



**Aviation Human Factors Division
Institute of Aviation**

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

**Reliability in Automated Aids for
Unmanned Aerial Vehicle Flight Control:
Evaluating a Model of Automation
Dependence in High Workload**

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

**Technical Report
AHFD-04-05/MAAD-04-1**

March 2004

Prepared for

**Micro Analysis and Design
Boulder CO**

Contract ARMY MAD 6021.000-01

Reliability in Automated Aids for Unmanned Aerial Vehicle Flight Control: Evaluating a Model of Automation Dependence in High Workload

Stephen R. Dixon & Christopher D. Wickens
University of Illinois at Urbana-Champaign

Abstract

Twenty-four participants flew a simulated unmanned aerial vehicle (UAV) through ten mission legs while searching for targets of opportunity and monitoring system parameters. Participants were assisted by automation which provided auditory alerts in response to system failures (SF). The auto-alerts were either 80% reliable or 60% reliable; the latter condition resulted in either a 3:1 ratio of false alarms to misses, or vice versa. Results indicated that the 80% reliable automation exceeded baseline (no automation) performance in the target search task. The two 60% reliable conditions did not provide any benefits to performance; both false alarms and misses hurt performance in the automated task *and* concurrent tasks of target monitoring. The current data, and those from a prior study with different reliability values, are evaluated against a computational model of automation compliance and reliance, and it is observed that these two effects are linearly related to automation false alarm and hit rate, and are relatively independent of each other. Implications for this study suggest that automated aids must be fairly reliable to provide global benefits, and data regarding the relative costs of misses versus false alarms on performance were equivocal, suggesting qualitatively different sorts of effects of reliance and compliance.

1.0 Introduction

Imagine one person supervising the mission requirements of two, four, or even a dozen aircraft simultaneously. This might not be too difficult if each aircraft is manned by an expert pilot who can handle the mundane tasks of the mission. However, when the aircraft are not piloted, as is the case with UAVs (unmanned aerial vehicles), the supervising task now also involves navigating the UAV, monitoring craft parameters for possible problems and correcting those problems when they occur during a mission, searching for possible targets, analyzing the targets upon detection, and reporting details of the targets to mission command (Dixon, Wickens, & Chang, 2003). These tasks are required of UAV pilots for *each* aircraft.

The military is currently employing different forms of automation to aid pilots in these tasks; however, very few automated aids are perfectly reliable. What happens when the automation fails and how does this affect pilot trust and human-automation performance? Previous research suggests that imperfect automation creates different states of overtrust, undertrust, or calibrated trust (Parasuraman & Riley, 1997), “cognitive laziness” (Skitka et al, 1999), “complacency” (Parasuraman, Molloy, & Singh, 1993), and performance loss (Dixon & Wickens, 2003; Molloy & Parasuraman, 1996; Rovira & Parasuraman, 2002; Rovira, Zinni, & Parasuraman, 2002; Wickens & Xu, 2002).

The literature is ambiguous on the issue of how unreliable the automation needs to be to cause performance to drop below that of no automation at all. Lee and See (in press) suggest that this level may be approximately 70-75%. Merlo et al. (2000) found benefits for target cueing at 70% reliability levels. Kantowitz et al. (1997) also found benefits at the 70% level for drivers who relied on highway traffic information. Galster et al. (2001) found benefits as low as 67% reliability, while Yaacov et al. (2002), and Lehto et al. (2000), both found benefits at 60%. Maltz and Meyer (2001) observed reliability benefits at approximately 90% in a signal detection task, but non significant costs (relative to the control condition) at 70%. Rovira et al. (2002) found some benefits to the concurrent tasks even at 50%. Still others found no benefit (Yeh & Wickens, 2001) or costs (Dzindolet et al., 1999) at 75% for target cueing. Maltz and Shinar (2003) observed costs of unreliable cuing across a range of reliability levels from as high as 90%, when the non-automated, uncued version of the task was relatively easy.

Most of the studies described above, with the exception of Rovira et al, were conducted in single task conditions. Dixon & Wickens (2003), the only study to focus specifically on both UAVs *and* a multiple-task environment, found benefits for an auto-pilot with 67% reliability, but costs for an auto-alerting system at the same reliability level. One possible reason for the discrepancy in values regarding this “cutoff point”, below which imperfect automation is worse than no automation at all, may relate to the distinction between trust and automation dependence on the one hand, and low and high workload on the other. Thus, under conditions of high workload, when resources are scarce (at a premium), an operator may depend upon imperfect automation, availing more resources to improve concurrent task performance even if the automation is not fully trusted, and such dependence will degrade performance of the automated task itself even as it helps concurrent tasks (e.g. Rovira et al., 2002). Consistent with this interpretation, Dixon & Wickens (2003), for example, found that the imperfect auto-alerts imposed more costs on the automated alerting task than on concurrent tasks, while finding no benefits to any of the tasks.

One model of automation developed by Parasuraman, Sheridan & Wickens (2000) outlines four stages of automation and how each of those stages processes data differently and provides certain unique automated benefits in reducing workload. These stages correspond to the four stages of human information processing. Stages 1-2 usually involve the functions of stimulus input, perception, synthesis, and analysis, while stages 3-4 assist response selection and response executions, respectively. Each of these stages offer different types of enhancements to the human-machine interface, and each mitigate performance differently.

Within stage 1-2 automation, unreliable aids will create false alarms and/or misses. A false alarm occurs when the aid indicates an event that does not actually happen in the real world. For example, the aid may infer that it has detected an enemy aircraft in the vicinity and gives an alarm to alert the pilot to the danger, when in fact there is no such enemy present. A miss occurs when the aid fails to notice a real event, as is the case when an enemy aircraft is present but the aid gives no warning. False alarms and misses have different consequences on pilot trust and human-automation performance. False alarms tend to cause distrust in the aid (Meyer & Ballas, 1997) and lead to ignoring the diagnosis, or a “cry-wolf syndrome” (Breznitz, 1983), while misses lead to reallocation of visual resources to the raw data in order to “catch” the automation miss (Cotté, Meyer & Coughlin, 2001).

Bliss (2003) noted that false alarm related incidents and accidents were still quite common in both civil and military aviation, and in some circumstances contributed to more aviation problems than did misses. Several other studies have also examined the different effects of false alarms and misses in a carefully controlled manner. For example, Gupta, Bisantz, & Singh (2001) had participants operate a virtual driving simulator, and measured operator reaction to an adverse condition warning system (ACWS), which indicated possible vehicle skids with a distinctive auditory alarm. Their results suggest that the alarms may have startled participants, and that false alarms had a more negative impact on operator trust than did misses. However, their experiment did not include true false alarms and misses; rather, a false alarm was defined as an early warning, while a miss was defined as a late warning. Lehto et al. (1997) also directly compared misses versus false alarms, in a decision aid that recommended (or failed to respond) drivers for safe passing on a two-lane road. However, the two bias conditions that they compared also differed in the overall sensitivity of the automated safe passing detection system, so that the advantage they observed of the miss-prone over the false alarm-prone system could have been attributed to its greater sensitivity, rather than to the setting of its decision criterion.

An important cognitive distinction, relevant to the contrast between miss-prone and false-alarm-prone stage 1 automation was introduced by Meyer (2001), in contrasting the terms “compliance” and “reliance”. (Meyer’s choice of the term “reliance” is slightly unfortunate here, because this term is also used in a broader sense of “dependence” on automation. We preserve his use of that term, and distinguish it from dependence). According to this dichotomy, **reliance** (in the Meyer sense) is when the operator depends on automation to alert him if there is a failure. A system with few (or no) misses will breed reliance. **Compliance** is when the operator responds immediately and infallibly to (complies with) the automated alert. A system with a low false alarm rate will breed compliance. Meyer (2001) has found that different experimental variables influence these two components of automation dependence. For example, close proximity of a

visual alert system to the raw data appears to breed compliance, even with an invalid alarm; whereas training and experience appears to breed more calibrated reliance.

While Meyer (2001) did not explicitly manipulate miss and false alarm rate, Cotté, Meyer, & Coughlin (2001) did so in experiments in which participants drove along an undivided single lane road in a virtual simulator, with warnings to brake when a possible dangerous event transpired. The authors concluded that the criterion setting with more false alarms relative to misses resulted in drivers having less compliance in the system when an alarm sounded, but more reliance that the system was safe when there was no alarm. The higher false alarm rate resulted in fewer responses to the alarm than the other condition, indicating reduced operator compliance. This reduced “compliance” was partly mitigated by providing visible causes of false alarms. Maltz & Shinar (2003) also examined the effects of miss and false alarm rate on compliance and reliance. They had participants search for military vehicles in both infrared and natural color pictures. Target cues were added to highlight possible targets, with nine levels of reliability, defined by conditions with 3 different probabilities of misses and conditions with 3 different probabilities of false alarms. Results indicated that increasing false alarm rates caused target detection performance to suffer along with a reduction in compliance, while increasing misses had no adverse affect on performance, and less of an effect on reliance. Cue dependence appeared to increase with the overall reliability of the cueing feature.

There seems to be a sort of “lore” abounding that false alarm prone systems, and their triggering of reduced compliance, are somehow worse than miss prone systems, with their fostering of reduced reliance. Surprisingly, however, this direct contrast appears to have been rarely analyzed in multi-task simulations where both performance and attentional demands of the two types of imperfection can be directly compared. The attention demands are particularly important in assessing reliance, since increased reliance upon automated alerts should avail more spare capacity in which to apply to the concurrent task. Dixon & Wickens (2003) made such a contrast by having pilots perform a high-fidelity UAV simulation under several conditions, including a baseline condition with no automation, a condition with perfectly reliable auditory auto-alerts to aid detection of system failures, a condition with 67% reliable auto-alerts that included false alarms (no misses), and a condition with 67% reliable auto-alerts that included misses (no false alarms). Pilots were required to fly to ten different command targets and report on the characteristics of those command targets. This involved reading and memorizing a set of fly-to coordinates and a report question about the command target (e.g. *Where are the helicopters located relative to the building?*). Pilots could refresh these flight instructions by pressing a repeat key. Simultaneously, they were expected to search for camouflaged targets of opportunity and report on them upon detection, while monitoring for possible system failures and correcting them when they occurred.

The results of this study revealed that the perfectly reliable auto-alerts benefited detection of the alerted system failures relative to baseline, but the two imperfect auto-alert conditions provided almost no benefit relative to baseline. In the false alarm condition, the only improvement was in system failure (SF) report accuracy, while target of opportunity detection times and SF detection rates both suffered relative to baseline performance during high workload situations, when the operator was involved in a concurrent 3D image manipulation task. From subjective ratings of trust provided by the pilots, the authors determined that the pilots were unable to effectively

calibrate their trust in the auto-alerts, rating the reliability of the aid to be much lower than it really was (36% vs. 67%). This poor calibration appears to have caused reduced compliance in the alerting system (Cotté et al, 2001), which resulted in pilots ignoring the automation aid during high workload trials when they were concurrently dealing with a target inspection. When an alarm went off, the operator probably assumed that it was yet another false alarm, and did not yield any visual resources from the target inspection task in order to double-check the system gauges to see if there really was a SF.

The miss condition also provided some benefit to SF report accuracy, but suffered relative to baseline for SF detection times during high workload, as well as the flight instruction recall task, as seen by the increase in the number of repeated requests for the command target instructions. In contrast to the finding by Gupta et al. (2001), pilots in the miss condition showed even less trust in the automation aid (20%) than in the false alarm condition (36%), suggesting that as pilots lost trust in the aid, they focused more visual attention on the SF gauges in order to catch the automation failures, indicating underreliance on the alerts. In fact, this reallocation of visual attention appeared to surpass even the baseline condition (with no aid), as reflected by the significant drop in concurrent task performance (detection of TOOs) below baseline.

In summary, many previous studies have found a higher penalty on performance for false alarms than for misses in single-task conditions (e.g. Cotté et al, 2001; Gupta et al, 2001; Maltz & Shinar, 2001); however, this was not the case for the earlier UAV study, which found a roughly equal penalty for both types of automation failures in a multiple-task environment. What is not known from these studies (or is ambiguous) is why the operator/pilot is so poor at calibrating trust, what exact level of reliability is needed to improve performance relative to baseline, what would happen when both false alarms and misses are included in a particular alerting aid during high workload (multiple-task) situations, and how attentional resources are allocated between tasks during system failures in a multiple-task environment.

While Dixon & Wickens (2003) used conditions with only false alarms or only misses, the current study included an 80% reliable condition with an equal number of false alarms and misses, as well as two 60% reliable conditions with a 3:1 ratio of false alarms to misses and vice versa. These two conditions are particularly important, as they allow us to determine how low compliance and reliance influence the response to both kinds of automation errors: automation misses and false alerts. Furthermore, participants in the current study were informed a priori as to the approximate level of reliability to see if they could calibrate their trust in the aid more affectively. We hypothesized that (1) 80% reliability would consistently improve performance above baseline, since it is above the “cutoff point” of 70% described earlier; (2) both 60% reliability conditions would degrade performance below baseline, since they are both below the “cutoff point” of 70%; (3) decrements due to unreliability would be more pronounced on the automated task than on concurrent tasks, a finding that would be consistent with the previous UAV study described above; and (4) miss-prone automation would disrupt concurrent tasks more than false-alarm prone automation, because of the former’s requirement for more continuous visual monitoring of SF status. A final objective of this study was to provide data which could, in conjunction with the data from Dixon & Wickens (2003), validate a computational model of operator dependence on and trust in imperfect automation.

2.0 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.

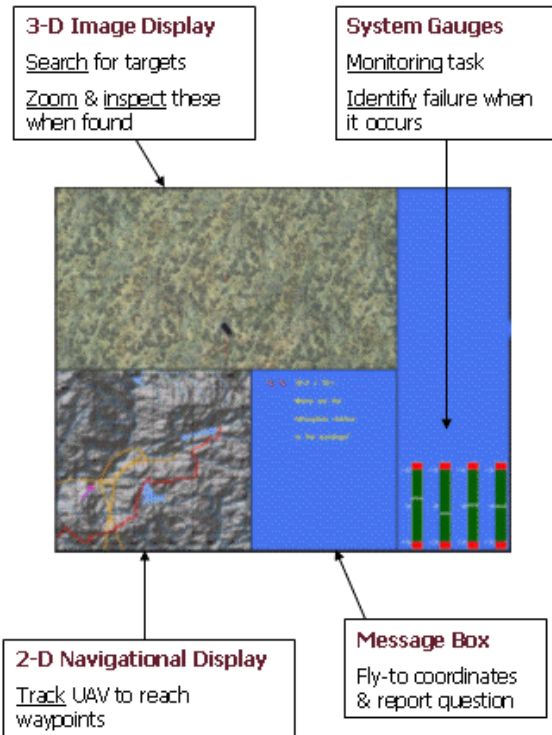


Figure 1. A UAV display with explanations for different visual areas.

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. 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. The A80 condition (A = automation; 80% reliable) failed by giving one false alarm (i.e. alarm with no actual SF), and one miss (i.e. a failure with no alarm) during each mission. These two automation failures occurring out of a possible 10 system failures that actually did occur, defined a $(1 - 2/10) = 0.80$ reliability level. The A60f condition (f = false alarm; 60% reliable) was less reliable than the A80 condition, and 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) was also less reliable than the A80 condition, but 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 their condition, they were told that the automation was either “fairly reliable” or “not very reliable”, as well as the bias setting (i.e. more false alarms or more misses). Ratings of subjective workload and trust were given by each pilot at the end of the mission. They answered the following questions regarding automation reliability: “*In general, how do you assess the trustworthiness of the auditory alarms?*”, “*how likely was the automation to provide an alarm when there was no SF (false alarm)?*”, and “*how likely was the automation to provide no alarm during a SF (miss)?*”

3.0 Results

One subject in the baseline condition was removed due to corrupted data. Table 1 presents an overview of the data.

Table 1. Overview of all performance measures. SE values are in parenthesis.

	MAN	A80	A60f	A60m
Tracking Error (MAE - meters)	84.25 (2.13)	84.45 (1.95)	82.75 (5.11)	85.76 (1.92)
Number of Repeats (per leg)	3 (1.12)	5.57 (1.72)	5.25 (1.65)	8.5 (1.59)
CT Detection Time (secs)	2.45 (.80)	1.96 (1.07)	4.16 (1.10)	4.11 (1.84)
TOO Detection Rate (%)	76 (5.0)	93 (7.4)	87 (7.1)	82 (7.2)
TOO Detection Time (secs) (High workload)	6.03 (1.99)	8.58 (2.82)	14.72 (2.63)	11.86 (5.51)
TOO Detection Time (secs) (Low workload)	6.04 (.91)	5.94 (1.28)	6.68 (1.20)	5.89 (1.24)
SF Detection Rate (%) (Low Load)	100 (1.9)	100 (2.8)	97 (2.7)	98 (2.7)
SF Detection Rate (%) (High Load)	79 (13.4)	69 (19.7)	50 (53)	75 (26)
SF Detection Time (secs) (Low Load)	2.17 (.48)	2.08 (.71)	2.50 (.19)	3.15 (.19)
SF Detection Time (secs) (High Load)	10.74 (2.25)	11.27 (3.31)	19.98 (3.19)	13.62 (3.20)
SF Report Accuracy (%)	88 (3.1)	97 (4.8)	98 (4.4)	94 (5.0)

3.1 Mission Completion

Tracking error was not affected by condition [$F(3, 27) = 1.24, p > .10$], a finding replicated in previous UAV experiments (Dixon & Wickens, 2003). As mentioned, pilots could refresh their memory of flight coordinates and report questions by pressing a repeat key. An ANOVA revealed that the number of repeats was affected by condition [$F(3, 25) = 3.56, p = .029$]. Further comparisons revealed that only the A60m condition (8.5 repeats) was statistically different from the baseline (3 repeats) condition [$t(12) = 3.09, p = .004$], replicating findings from Dixon & Wickens (2003). The A60m condition also performed worse than the A80 condition [5.57 repeats; marginally significant: $t(13) = 1.64, p = .09$] and the A60f condition [5.25 repeats; marginally significant: $t(14) = 1.59, p = .052$]. No other comparisons were significant [all $p > .10$]. This cost to the A60m condition indicates that the system gauges required more visual and cognitive resources from pilots in order to catch the frequent automation misses, to the detriment of recalling the flight instructions.

3.2 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 2 presents TOO detection rates across condition.

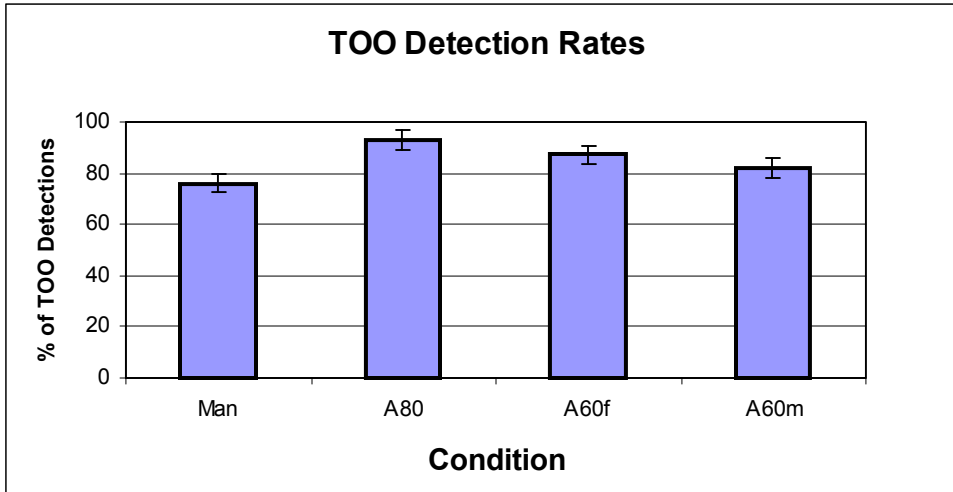


Figure 2. TOO detection rates across condition.

Planned comparisons between the four conditions revealed that only the A80 condition differed significantly from the baseline condition [$t(13) = 2.15, p = .025$], showing improved performance. All other comparisons were not significant [$p > .10$].

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 3. An ANOVA revealed a marginally significant main effect of condition [$F(3, 23) = 2.93, p = .055$], as well as a significant main effect of load [$F(1, 23) = 31.56, p < .01$]; however, an interaction between condition and load [$F(3, 23) = 4.82, p = .01$] indicates that the condition effect was only present at high load (i.e. when the pilot had engaged in SF detection and report).

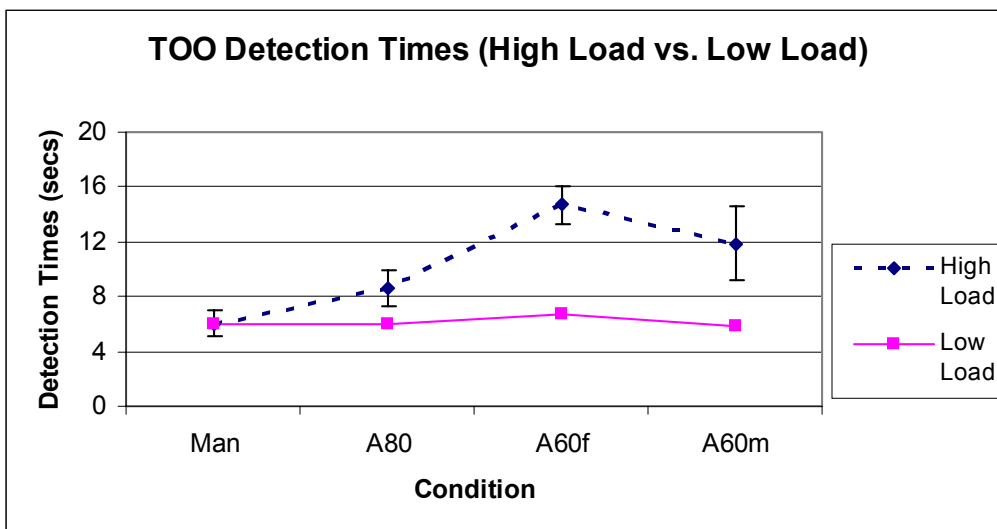


Figure 3. TOO detection times across condition and workload. SE bars are included.

Planned comparisons indicated that the A80 condition did not differ from baseline, whereas both the A60f and the A60m did so [$t(11) = 3.57, p < .01$; $t(10) = 2.04, p < .05$], respectively. While the two conditions did not differ significantly from each other [$t(11) = 1.04, p > .10$], the trend toward greater decrement with the A60f condition is consistent with Dixon & Wickens (2003), who found that the false alarm condition imposed a higher cost on concurrent TOO detection at high load than did the miss condition, and is probably due to pilot mistrust in the automation once it dropped below an “acceptable” level of reliability. This mistrust caused pilots to put more visual and cognitive resources into the SF task, to the detriment of the concurrent TOO task, causing them to take 6-8 seconds longer to detect the TOOs. Importantly, the current results suggest that the miss condition (here at 60% reliability) imposed a cost relative to baseline, whereas in Dixon & Wickens (2003), with 67% reliability, there was no cost.

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. Figure 4 presents CT detection times across condition.

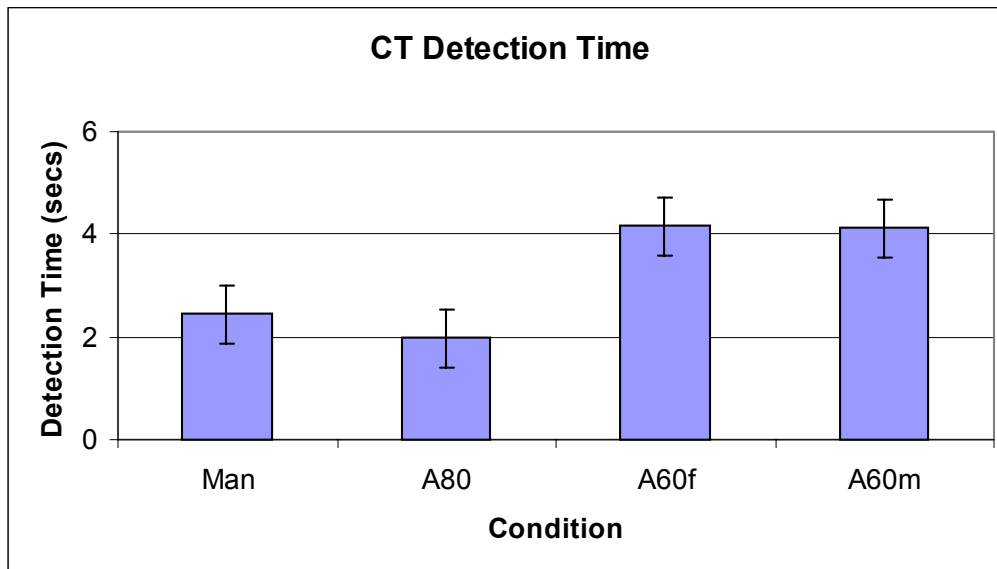


Figure 4. CT detection times across condition. SE bars are included.

An ANOVA revealed a main effect of condition [$F(3, 27) = 6.16, p < .01$]. Further comparisons reveal that both the A60f [$t(13) = 3.40, p < .01$] and the A60m [$t(13) = 2.21, p = .02$] conditions suffered relative to baseline, with an average increase of 2 seconds to reaction times. All other comparisons were not significant [$p > .10$]. Unlike Dixon & Wickens (2003), who found that auto-alerts with 67% reliability did not cause decrements in CT detection time performance, the current level of reliability (60%) appears to result in poorer performance relative to baseline for both types of automation failure. Condition did not affect accuracy or speed of CT report [all $p > .10$].

3.3 System Failures (SF)

An ANOVA on SF detection rates revealed that higher load reduced detection rates [F(1, 27) = 21.46]; however, there was no main effect of condition [F(3, 27) < 1.0], nor a statistical interaction between condition and load [F(3, 27) < 1.0]. Interestingly, the load effect described above is, in the A60M condition, constrained to the first automation failure (this is the only condition with more than one automation miss, and thus the only condition in which first failure effects could be assessed); that is, *if* a pilot missed a SF during the mission, and was informed of this by an alarm sounding to indicate that the pilot had missed the SF, they did not miss *any* SFs afterwards.

An ANOVA on SF detection times (including only the Man, A80, A60f, and A60m data) shown in Figure 5 reveals that higher load increased detection times [F(1, 27) = 93.3, p < .001]. The main effect of condition [F(3, 27) = 3.62, p = .026], can only be interpreted in the context of the interaction between load and condition [F(3, 27) = 3.06, p = .045], which reveals that the A60f condition suffered more due to high load than the other conditions.

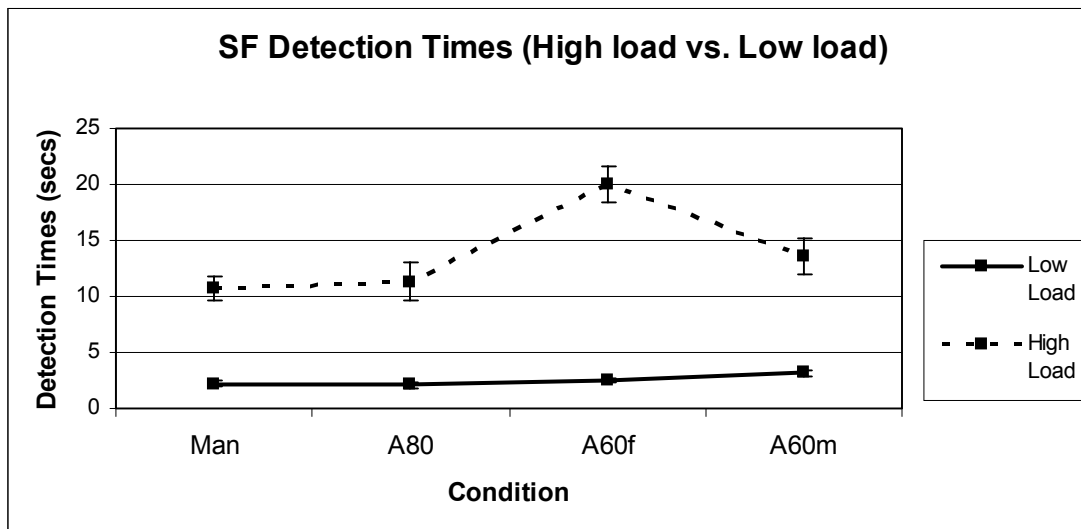


Figure 5. SF detection times across condition and workload

Figure 5 reveals that at high workload, neither A80 [t(13) < 1.0], nor A60m [t(13) < 1.0], were significantly longer than manual, while A60f was longer than manual [t(13) = 2.19, p < .05], A80 [t(14) = 3.08, p < .01], and A60m [t(14) = 2.16, p < .05]. 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)	26.05 sec (1.83)	23.29 sec (2.77)
	Alarm (correct)	13.93 sec (4.85)	3.96 sec (1.17)

The data in table 2 reveal the clear slowing for RT when the alarm “missed” the SF event, indicating that in both conditions, pilots had relied heavily upon the automation, and their detection suffered when it failed. Correct alerts were responded to more rapidly with the miss prone automation (mean = 3.96) than the false alarm-prone automation (mean = 13.93) [$p < .05$], reflecting the pilots’ immediate *compliance* with the auditory alert (Meyer, 2001) in the former condition, in contrast to the false-alarm prone condition, where pilots were less likely to interrupt target inspection to deal with the alarms. We also infer that greater compliance in the miss condition is coupled with an ongoing greater awareness of the SF gauges, fostered by a reduced *reliance* on that automation, and causing the greater disruption to memory recall described previously. The difference between reliance and compliance effects is explored in greater detail in the modeling section below.

The accuracy of the SF reports was not affected by load [$F(1, 23) < 1.0$] or condition [$F(3, 23) = 2.16, p > .10$].

3.4 Subjective Measures

3.4.1. *Subjective Workload.* At the end of each mission, pilots were asked to give subjective ratings of workload for the various tasks. There were 9 tasks which were rated: Initial heading, monitoring course, reading and recall of flight instructions, monitoring for SFs (visual aspect), monitoring for SFs (auditory aspect), correcting and reporting SFs, searching for TOOs, reporting TOOs, and reporting CTs. Figure 6 presents these subjective ratings across task.

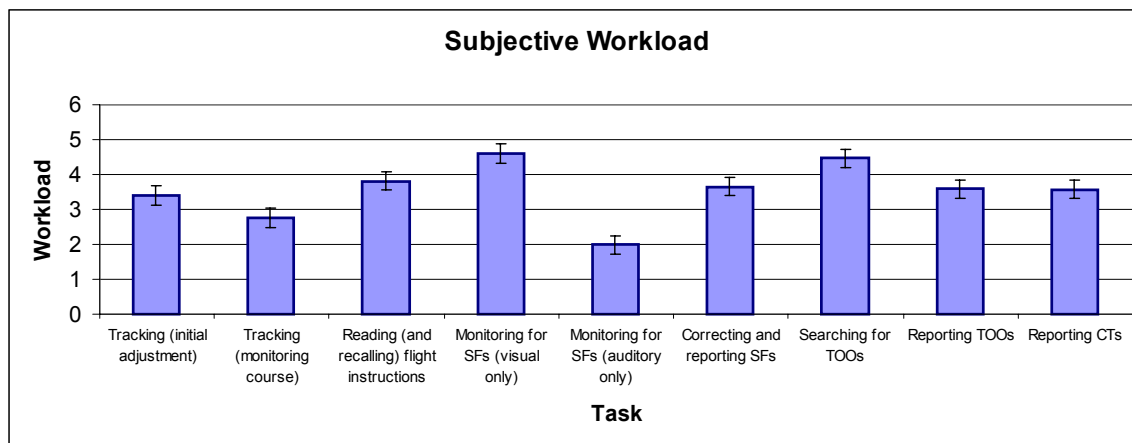


Figure 6. Subjective workload ratings across task. SE bars are included.

An ANOVA revealed a main effect of task [$F(8, 152) = 13.81, p < .01$]; however, there was no main effect of condition [$F(2, 19) < 1.0$], nor was there an interaction between condition and task [$F(16, 152) = 1.47, p = .12$], indicating that pilots in each condition found the tasks to be similar in difficulty; that is, no condition facilitated a perceived reduction in workload.

As with previous UAV studies (e.g. Dixon & Wickens, 2003), pilots appeared to find visual SF monitoring and TOO target search/report to be the most difficult tasks, while auditory SF monitoring was considered the easiest task.

3.4.2. *Subjective ratings of trust.* Pilots were surprisingly accurate in their overall assessment of the automation reliability, as seen in figure 7.

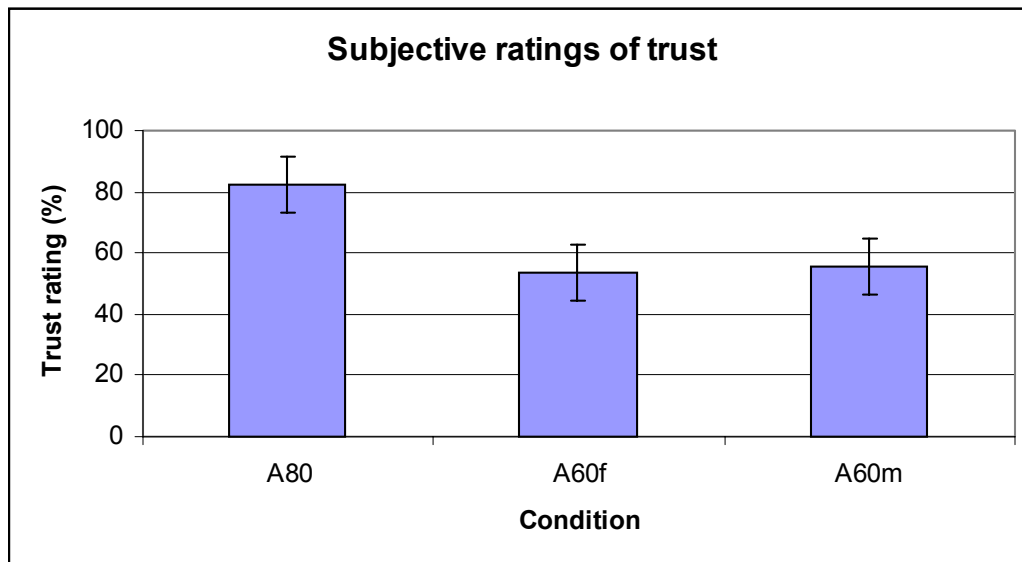


Figure 7. Pilot ratings of trust in automation aid. SE bars are included.

An ANOVA revealed a main effect of condition [$F(2, 19) = 4.13, p = .03$], indicating that the A80 condition facilitated better ratings of trust from the pilots. While Dixon & Wickens (2003) concluded that pilot trust in the automation was poorly calibrated, and more so in the “miss” condition than in the “false alarm” condition, the current data suggest that pilots more accurately calibrate their trust in the aid when they are notified prior to the mission that the automation would be either “fairly reliable” or “not very reliable”, as well as being alerted to which way the bias is set (i.e. conservative or liberal).

In the A80 condition, the rate of false alarms and misses were both 10%, and pilots rated these as 6.5% and 15%, respectively. In the less reliable conditions, pilots were not as accurate in their assessment of false alarm and miss rates, giving scores of 53% and 39%, respectively, in the A60f condition, when the actual rates were 30% and 10%. In the A60m condition, pilots gave ratings of 23% and 48% for false alarms and misses, respectively, when the actual rates were 10% and 30%. While the numbers were off by as much as 20%, it did appear that pilots were aware of which direction the bias was set (i.e. conservative or liberal). This awareness was probably largely due to giving them prior knowledge of this fact.

3.5 Modeling of Automation Dependence: Reliance and Compliance

The current simulation results, coupled with those of Dixon & Wickens (2003) provide us with an ideal opportunity to evaluate a computational model of automation dependence (reliance and compliance). We refer to this study below as “UAV3”. Table 3 shows the joint miss and false alarm rate of six different conditions that were run across the two experiments. Within each condition it is possible to assess measures of :

- **Reliance**, as indexed by (1) the performance on concurrent tasks (here we look at both TOO accuracy and detection rate, as well as frequency of use of the memory refresh key higher reliance → better performance, and less use of the memory refresh), and (2) the time required to respond to an unannounced failure (e.g., RT to an automation “miss”: higher reliance → longer RT).
- **Compliance**, as indexed by (1) the response time and accuracy to an announced system failure (a true alert: higher compliance → shorter RT), and (2) the time required to complete a TOO response *when an SF alert occurs after the TOO inspection has begun*. (Higher compliance → longer TOO response).

Table 3. Dependent measure values as a function of the P(m) and P(FA).

		P(m)				
		0	1	3	5	
P(FA)	0				15.08	Red = RT to SFmiss
		5.32			6.59	Blue = RT to TOO (not SFc)
		3.21			13.12	Pink = RT to SFs (high wl)
		0.07			0.09	Purple = SF miss rate (hi wl)
		0.24			0.30	Black = TOO miss rate (not SFc)
		2.25			5.25	Green = Repeats
	1		27.12	23.29		
			5.95	5.90		
			11.27	13.62		
			0.21	0.25		
			0.09	0.14		
			6.50	8.50		
	3		26.05			
			6.69			
			19.99			
			0.50			
			0.14			
			5.57			
	5		5.38			
			11.00			
		0.31				
		0.15				
		3.04				

Table 3 presents the values of these measures within each of the cells (note however that RT to SF misses cannot be assessed in the condition with no misses). To the extent that reliance and compliance are independent components of automation dependency, and that operators are perfectly calibrated to true reliabilities, we predict that (a) the above two vectors of reliance and compliance performance measures should be linearly effected by miss rate and false alarm rate; (b) each vector should be unaffected by the other term. That is, compliance measures should be unaffected by miss rate, and reliance measures should be unaffected by false alarm rate.

Figure 8 presents the four dependent measures of reliance as a function of miss rate, while figure 9 does so for the two measures of compliance as a function of false alarm rate. To the extent that the pure model is validated, all functions should be relatively linear (note that we have chosen to plot data as a function of the number of “alarm errors”—misses or false alarms—rather than as a percent, or reliability number, because of the difficulty in precisely identifying the probability of a false alarm (Wickens & Hollands, 2000). Shown also in the figure are the linear correlations between reliability (# of automation errors) and the dependent variables.

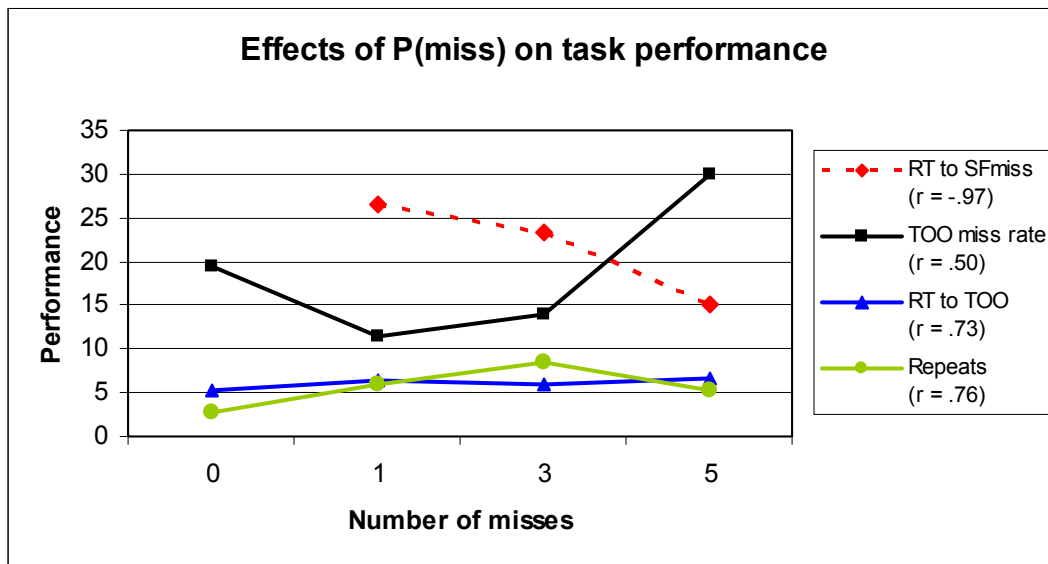


Figure 8. Task performance as a function of the number of automation misses. Note that the some of the data points actually represent the pooled data from two different conditions which, nevertheless had the same X axis value, as shown in table 3. The correlations are based on the y axis values of all the conditions separately, not the averaged values plotted in the figure.

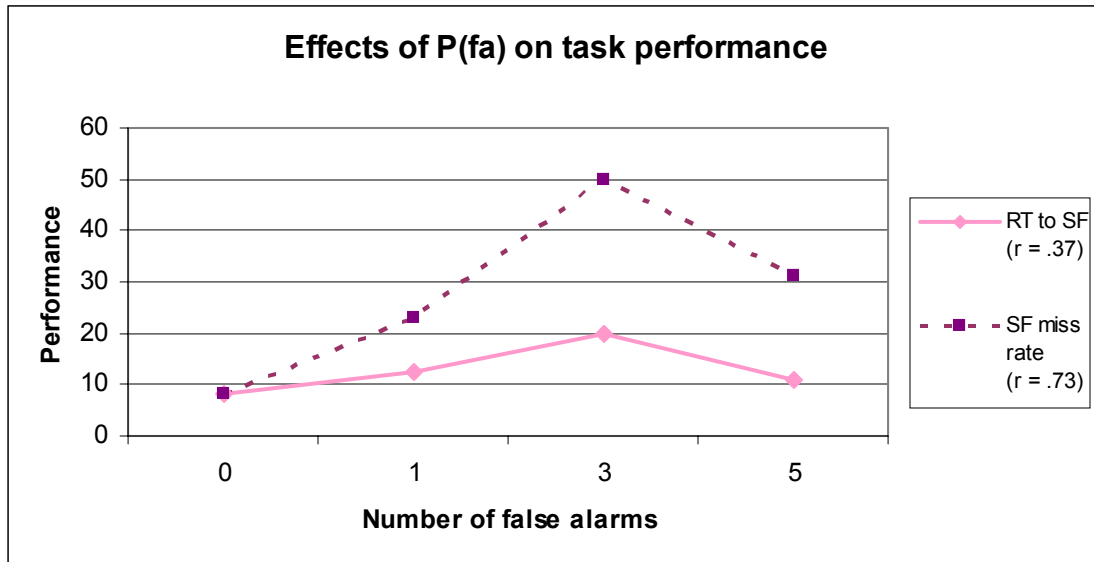


Figure 9. Task performance as a function of the number of automation false alarms.

Examination of figure 8 reveals that all three measures of reliance show a correlation in the expected direction (lower miss rate \rightarrow greater reliance \rightarrow better concurrent task performance and slower response to the rare automation miss); and these correlations are relatively high. Indeed only one data point seriously departs from monotonicity, which is the data point with perfect automation, taken from the previous study UAV3, showing a higher than predicted TOO (concurrent task) miss rate, from one who was relying heavily upon the highly reliable SF alerts. Indeed the other two TOO miss rate data points that are contributed by the UAV3 data are also higher than the linear regression model would predict, suggesting the possibility that, in UAV3, pilots were giving less attention to the 3D image window (where the TOOs appeared) than they should have been (and/or less than they did in UAV4).

The two measures of compliance, shown in figure 9, were assessed at high workload when the participants' attention was heavily engaged in manipulating the 3D image (and therefore might be more reluctant to leave the image inspection task, and switch to the alerted system display). Here again, the correlations are in the expected direction: higher FA rate \rightarrow less compliance \rightarrow slower and less accurate response to the SF alerts. Here also, as with the one TOO reliance measure in figure 8, a close model fit is thwarted by a UAV3 data point where, for FA=5, performance is better (more compliance) than one might otherwise predict, from the skeptical participant, mistrustful of a false-alarm prone system. Why this is the case remains somewhat of a mystery except for the possibility that participants in the previous experiment (UAV3) had an "attentional set" that placed relatively greater priority on SFs, rather than TOOs, compared to the participants in UAV4. We note that this trend is at least consistent with the relatively poorer performance on TOO detections in UAV3 than UAV4, as reflected by the reliance measure of TOO detection rate, discussed above.

4.0 Discussion

The data from the current experiment addressed two general issues: (A) will imperfect automation with a reliability of 80% provide benefits relative to a baseline condition, given that it is above the “cutoff” of 70% where Lee & See (in press) have estimated that benefits are lost? (B) Which type of automation imperfection in an alerting system is more problematic: a miss-prone or a false alarm-prone system, as this may be created by adjusting the decision criterion? Both questions were evaluated with respect to performance on the task that was itself automated—detecting system failures—and on the attention allocated to that task, as reflected by performance on concurrent tasks, and this allocation was examined under both low and high workload. Furthermore, both issues were evaluated in the context of other studies that have examined these issues, in particular, the most similar study, that carried out by Dixon & Wickens 2003. The two issues were expressed as four specific hypotheses, identified by number below.

Regarding the first question, as predicted, by H1, the 80% reliability condition produced no costs to any aspects of performance, and even yielded a benefit to TOO monitoring detection rate. The benefit to the concurrent task, but absence of benefit to the automated task itself, is consistent with H3 and with the pattern in previous findings (Rovira et al, 2002; Dixon and Wickens, 2003), whereby imperfect automation provides greater benefits (or lesser costs) to concurrent tasks than to the task that is itself imperfectly automated, consistent with H2. The current results also indicated that when the imperfection level dropped below the 70% level, to the 60% value characteristic of both the miss and false alarm prone alerting systems, performance was degraded on the automated task and, in some cases, the concurrent tasks. Thus the data are consistent with Lee & See’s (in press) conclusion of an approximate threshold of 70-80% (with many qualifications, such as whether the response is to a first failure, or a steady state response of a well calibrated observer; Wickens, 2000). The issue of *how* performance was degraded differentially by the two kinds of automation failure is the focus of the second issue and H4, to which we now turn.

The contrast between misses and false alarms was less consistent in the current data than was the issue of the performance at the 80% level, and this consistency was lacking both within the different dependent variables of the present study, and with prior research (e.g., Dixon and Wickens, 2003, Maltz & Shinar, 2003). Regarding concurrent task performance, the results were most externally consistent with H4, although internally inconsistent. Replicating Dixon & Wickens 2003, we again found that misses imposed a greater costs than false alarms on memory for the command targets, as reflected in use of the repeat key, whereas false alarms imposed a greater cost than misses on the time taken to detect the targets of opportunity. In accounting for this difference, we can distinguish between the effects of imperfection on an **enduring attentional state**, from those effects on a **particular event**. Thus the miss-prone system forces the pilot to continuously monitor the SF gauges—an enduring state—more so even than in the baseline condition. Such allocation of resources, which prevented any cost from accruing in detecting system failures themselves (figure 5), imposed a cost on memory for the Command Target information, requiring more frequent use of the repeat key. In contrast, the false-alarm prone system has little effect on the enduring attentional state, as the pilot is aware that if there *is* an alarm, s/he will be alerted. However when an alarm does occur, the requirement to assess the raw data, and discriminate whether those data indicate a failure or not, imposes an added delay in

detecting a concurrent TOO, a requirement not imposed with the baseline nor the miss-prone system.

The effects of miss versus false alarm prone automation on the SF task itself are a little more puzzling, and somewhat less consistent. First, the false alarm prone system had no detrimental effects on SF detection rate in the current study, but was shown to disrupt detection rate in Dixon & Wickens 2003 (although table 1 reveals the effect here was in the same direction – 50% detection rate with A60f and 75% with A60m; but a non-significant difference because of the very high variance). Perhaps this difference may have been a consequence of the better calibration between perceived and true reliability achieved in the present study, a calibration which is deemed to be beneficial. A second inconsistency was observed in SF RT (figure 5), where the false alarm prone system clearly led to longer responses than baseline (under high workload), whereas in the prior study, there was no penalty for false alarm proneness on SF detection times. This inconsistency can be readily explained by the single long RT observed in the single automation miss trial that occurred (table 2). When this is removed there is no difference in RT between the two conditions. The importance of this single data point emerges in the analysis of reliance in the modeling effort described below.

Given the lack of full consistency, both internal and externally with the prior UAV study, it is difficult to draw firm conclusions regarding which sort of criterion setting is more problematic in the multi-task UAV context. This ambiguity in turn makes it difficult to resolve the extent to which the current results are consistent with the only other study that clearly, and without confound, manipulated the two types of alerts (Maltz & Shinar, 2003), a study that suggested increasing false alarm rate was more detrimental than increasing miss rate. It is apparent that such differences may be highly contingent upon particular strategies adopted by participants, and these may be quite task and instruction-dependent. However we do believe that the dichotomy of effects on enduring attentional state, versus those on specific event response, and between reliance and compliance, are useful to help distinguish the effects on SF performance and on concurrent performance. These dichotomies are highlighted by the modeling efforts.

As noted above, a useful way to consider the difference between the two conditions is in terms of the dichotomy between reliance, as induced by a miss-free alerting system, and compliance, as induced by a false-alarm free system (Meyer, 2001; Maltz & Shinar, 2003). This difference highlights the challenge imposed by the “apples and oranges” comparison between these two types of stage 1 automation imperfections. By combining the results of the two experiments (a reasonable step, since all reliability conditions were between-subjects, and the procedures followed in the two experiments were essentially identical), we were able to accomplish three goals. First, we believe this is the first effort to validate a quantitative (computational) model of automation, in order to examine the effects of different degrees of reliability on compliance and reliance. Second, replicating the work of Maltz & Shinar (2003), and Meyer (2001), we were able to establish the degree of independence of the two phenomena. Third, we were able to extend the comparative examination of reliance and compliance to the effects on attention and dual task performance, in a way that the prior studies have not done. While such dual task work was carried out by Rovera et al (2001), their work did not invoke the reliance/compliance distinction.

By collectively aggregating the four indices of **reliance** (concurrent performance speed and accuracy, CT memory, and delay in responding to automation failures; Figure 8), we were able to establish that these were reasonably linearly correlated with miss rate and, importantly, were relatively unaffected by (independent of) whether false alarms were high or low. We do note that the dual task measures that best indicated reliance were those collected in high workload.

By collectively aggregating the two indices of **compliance** (response to SF alerts: Figure 9), we also obtained reasonable linearity with false alert rate and, most importantly, observed a fair degree of independence of compliance from reliance, as the latter was driven by the SF miss rate. As a result, we were thereby again able to establish the relative independence of the two cognitive concepts, as Meyer (2001) and Maltz & Shinar (2003) have done. Here we established that independence in a dual task environment, where reliance becomes a critical variable.

5.0 Conclusion

A model of attention and performance with imperfect stage 1 automation assumes that the designer of alerting systems may not be able to perfect the overall reliability of the system, thereby creating an overall automation error rate. However the designer can control the “threshold” of the alert, (like adjusting the parameter “beta” in signal detection), thus affecting the relative balance between false alerts and misses. Each of these error rates feeds, relatively independently, into a set of compliance and reliance effects; influencing human response both at the time that alerts appear (influenced by compliance) and during the remainder of the time, when failures (both detected and undetected by automation) are monitored. Reliance heavily affects attentional state during this remaining time. Compliance influences the propensity to switch attention, away from an ongoing task, when an alert occurs, reflecting the classic “mistrust” affect engendered by false alerts (Bliss, 2003, Breznitz, 1983; Sorkin, 1989). At least across the range of reliability explored, reliance and compliance behavior appears to be relatively linearly calibrated with true failure rate, and at reliability values above 75%, such automation appears to provide benefits.

Acknowledgments

The authors also wish to acknowledge the support of Ron Carbonari and Jonathan Sivier in developing the UAV simulation, and Dervon Chang and Bobby Bernard for helping with data collection.

References

- Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, N.J.: Lawrence Erlbaum.
- Cotté, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger driver’s reliance on collision warning systems. *Proceedings of the 45th Annual Meeting of the Human Factor Society* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.

- Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the 34th Annual Meeting of the Human Factors Society* (pp. 1524-1528). Santa Monica, CA: Human Factors Society.
- Dixon, S. & Wickens, C.D. (2003). *Imperfect Automation in Unmanned Aerial Vehicle Flight Control*. (AHFD-03-17/MAAD-03-1). Savoy, IL: University of Illinois, Aviation Research Lab.
- Dixon, S.R., Wickens, C.D., & Chang, D. (2003). Comparing quantitative model predictions to experimental data in multiple-UAV flight control. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 339-343). Santa Monica, CA: Human Factors and Ergonomics Society.
- Fisher, D.L., & Tan, K.C. (1989). Visual displays: The highlighting paradox. *Human Factors*, 31(1), 17-30.
- Galster, S. M., Bolia, R. S., Roe, M. M., & Parasuraman, R. (2001). Effects of automated cueing on decision implementation in a visual search task. *Proceedings of the 45th Annual Meeting of the Human Factor Society* (pp. 321-325). Santa Monica, CA: Human Factors and Ergonomics Society.
- Hammer, J. M. (1999). *Human factors of functionality and intelligent avionics*. In Garland, Wise, & Hopkin (Eds.), *Handbook of human factors in aviation*. (pp. 549-565). Mahwah, NJ: Lawrence Erlbaum.
- Jones, P. M., Wilkins, D. C., Bargar, R., Sniezek, J., Asaro, P., Danner, N., Eychaner, J., Chernyshenko, S., Schrah, G., Hayes, C., Tu, N., Ergon, H., & Lu, L. (2000). CoRAVEN: *Knowledge-based support for intelligent analysis*. In M. E. Benedict (Ed.), *ARL Federated Laboratory 4th Annual Symposium: Advanced Displays & Interactive Displays Consortium* (pp. 89-94). University Park, MD: US Army Research Laboratory.
- Kantowitz, B., Hanowski, R., & Kantowitz, S. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors*, 39, 164-176.
- Lee, J. D., & See, K. A. (in press). Trust in automation: Designing for appropriate reliance. *Human Factors*.
- Lehto, M. R., Papastavrou, J. D., Ranney, T. A., & Simmons L. A. (2000). An experimental comparison of conservative versus optimal collision avoidance warning system thresholds. *Safety Science*, 36-3, 185-209.
- Lorenz, B., Nocera, F. D., Röttger, S., & Parasuraman, R. (2001). The effects of level of automation on the out-of-the-loop unfamiliarity in a complex dynamic fault-management task during simulated spaceflight operations. *Proceedings of the 45th Annual Meeting of the Human Factor Society* (pp. 44-48). Santa Monica, CA: Human Factors and Ergonomics Society.
- Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors*, 45, 281-295.
- Merlo, J. L., Wickens, C. D., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. *Proceedings of the 4th Annual Army Federated Laboratory Symposium* (pp. 27-31). College Park, MD: Army Research Federated Laboratory Consortium.
- Metzger, U., & Parasuraman, R. (2001). Conflict detection aids for air traffic controllers in free flight: Effects of reliable and failure modes on performance and eye movements.

- Proceedings of the 11th International Symposium on Aviation Psychology* (pp. 1-5). Columbus, OH: The Ohio State University.
- Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, 43, 563-572.
- Meyer, J., & Ballas, E. (1997). A two-detector signal detection analysis of learning to use alarms. *Proceedings of the 41st Annual Meeting of the Human Factor Society* (pp. 186-189). Santa Monica, CA: Human Factors and Ergonomics Society.
- Molloy, R. & Parasuraman, R. (1996). Monitoring an automated system for a single failure: vigilance and task complexity effects. *Human Factors*, 38, 211-322.
- Mosier, K.L., Skitka, L.J., & Korte, K.J. (1994). Cognitive and social psychological issues in flight crew/automation interaction. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends* (pp. 191-197). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, 3(1), 1-23.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics*, 30(3), 286-297.
- Rovira, E., & Parasuraman, R. (2002). Sensor to shooter: Task development and empirical evaluation of the effects of automation unreliability. Paper presented at the *Annual Midyear Symposium of the American Psychological Association*, Division 10 (Military Psychology) and 21 (Engineering Psychology). Ft. Belvoir, VA.
- Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. In *Proceedings of the 26th Annual Meeting of the Human Factors and Ergonomics Society*. (pp. 327-331). Santa Monica, CA: Human Factors and Ergonomics Society.
- Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, 43(4), 573-583
- Skitka, L., Mossier, K., Burdick, M. (1999). Does Automation Bias Decision-making? *International Journal of Human-Computer Studies*, 51(5).
- Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, 10(1), 85-97.
- Wickens, C. D. (2000). *Imperfect and unreliable automation and its implications for attention allocation, information access, and situation awareness* (Technical Report ARL-00-10/NASA-00-2). University of Illinois at Urbana-Champaign, Savoy, IL: Aviation Research Lab.
- Wicken, C.D. & Kessel, C (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 24-34.
- Wicken, C.D. & Kessel, C (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology*, 6(3), 564-577.

- Wickens, C. D., Kroft, P., & Yeh, M. (2000). Data base overlay in electronic map design: Testing a computational model. *Proceedings of the IEA 2000/HFES 2000 Congress*. Santa Monica, CA: Human Factors & Ergonomics Society
- Wickens, C.D. & Xu, X. (2002). *Automation trust, reliability and attention*. (AHFD-02-14/MAAD-02-2). Savoy, IL: University of Illinois, Aviation Research Lab.
- Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors*, 44(1), 44-50.
- Yaacov, A.B., Maltz, M., & Shinar, D. (2003). Effects of an in-vehicle collision avoidance warning system on short- and long-term driving performance. *Human Factors*, 44(2), 335-342.
- Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: The effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3), 355-365.
- Yeh, M., Wickens, C. D., & Seagull, F. J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors*, 41(4), 524-542.