

PILOT DEPENDENCE ON IMPERFECT DIAGNOSTIC AUTOMATION  
IN SIMULATED UAV FLIGHTS: AN ATTENTIONAL VISUAL SCANNING ANALYSIS.

Christopher Wickens, Stephen Dixon, Juliana Goh and Ben Hammer  
University of Illinois, Aviation Human Factors Division  
Savoy, Illinois

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. 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, either as military UAVs must transition through civilian airspace, or as UAVs are called upon to perform non-military functions such a border surveillance or cargo transport. 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, as in most intended civilian applications, or two or more UAVs, as envisioned in many military applications, there are two major factors that mitigate the effectiveness of automation, in UAV control (as well as its effectiveness other aviation systems).

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), 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.

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”. 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 in aviation, such as conflict and collision alerts (Bliss, 2003; Pritchett, 2001), 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” 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), (b) provide some attention to the “raw data” (to the extent that the alerting system may be miss-prone) and (c) perform a host of attention demanding concurrent tasks.

Some more specific description of what these costs are, emerges from a treatment of alert systems developed by Meyer (2001, 2004; Maltz & Shinar, 2003), who distinguishes between two cognitive states of human dependence on alerting automation: **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, 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 setting that generates automation misses will reduce reliance, at the expense of visual attention to concurrent tasks.

**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 frequent type of setting) will reduce compliance, even if this setting has minimal effect on reliance.

In a pair of UAV simulation experiments, Dixon and Wickens (2004; Dixon Wickens and Chang, 2005, in press) varied the auditory alerting threshold as well as the overall reliability of system monitor gauges in their simulated UAV. 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). however we had no direct measures of the allocation of visual attention, as revealed through visual scanning measures. 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 these scan patterns as four groups of pilots monitored 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

39 student pilots from the Institute of Aviation volunteered to participate in the experiment. They were paid \$9.00/hour. Each pilot flew the UAV through ten different mission legs (one practice, 9 experimental), while completing three goal-oriented tasks commonly associated with UAV flight control: mission completion, target search, and systems monitoring. They used the interface shown in figure 1. At the beginning of each mission leg, pilots obtained flight instructions via the Message Box, including fly-to coordinates and a report question pertaining to the next command target (CT). These instructions were present for 15 seconds; in case the pilot forgot the instructions, pressing a repeat key refreshed the flight instructions for an additional 15 seconds.

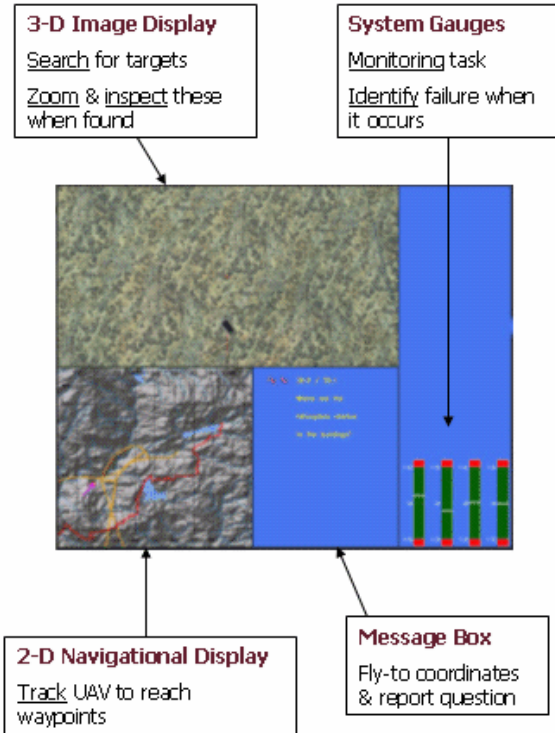


Figure 1: the Experimental Display.

Once pilots arrived at the CT location, they loitered around the target, manipulated a simulated 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?*). This challenging CT report demanded motor, visual and cognitive resources (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 camouflaged. TOOs could occur during simple tracking (low workload) or during a pilot response to a system failure as described below (high workload).

Concurrently, pilots were also required to monitor the system gauges for possible system failures (SF). This was the “automated task”. SFs were designed to fail either during simple tracking (low workload) or during TOO/CT inspection (high workload). The SFs lasted 30 seconds, after which the screen flashed bright red and a salient auditory alarm announced that the pilot had failed to detect the SF. There were a total of 10 SFs, with never more than two SFs occurring during any mission leg.

Automation aids, in the form of auditory auto-alerts during SFs, were provided for three of the four conditions. The **A100** condition (A = automation, 100% reliable) never failed to alert pilots of SFs. The **A60f** condition (f = false alarm, 60% reliable) failed by producing 3 false alarms and 1 miss out of the 10 SFs. The **A60m** condition (m = miss) failed by failing to notify pilots of a system failure on 3 of the 10 SFs, while generating a single false alarm. The final condition was a **baseline** manual condition, with no automation aid to assist pilot performance.

Pilots were not aware of the precise level of reliability of each automation aid; they were simply told that the automation was either “perfectly reliable” or “not perfectly reliable” and which way the threshold was set (i.e., whether the automation would produce false more false alarms or misses).

## Results

**Primary Task performance.** The pilots’ primary task was to fly the UAV to the command targets and make the report. Neither tracking accuracy nor CT report were much effected by automation reliability level, nor did this level effect pilot’s memory for the CT information (as implicitly measured by the use of the “repeat” key). Hence pilots optimally protected this most important task from resource competition imposed by other tasks.

**TOO monitoring.** Prior studies had shown that this “secondary” task of monitoring the 3D image window was sensitive to the demands imposed by imperfection of the automation (Dixon & Wickens, 2004). Table 1a shows performance on the TOO task as a function of condition.

We focused our analysis on TOO responses that only occurred under low workload conditions, in which a system failure had not occurred (i.e., during the period of reliance) and observed the trend in both accuracy and speed to be degraded with less reliable automation, particularly in the miss-prone condition [although this trend was not significant for RT, and only marginally so for detection rate ( $F_{3, 26} = 2.31, p=.10$ )].

Table 1. TOO and system failure monitoring/detection performance.

	Baseline	A100	A60F	A60M
(a) <u>TOO</u>				
(low workload)				
Acc (%)	89.00	82.00	75.00	61.00
RT (S)	6.05	6.50	7.64	10.10
(b) <u>System Failure</u>				
Low workload				
Acc	100.00	100.00	100.00	100.00
RT	7.19	2.18	3.02	3.36
High workload				
Acc	92.00	100.00	42.00	58.00
RT	11.46	4.82	23.28	14.77

**System Failure Detection.** Analysis of the system failure data revealed no effects of automation reliability at low workload, but that response times were faster when any type of automation was present,  $F(3,26) = 5.40, p < .01$ . Importantly, highly significant effects of reliability emerged at high workload, as revealed by the significant load X condition interaction in both Accuracy,  $F(3,26) = 7.91, p < .01$ , and RT,  $F(3,26) = 9.65, p < .01$ . Our particular interest was in the differential cost between miss-prone and false alarm-prone automation, where, in the high workload condition, both accuracy ( $t=1.96, p=.04$ ), and RT ( $t=3.53, p<.01$ ) demonstrated a greater cost in the false-alarm prone condition than in the miss-prone condition.

Thus the emerging picture is one in which performance on both tasks suffers when automation reliability degrades, but SF performance degrades more severely, particularly in high workload, and with false-alarm prone automation, whereas monitoring of the 3D image window for TOOs degrades only slightly, and even then only in the miss-prone condition. Thus we now ask whether visual scanning behavior, a direct manifestation of attention allocation and switching, can provide any insight as to the role of reliance and compliance in mediating the above effects.

**Visual Attention allocation.** Table 2 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.

Table 2. 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	58.7	56.4	45.5
2D	36.7	39.2	32.2	35.1
SF	13.0	5.7	11.3	18.6
MB	4.1	6.6	9.0	11.9

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 nav 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 baseline 50%) from having auditory alerts, whether these were fully reliable [100A,  $t(14) = 2.05, p < .03$ ], or imperfect, but having few misses [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 significant [ $t(13) = 1.7, p = .06$ ], indicating the shift in attention to concurrent tasks, fostered by a designer's decision to change the alerting threshold.

Scanning to the 2D image display, hosting information for primary task navigation performance did not differ significantly between conditions, indicating how pilots treated this display which hosted primary task information, as 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) greatly reduced the visual attention required, relative to baseline [ $t(13) = 3.97, p<.01$ ], the miss-prone automation condition required far more visual attention to this display, as expected given that pilots are, presumably, paying more attention to the "raw data" compared to the false alarm prone condition [ $t(13) = 2.05, p=.03$ ], 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 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.

**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 occurred. These measurements were computed by hand from a time-file of scanning across the 4 AOIs. The data for these “scan RT’s” are shown in Table 3 when the alerts occurred during the high workload period while the pilot was engaged in image scanning:

Table 3. 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
19s	4.5s	16 s	4.0s

A one way ANOVA on these data revealed a highly significant effect of condition,  $F(3,29) = 5.806, p = .004$ , 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 times.

### Discussion

The current results extended the previous findings of imperfect diagnostic automation in UAVs (Dixon & Wickens, in press) to consider the explicit response of pilot attention, underlying the two inferred

constructs of reliance and compliance. These two constructs characterize a pilot’s response to 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 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 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 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 significant operational importance.

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 rates 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. (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.
- 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 40<sup>th</sup> 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 47<sup>th</sup> 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.