minutes.

The SF task was the task served by automation. Automation aids, in the form of auditory auto-alerts during SFs, were provided for three out of the four conditions in this between-subjects design. The A100 condition (A = automation; 100% reliable) never failed. The A60f condition (f = false alarm; 60% reliable) was created by imposing three automation false alarms and one automation miss (4 automation failures), out of the 10 system failures that actually occurred; hence the reliability is considered to be 0.60. The A60m condition (m = miss; 60% reliable) resulted in more misses than false alarms (3:1 ratio). During a false alarm, the pilot was instructed to ignore the warning after cross-checking with the raw data to confirm the inaccuracy of the alarm. During a miss, the pilot was instructed that he or she was still responsible for “catching” the SF and correcting it. The final condition was a baseline condition (Man), with no automation aid to assist pilot performance.

Pilots were not aware of the precise level of reliability provided by each automation aid; however, depending on the experimental condition, they were told that the automation was either “fairly reliable” or “not very reliable”, as well as, in the latter case, the bias setting (i.e., more false alarms or more misses).

Eye-tracking data were collected from each pilot during the mission.

Results

For the most part, three planned comparisons were used to assess statistical effects: a) Baseline vs. the combination of A67f and A67m in a planned comparison (i.e., weights of -1, 0.5, 0.5), b) Baseline vs. A100, and c) A67f vs. A67m. Because only three a priori comparisons were made, familywise error rates were not adjusted (see Keppel, 1982, for more details). Two subjects (one in the A60f condition and one in the A60m condition) were removed due to corrupted data files. Note that because of these missing data, and unavoidable missing data points (e.g., if a target doesn't come into view on the 3D display, then a participant has no chance to detect it; or if a participant does not detect a target, then there are no data for the target detection times), the degrees of freedom in the following comparisons are sometimes less than the maximum value. Table 1 presents an overview of the data. Note that high workload for the TOO detection task refers to those times when a TOO entered the 3D image window while the pilot was engaged in responding to a system failure, while high workload for the SF task refers to those times when the SF occurred while a pilot was zooming and inspecting a target (TOO or CT).

Table 1. Overview of all performance measures. Standard errors are in parentheses.

	Baseline	A100	A60f	A60m
Tracking Error (MAE - meters)	82.65 (0.41)	82.65 (0.28)	84.62 (1.09)	83.09 (0.87)
Number of Repeats (per leg)	8.38 (2.45)	6.43 (0.95)	14.86 (4.29)	16.86 (4.83)
CT Detection Time (secs)	5.68 (1.60)	7.34 (1.53)	5.44 (1.83)	7.33 (1.47)
TOO Detection Rate (%) (Low workload)	88.75 (3.00)	82.00 (6.00)	75.00 (5.00)	61.00 (14.0)
TOO Detection Rate (%) (High workload)	37.50 (16.0)	42.86 (13.0)	57.14 (17.0)	28.57 (18.0)
TOO Detection Time (secs) (Low workload)	6.05 (0.87)	6.50 (1.38)	7.64 (1.89)	10.10 (1.51)
TOO Detection Time (secs) (High workload)	9.89 (2.56)	7.46 (3.45)	8.66 (2.97)	12.03 (1.23)
SF Detection Rate (%) (Low Load)	100 (0.00)	100 (0.00)	100 (0.00)	100 (0.00)
SF Detection Rate (%) (High Load)	91.67 (3.00)	100 (0.00)	41.67 (14.0)	58.33 (12.0)
SF Detection Time (secs) (Low Load)	7.19 (1.18)	2.18 (0.18)	3.02 (0.95)	3.36 (0.63)
SF Detection Time (secs) (High Load)	11.46 (2.04)	4.82 (0.93)	23.28 (2.18)	14.77 (2.69)

Mission Completion.

Planned comparisons revealed no significant differences in tracking error, [all $p > .10$]; however, a post hoc comparison between the A60f and Baseline, [$t(13) = 1.79$, $p < .05$], and between the A60f and A100 condition, [$t(13) = 1.86$, $p < .05$], indicated higher tracking error for the condition with false alarms, although it is not clear that this difference of less than 3% has any practical significance.

Planned comparisons revealed a significant increase in the use of the Repeat button (to refresh flight instructions) for the unreliable automation conditions over Baseline, [$t(13) = 1.81$, $p < .05$]. The other two comparisons revealed no significant differences, [all $p > .10$]. This indicates that pilots found it more difficult to retain information in memory due to having to put more attention into the systems monitoring task.

Targets of Opportunity (TOO) and Command Targets (CT)

TOOs occurred in both low workload (with no other task besides simple tracking) and, less frequently, in high workload (with a concurrent SF task). TOO detection rates refer to what

percentage of TOOs that pilots were able to detect. Figure 3 presents TOO detection rates across condition.

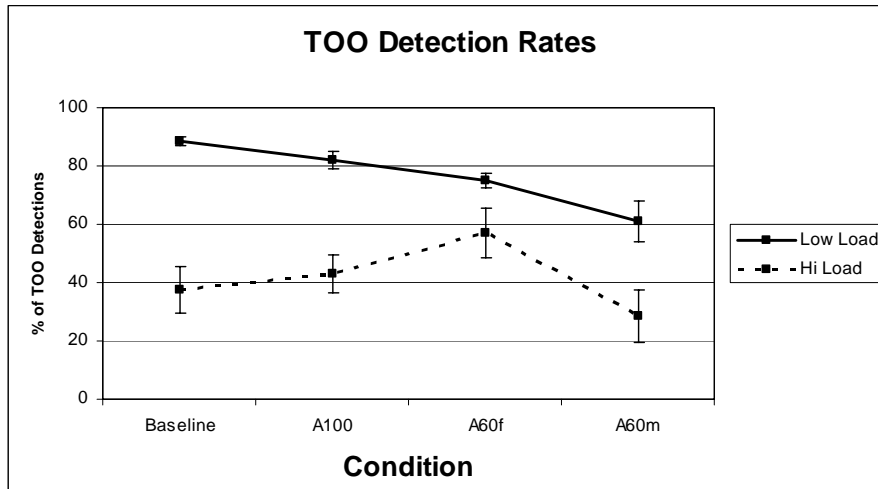


Figure 3. TOO detection rates. SE bars are included. The large SE bars for the high workload trials reflects the small number of such trials that occurred.

As shown in Figure 3, there appear to be effects for both low and high workload situations. In the low workload trials, the Baseline condition produced significantly more TOO detections than the unreliable automation conditions, [$t(13) = 2.67, p < .01$]. The other two comparisons were not significant, [Baseline vs. A100: $t(14) = 1.01, p > .10$; A60f vs. A60m: $t(12) < 1.0$]. In high workload trials, there were no significant differences between any of the conditions. A post hoc analysis of variance for workload x condition for the A60f and A60m conditions revealed a significant effect of workload, [$F(1, 12) = .73, p = .05$], but not of condition, [$F(1, 12) = 1.56, p > .10$], with no significant interaction, [$F(1, 12) < 1.0$].

TOO detection times refer to how long it took pilots to detect the TOO once it had entered the 3D display, as shown in Figure 4.

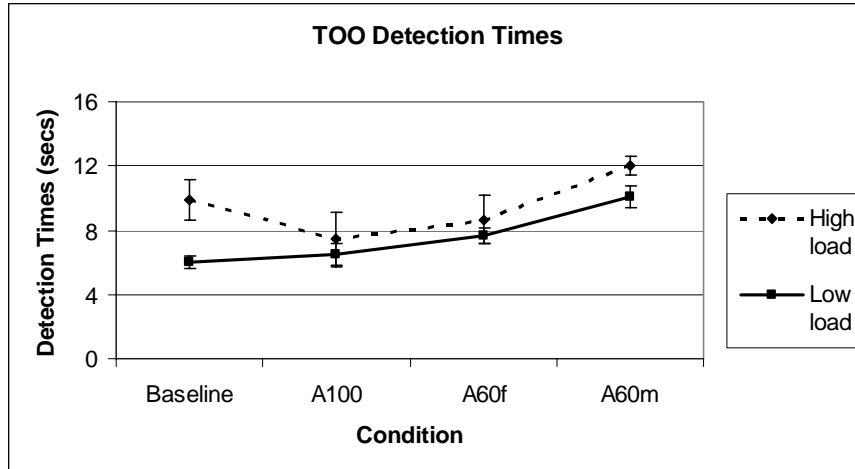


Figure 4. TOO detection times. SE bars are included.

Figure 4 reveals that, in low workload trials, there is a trend towards increased detection times as the automation becomes unreliable. Planned comparisons indeed showed that the unreliable automation conditions resulted in longer detection times relative to Baseline, [$t(13) = 2.16, p < .05$]. The other two planned comparisons were not significant. In high workload trials, there were no significant differences between any of the conditions, [all $p > .10$]. A post hoc analysis of variance for workload x condition for the A60f and A60m conditions revealed a marginally significant effect of workload, [$F(1, 6) = 4.34, p = 0.08$], but no effect of condition, [$F(1, 6) < 1.0$], and no interaction, [$F(1, 6) = 2.71, p > .10$].

As with TOOs, command target (CT) detection times were measured by how long pilots took to detect the CT once it entered the 3D display. Planned comparisons revealed no significant differences between any of the conditions, [all $p > .10$].

System Failures (SF)

System failure detection performance was measured by both SF detection rates and detection times. Figure 5 presents the SF detection rates as a function of workload.

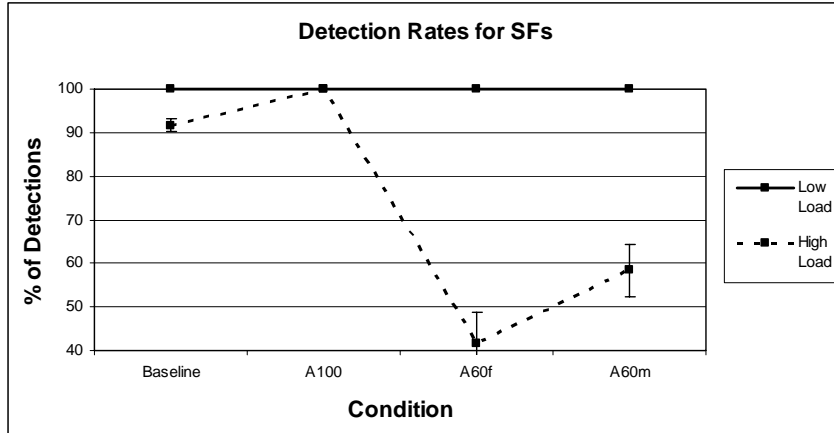


Figure 5. SF detection rates. SE bars are included.

As can be seen in Figure 5, there were no differences in performance between conditions during the low workload trials, so statistical analyses only focused on the high workload trials, when resources are assumed to be scarce. Planned comparisons revealed the Baseline condition performed better than the unreliable automation conditions, [$t(13) = 3.01, p < .01$]. There was no significant difference between the Baseline and 100% reliable conditions, [$t(14) = 1.0, p > .10$]. The false alarm condition performed worse than the miss condition, [$t(12) = 1.96, p < .05$], indicating that false alarms had a more detrimental effect on SF detection performance than did misses. This finding replicates the *trends* (not statistically significant) seen in previous UAV studies (Dixon & Wickens, 2004), where false alarm prone automation always seems to result in fewer SF detections.

Figure 6 presents SF detection times as a function of workload.

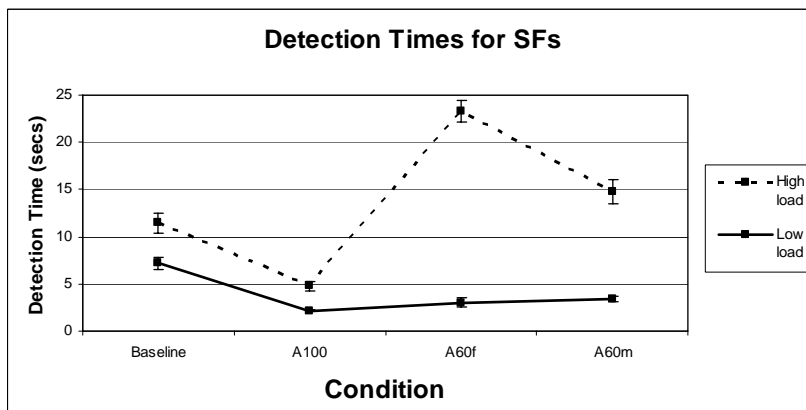


Figure 6. SF detection times. SE bars are included.

Figure 6 reveals that in low workload the Baseline condition resulted in longer SF detection times than the unreliable automation conditions, [marginally significant: $t(13) = 2.23, p < .05$], as well as the 100% reliable condition, [$t(14) = 3.59, p < .01$]. The two unreliable conditions did not differ statistically, [$t(12) < 1.0$]. In the high workload trials, the Baseline condition performed better than the two unreliable automation conditions, [marginally significant: $t(13) = 1.50, p = .08$], but worse than the 100% reliable condition, [$t(14) = 3.01, p < .01$]. The false alarm condition resulted in longer detection times relative to the miss condition, [$t(12) = 3.55, p < .01$]. This data indicates that the false alarm condition not only resulted in fewer SF detections, but also in longer SF detection times than the other conditions. These findings are very similar to those found in previous UAV studies (see Dixon & Wickens, in press).

We note that each of the 60% condition means is actually composed of two different components: responses when an alert correctly sounded, and those when the alert failed to sound. Table 2 shows the resulting four means, within the high workload condition.

Table 2. Component means in the A60f and A60m conditions. SE is in parentheses.

		CONDITION	
		A60f	A60m
EVENT	Miss (failure)	28.67 sec (1.42)	18.10 sec (4.43)
	Alarm (correct)	19.16 sec (3.20)	5.71 sec (2.37)

As was shown in previous UAV studies (Dixon & Wickens, in press), the data in Table 2 reveal a clear slowing of response times when the automation failed to alert the SF event. This indicates that in both conditions, pilots appeared to rely on the automation and their performance detection suffered because of its failures. Noteworthy is the fact that this prolongation was significantly greater when the automation miss rate was low (A60f; 28.7 sec), than when the automation miss rate was high (A60m, 18.1 sec), thus clearly establishing the “complacency” effect, enhanced by increased reliance. During situations with correct automation alerts, pilots responded more rapidly with the miss-prone automation (mean = 5.71) than they did with the false-alarm prone automation (mean = 19.16), [$t(12) = 3.38, p < .05$], indicating the pilots’ more immediate compliance to the automation alert (Meyer, 2001) in the former condition.

Eye-Tracking Data

Percentage Dwell Time (PDT). Table 3 provides a measure of the percent dwell time (PDT) that the eyes spent within each of the four areas of interest (AOI) on the workstation. The data are only reported during steady state (low workload) monitoring, not during the high workload segments involving zooming and panning of the 3D image window to identify detected targets. It is during this low workload period that pilots **rely** upon automation to alert them if such a system failure occurs. These data are presented graphically in Figure 7.

Table 3. Percentage Dwell Time that visual fixation is spent for the four experimental conditions within each area of interest (AOI): 3D image display where the TOOs were located, the 2D navigation display, the System failure monitoring gauges, and Message Box.

	Baseline	A100	A60F	A60M
<u>AOI</u>				
3D (TOO)	50.0 (3.21)	58.7 (2.83)	56.4 (3.59)	45.5 (5.29)
2D	36.7 (2.74)	39.2 (2.45)	32.2 (5.19)	35.1 (4.38)
SF	13.0 (1.35)	5.7 (1.26)	11.3 (1.62)	18.6 (3.18)
MB	4.1 (0.75)	6.6 (1.34)	9.0 (1.64)	11.9 (2.64)

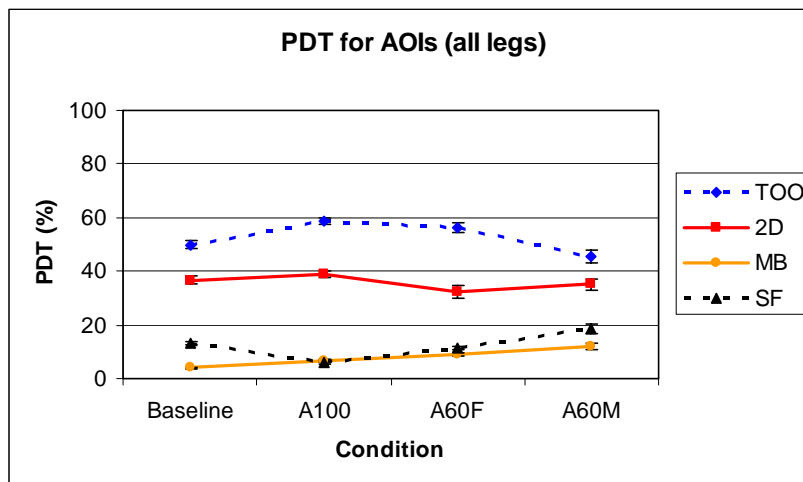


Figure 7. Percentage Dwell Time that visual fixation is spent for the four experimental conditions within each area of interest (AOI). SE bars are included.

Figure 7 reveals that there are changes in visual attention allocation based on what type of automated aid is available. A 2 way (AOI X condition) ANOVA carried out on the PDT data revealed a significant effect of AOI, $F(3, 78) = 155.75, p < .001$. The 3D image window, hosting the most demanding surveillance and detection task demanded the most visual attention, the 2D navigational display, hosting the most important task (command target location information) required around a third of the pilot's attention, and the two remaining AOIs demanded the least. Importantly, the significant AOI X condition interaction, $F(9,78) = 2.41, p = .05$, reflected automation reliance. Here we see that visual attention to the TOO window benefited (relative to the baseline value of 50%) from having auditory alerts, whether these were fully reliable (59%) [100A, $t(14) = 2.05, p < .03$], or imperfect, but having few misses (56%) [60F, $t(13) = 1.34, p = .10$]. However miss-prone automation drew as much, if not more, visual attention away from the 3D window (45.5%) as this window received in the baseline condition (50%). While this

decrease from the baseline was not significant, the difference between miss prone and false alarm prone automation was marginally significant [$t(13) = 1.7, p = .06$], reflecting the shift in visual attention away from concurrent tasks, fostered by a designer’s decision to increase the threshold in a way that produces more misses.

Scanning to the 2D image display did not differ significantly between conditions, indicating how pilots treated this display, which hosted primary task information, of utmost priority. However, scanning to the SF gauges themselves reflected an expected pattern, opposite to that of the 3D image window. While perfect automation ($A100 = 5.7\%$) greatly reduced the visual attention required, relative to baseline (13%) [$t(13) = 3.97, p < .01$], the miss-prone automation condition required far more visual attention to this display (18.6%), as expected given that pilots are, presumably, paying more attention to the “raw data” compared to the false alarm prone condition (11.3%) [$t(13) = 2.05, p = .03$], which did not differ from baseline. An additional feature is that pilots paid even more attention (18%) in the miss-prone condition, than in the non-automated baseline (13%, $t = 1.71, p < .05$), a cost that, as we saw above, bought them nothing in terms of better SF detection performance. There was no difference in scanning to the message box across conditions.

One might not have expected the false alarm rate to influence automation reliance, and indeed it did not appear to influence the measures of the residual attention to the 3D image window where the TOOs appeared. However, somewhat surprisingly, the higher FA rate *did* compel more attention to the SF display than the fully reliable automation condition, and induced no less attention there than the baseline condition. Thus no attention was “saved” by FA-prone automation relative to the baseline, in spite of the fact that nearly all failures were alerted. Thus, the general distrust induced by false alarms may have led to pilot suspicion that such a system requires further monitoring.

Mean Dwell Duration (MDD). Table 4 provides a measure of the mean dwell duration (MDD) that the eyes spent within each of the four areas of interest (AOI) on the workstation. As with the PDT data, these data are only reported during steady state (low workload) monitoring. These data are presented graphically in Figure 8.

Table 4. Mean Dwell Duration (MDD) for the four experimental conditions within each area of interest (AOI).

	Baseline	A100	A60F	A60M
<u>AOI</u>				
3D (TOO)	2.06 (0.21)	3.02 (0.34)	2.48 (0.20)	2.21 (0.35)
2D	1.95 (0.14)	2.14 (0.20)	1.76 (0.18)	2.02 (0.32)
SF	0.51 (0.05)	0.72 (0.08)	0.67 (0.04)	0.72 (0.18)
MB	0.84 (0.08)	0.87 (0.12)	0.95 (0.08)	1.29 (0.10)

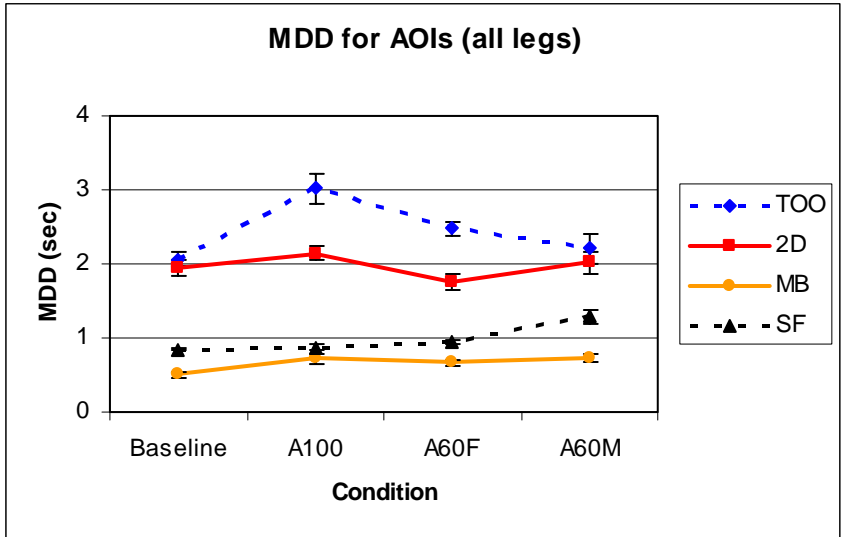


Figure 8. Mean Dwell Duration (MDD) for the four experimental conditions within each area of interest (AOI). SE bars are included.

A 2 way (AOI X condition) ANOVA carried out on the MDD data revealed a marginally significant effect of condition, $[F(3, 26) = 2.50, p = .08]$, a highly significant effect of AOI, $[F(3, 78) = 75.49, p < .001]$, and a non-significant trend towards an interaction, $[F(9, 78) = 1.56, p = .14]$. When comparing Figures 7 and 8, we can see the same pattern of visual behavior to the SF and TOO displays, as the beta criterion of the unreliable automation is adjusted from false alarms to misses.

The fact that the profile of the graphs in Figure 8 roughly parallel that in Figure 7, in the amount of time attention dwelt on each AOI when it was visited.

Visual Scan Response time. We inferred that compliance would be related to the speed with which visual attention moved to the SF gauges from wherever it was located at the time that the alert sounded. These measurements were computed by hand from a time-file of scanning across the 4 AOIs. The data for these “scan RTs” are shown in Table 5 when the alerts occurred during the high workload period while the pilot was engaged in image scanning. Figure 9 presents these same data in graphical form.

Table 5. Scan RTs in seconds (baseline scans represent the delay between the SF and the first look at the display. All others represent the delay between the auditory alert and the first look).

Baseline	A100	A60F	A60M
19.0 (2.00)	4.50 (1.13)	16.0 (2.78)	4.00 (2.48)

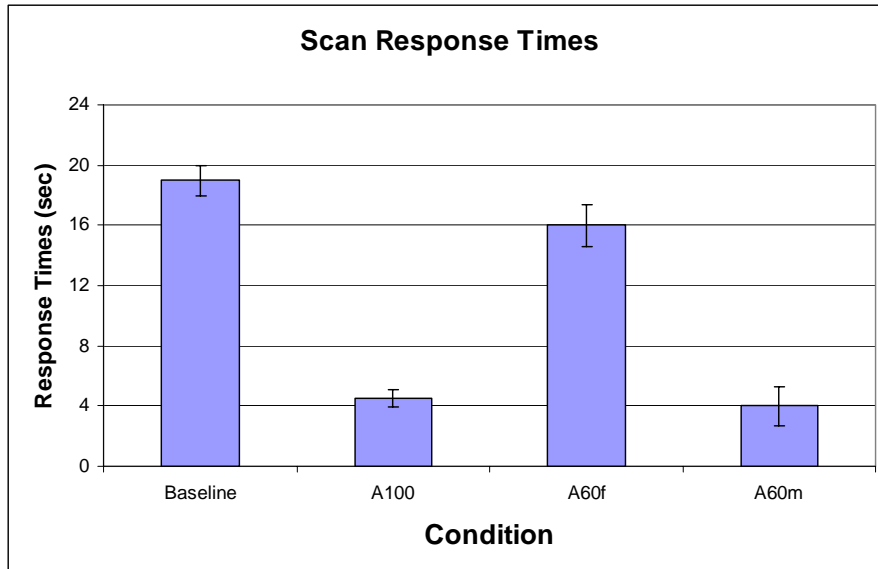


Figure 9. Scan RTs in seconds. SE bars are included.

A one way ANOVA on these data revealed a highly significant effect of condition, $F(3,29) = 5.806, p < .01$, revealing that looks were as rapid in the miss-prone condition, as in the perfect automation condition (pilots' perfectly complying with the alerts), but were as slow in the false-alarm condition as were the unaided glance response times.

Discussion

The current results extended the previous findings of imperfect diagnostic automation in UAVs (Dixon & Wickens, 2004, in press) to consider the explicit response of pilot visual attention, underlying the two inferred constructs of reliance and compliance. These two constructs characterize a pilot's dependence on automation that has a low miss rate and a low false alarm rate respectively.

As in the previous study, we found that an increasing miss rate produced a marginal loss in concurrent task performance. In the current data we noted that this loss was paralleled (and presumably caused) by a re-allocation of visual attention away from the 3D image window, toward the raw data hosted within the SF display (i.e., toward the oscillating bars representing system parameter health).

Also as in the previous study, we found that an increasing automation false alert rate, while having little effect on concurrent task performance (or attention allocation to the concurrent task), yielded a pronounced loss in SF detection performance in high workload, causing misses of some true alerts, and substantial delays in responding to all alerts. Interestingly, the increase in mean response time from the perfect automation condition to the A60F condition was 19 sec (Table 1b), whereas the increase in mean scan RT was only 11.5 sec (Table 3). Such a difference between the two measures indicates that, when false alarm rate was high, alert-driven looks to the display were followed by an additional 7.5 seconds of examining

the raw data to assure that the alert was indeed a true one, before an overt response was given. Overall, this delay, reflecting the cost of false-alarm prone automation, is of significant duration to be of considerable operational importance and could, for example, considerably compromise an operator's ability to intervene with a manual response to restore UAV stability.

The current data reinforces the notion that imperfect automation effects can be well modeled by their influence on pilot attention, and that such effects can be profound if automation reliability is allowed to drop to levels of around 60%, well below the threshold of approximately 70% reliability revealed to determine when automation is no longer useful (Wickens & Dixon, 2005). While such a reliability level may seem, at first glance, to be unrealistically low, it should be noted that in many aviation circumstances diagnostic automation is asked to **predict** events in a probabilistic world, plagued by future uncertainties in such variables as human response, or turbulence (Xu, Rantanen & Wickens, 2005; Thomas, Wickens & Rantanen, 2003; Krois, 1999). Under such circumstances, reliability rates not unlike those examined here, may be expected. It is therefore important that the consequences of these rates to pilot/supervisor performance are well understood.

Acknowledgments

This research was sponsored by subcontract #ARMY MAD 6021.000-01 from Microanalysis and Design, as part of the Army Human Engineering Laboratory Robotics CTA, contracted to General Dynamics. David Dahn was the scientific/technical monitor. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Army. The authors also wish to acknowledge the support of Ron Carbonari and Jonathan Sivier (in developing the UAV simulation).

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. 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 model of automation dependence in high workload. *Human Factors*.
- Dixon, S. R., Wickens, C. D., & Chang, D. (2005, in press). Mission control of unmanned air vehicles: A workload analysis. *Human Factors, 47*.
- 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.
- Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.

- Krois, P. (1999, July 25). *White Paper: Human factors assessment of the URET conflict probe false alert rate*. Washington, DC: Federal Aviation Administration.
- 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*.
- Parasuraman, R. M., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced "complacency". *International Journal of Aviation Psychology*, 3, 1-23.
- Pritchett, A. (2001). Reviewing the role of cockpit alerting systems. *Human Factors & Aerospace Safety*, 1, 5-38.
- Schutte, P. C., & Trujillo, A. C. (1996). Flight crew task management in non-normal situations. *Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 244-248). Santa Monica, CA: HFES.
- Talleur, D.A., & Wickens, C.D. (2003). The effect of pilot visual scanning strategies on traffic detection accuracy and aircraft control. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University.
- Thomas, L.C., Wickens, C.D., & Rantanen, E.M. (2003). Imperfect automation in aviation traffic alerts: A review of conflict detection algorithms and their implications for human factors research. *Proceedings of the 47th Annual Meeting of the Human Factors & Ergonomics Society*. Santa Monica, CA: HFES.
- Wickens, C. D., & Dixon, S. (2002). *Workload demands of remotely piloted vehicle supervision and control: (I) Single vehicle performance* (ARL-02-10/MAD-02-1). Savoy, IL: University of Illinois, Aviation Research Laboratory.
- 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., 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.
- Xu, X., Rantanen, E. & Wickens, C. D. (2005). Effects of conflict warning system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Proceedings of the International Symposium on Aviation Psychology*. Dayton, OH: Wright State University.