Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload

Stephen R. Dixon and **Christopher D. Wickens,** University of Illinois, Aviation Human Factors Division, Savoy, Illinois

Objective: Two experiments were conducted in which participants navigated a simulated unmanned aerial vehicle (UAV) through a series of mission legs while searching for targets and monitoring system parameters. The goal of the study was to highlight the qualitatively different effects of automation false alarms and misses as they relate to operator compliance and reliance, respectively. **Background:** Background data suggest that automation false alarms cause reduced compliance, whereas misses cause reduced reliance. **Method:** In two studies, 32 and 24 participants, including some licensed pilots, performed in-lab UAV simulations that presented the visual world and collected dependent measures. **Results:** Results indicated that with the low-reliability aids, false alarms correlated with poorer performance in the system failure task, whereas misses correlated with poorer performance in the concurrent tasks. **Conclusion:** Compliance and reliance do appear to be affected by false alarms and misses, respectively, and are relatively independent of each other. **Application:** Practical implications are that automated aids must be fairly reliable to provide global benefits and that false alarms and misses have qualitatively different effects on performance.

INTRODUCTION

Unmanned aerial vehicles (UAVs) are now commonly used to fulfill military reconnaissance missions without endangering human pilots. The current study considered the role of imperfect automation in buffering multitask interference, as a single UAV pilot may be called upon to perform the multiple tasks required of UAV supervision and control.

Imperfect Automation

Previously, Dixon, Wickens, and Chang (2005) employed a perfectly reliable auditory autoalert system to aid pilots in detecting system failures during simulated military reconnaissance missions, and they found that these autoalerts improved performance in the automated task with no performance loss in either of two concurrent tasks. Unfortunately, these types of alerting aids are rarely entirely reliable; subsequently, questions arise as to the effect of unreliable automation on pilot trust, dependence, and human-automation performance. Imperfect automation has been shown to create different states of overtrust, undertrust, or calibrated trust (Parasuraman & Riley, 1997), "complacency" (Metzger & Parasuraman, 2005; Parasuraman, Molloy, & Singh, 1993), and performance loss (Molloy & Parasuraman, 1996).

In spite of such reported problems, imperfect automation clearly can assist human operator performance (e.g., Galster, Bolia, Roe, & Parasuraman, 2001; St. John & Manes, 2002; Yeh, Merlo, Wickens, & Brandenburg, 2003), particularly in circumstances when human resources to the unaided task are insufficient (e.g., Maltz & Shinar, 2003; Yaacov, Maltz, & Shinar, 2003) and, therefore, the human must depend upon the automation. Such resource scarcity may result either when the task itself is difficult (Maltz & Shinar, 2003) or when the automated task is carried out in a multitask context (C. D. Wickens & Dixon, 2005).

Report Documentation Page

1. REPORT DATE **2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Illinois,Aviation Human Factors Division,1 Airport Road,Savoy,IL,61874**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**13**	

Diagnostic Failures: Misses and False Alarms

The focus of the current study was on imperfect automation diagnostic alerting systems, in which the automation attempted to distinguish two possible states of the world: a "safe" state and a "dangerous" one (Swets & Pickett, 1982). The sources of imperfection in such systems relate to imperfect sensors and algorithms as well as to noisy or probabilistic data in an uncertain world. The performance of such systems can generally be represented in the framework of signal detection theory (Green & Swets, 1988; T. D. Wickens, 2002), whereby the consequences of the imperfection show up as automation misses and/or false alarms.

In application, the automation designer typically has the opportunity to set "beta" (the threshold of the alerting system) in a way that will trade off the relative frequency of these two kinds of automation errors. At issue is where this trade-off should optimally be set. If the output of the automatic diagnostic process directly triggers a decision, then the optimal criterion could easily be calculated by applying some expected value algorithm to the consequences of the two sorts of resulting actions. However, this process becomes complicated when the human operator also has parallel access to the same perceptual "raw data" processed by the automation, bringing qualitatively different strengths of perceptual analysis to bear. Here the optimal setting may vary (Sorkin & Woods, 1985). In such cases, an automation miss may not inevitably create a total system miss if the human is somewhat vigilant of the raw data. Furthermore, in those multitask situations in which automation dependence is critical because of high workload, the costs to total system performance must also account for the costs (of automation misses and/or automation false alarms) to human performance on concurrent tasks.

There is some evidence that the generic costs of alerting system false alarms may be greater than those of misses. For example, Bliss (2003) found that pilots reported more than twice as many alert-related aviation incidents related to false alarms as compared with those related to misses (although this disparity may reflect a higher base rate of false alert events). Maltz and Shinar (2003) observed a similar asymmetry in their laboratory data.

Furthermore, false alarms are well known to cause annoyance, to lead to unnecessary evasive actions, and, in the worst-case scenario, to lead to sufficient distrust of the automated system that true alarms are ignored – the "cry wolf" syndrome (Breznitz, 1983; Parasuraman & Riley, 1997; Sorkin, 1989). Despite such evidence, it is important to note that in many situations, misses may be more costly than false alarms (e.g., air traffic control) and that experts may be more accepting of false alarms than of misses (Masalonis & Parasuraman, 1999).

Reliance Versus Compliance

The qualitative distinction between the two kinds of diagnostic imperfections is important because of the recent dichotomization of two very different cognitive states – reliance and compliance – that are associated with automated diagnostic systems committing one or the other type of error, particularly under conditions of high workload (Maltz & Shinar, 2003; Meyer, 2001, 2004). We consider these two states to be two different manifestations of automation dependence, a dependence that will be inversely related to automation reliability in resource-scarce circumstances. Here *reliance* refers to the human operator state when the alert is silent, signaling "all is well." Reliant operators will have ample resources to allocate to concurrent tasks because they rely on the automation to let them know when a problem occurs on the automated task. Miss-prone automation will degrade reliance, particularly under high workload, and as a result should lead to decrements in concurrent tasks. In forcing the operator to pay closer attention to the raw data of the alerted domain, there should be more effective detection of those (now more frequent) misses made by the automation system. Conversely, highly reliable, low-miss automation, although availing ample resources for concurrent tasks, should leave the operator quite vulnerable to the rare automation misses during high workload – the "complacency" effect (Bainbridge, 1983; Molloy & Parasuraman, 1996; Parasuraman et al., 1993).

In contrast, *compliance* describes the operator's response when the alarm sounds, whether true or false. A compliant operator will rapidly switch attention from concurrent activities to the alarm domain (and possibly immediately initiate an alarm-appropriate response, such as leaving the

building upon hearing a fire alarm). Automation that is prone to false alarms will degrade compliance, the consequences of which are a delayed response (or possibly, no response at all) to a true alarm (Breznitz, 1983; Sorkin, 1989).

Although the research of Meyer (2004) has suggested that the two may be somewhat independent states, with separate factors affecting reliance and compliance, these two constructs have not been separately and quantitatively evaluated within a multitask context in which resources are scarce and the threshold of an alarm system is systematically varied to alter the two types of automation errors. Maltz and Shinar (2003) imposed such variation but did so within a single-task context in which resource demand (and automation dependence) was created by a more demanding task. Unfortunately, no study of imperfect alert automation has systematically varied the threshold of the alert system within a dual-task context, where the consequences of allocating resources to the secondary task can be assessed.

The Current Study

Because of its inherent multitask nature (Dixon et al., 2005) and ecologically valid properties, UAV simulation provided an ideal test bed for two experiments that examined the issues of imperfect automation in dual-task settings. In both experiments, participants conducted simulated reconnaissance missions in which they were responsible for navigating an UAV to 10 different command targets and for reporting details of those targets to mission command (Dixon et al., 2005). This was considered the primary task. Simultaneously, they were required to search for possible targets of opportunity (TOOs) along the way. Upon detecting targets, a high-workload camera zoom and inspect task was engaged. This was considered the secondary or concurrent task, upon which the hypothesized effects of reliance could be observed. Participants also had to monitor on-board system parameters for possible failures. This was considered the imperfect diagnostic automation task supporting the primary task, given that an auditory automation-alert aid was sometimes available to indicate when these system failures had occurred.

In Experiment 1, this aid was either perfectly reliable or 67% reliable (producing either false alarms or misses in two different conditions). A

fourth condition, with no automation, provided baseline data with which to compare these automated conditions. In Experiment 2, participants were assisted by the same autoalert aid but with reliability levels of 80% (producing both a false alarm and a miss) and two conditions of 60% (producing both false alarms and misses at a 3:1 ratio, and vice versa). The multiple levels of automation reliability achieved by varying miss and false alarm rate independently across the two experiments provide data to validate a computational model of dependence on imperfect automation. More specifically, our experiments address four hypotheses:

H1: The symptoms of automation dependence (benefits if correct, costs if incorrect) will emerge primarily at high workload. Automation imperfection driven by misses and false alarms would show qualitatively different effects as reflected by measures of reliance and of compliance respectively.

H2: Indices of high reliance will decrease with increasing miss rate. High reliance is indicated by good target-of-opportunity performance and command target memory and also by a slow response to the rare system failure miss when automation is reliable.

H3: Indices of high compliance will decrease with increasing false alarm rate. High compliance is indicated by rapid and accurate responses to all alerts, whether true or false.

H4: The two vectors of reliance and compliance will show relative independence from each other.

METHODS: EXPERIMENT 1

Participants

Thirty-two undergraduate and graduate students received $8/hr, plus bonuses of $20, $10, and $5, for first-, second-, and third-place finishes, respectively, out of groups of 8 participants. Participants were made aware of the incentives and told how the overall task performance would be calculated. Twenty of the participants were licensed pilots, who were equally distributed across conditions.

Apparatus

The experimental simulation ran on an Evans and Sutherland SimFusion 4000q system. The UAV

display was generated on an OPENsim Graphics card on a Hitachi CM721F 19-inch (48-cm) monitor, using 1280 × 1024 resolution. Figure 1 presents a sample display for a single UAV.

As shown in Figure 1, the experimental environment was subdivided into four separate windows. The top left window contained a 3-D egocentric image view of the terrain directly below the UAV (6000 feet altitude). During regular tracking periods, the operator could only view straight down to the ground. During a loiter pattern, the operator was able to zoom and to extend the viewing angle from 0° to 90° along both the x and y axes. The bottom left window contained a 2-D top-down map of the 20 × 20 mile (32 × 32 km) simulation world. Coordinates from 0° to 90° were placed along the x and y axes for navigation purposes. The bottom center window contained the message box, with "fly to" coordinates and command target (CT) report questions. These flight instructions were present for 15 s and could be refreshed for another 15 s at any time during the mission by pressing a "repeat" key. The bottom right window contained the four system gauges for the system failure monitoring task. 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 system failure occurred when one of the white bars moved into a red zone, indicated in gray at the tops and bottoms of the gauges in Figure 1. Participants used a Logitech Digital 3-D joystick to manipulate the aircraft/camera and an X-Key 20-button keypad to indicate responses.

Procedure

Each participant flew one UAV through 10 consecutive mission legs. During each leg, the participant completed three goal-oriented tasks that are commonly associated with UAV flight control: mission navigation and command target inspection, target of opportunity (TOO) search, and systems monitoring. At the beginning of each mission leg, participants obtained their flight instructions for that leg via the message box. Once participants arrived at the CT location, they loitered around the target, manipulated a camera for closer target inspection via a joystick, and reported back relevant information to mission command (e.g., "What weapons are located on the south side of the building?").

Along each mission leg, participants were also responsible for detecting and reporting the low-salience TOOs, a task similar to the CT report except that the TOOs were much smaller (1°–2° of visual angle) than the CT report objects and were
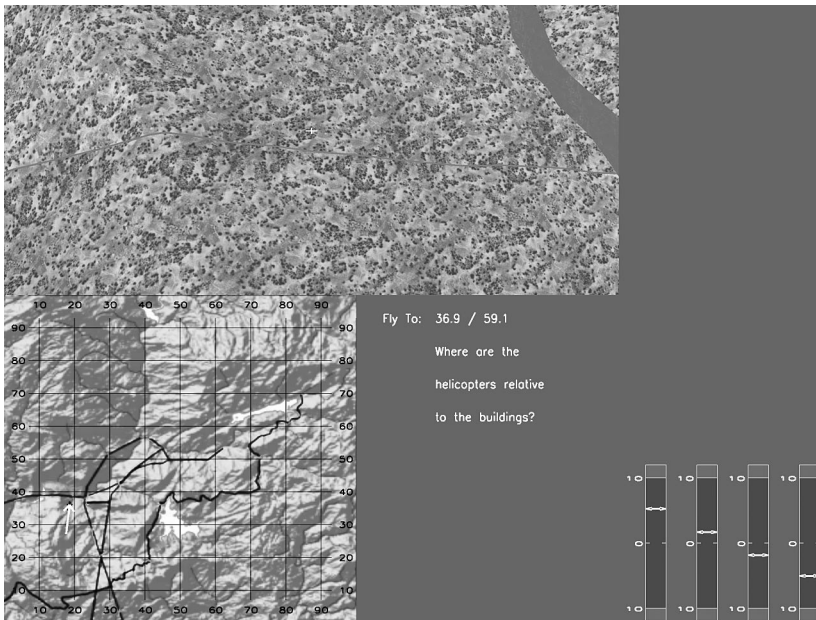

Figure 1. UAV simulation display. The actual display was larger, had better resolution, and was color-coded.

camouflaged. This was considered the secondary or concurrent task. TOOs were located randomly somewhere in the middle 60% of each leg; if a TOO was found, a report response with zooming and panning was required, much like the CT report. TOOs could become visible during simple tracking (low workload) or during a participant response to a system failure (high workload). These two types of TOOs occurred, respectively, with a ratio of roughly 4:1. Upon making a TOO report, the UAV was reoriented by the pilot to continue its original trajectory toward the command target.

Concurrently, participants were also required to monitor the system gauges for possible system failures (SFs), which were designed to fail during either simple tracking (i.e., low workload: easy concurrent task) or TOO and CT zoom/loiter inspection (i.e., high workload: difficult concurrent task). SFs lasted only 30 s, after which the screen flashed bright red and a harsh auditory alarm announced that the participant had failed to detect the SF. There were a total of 10 SFs, with no more than 2 occurring during any mission leg. SFs were temporally separated by 4 to 10 min. Some SFs were alerted with an automated auditory warning system (i.e., a tone).

Design

The auditory autoalerts for the SFs were provided for three out of the four conditions, using a between-subjects design (8 participants/group). The A100 condition (A = automation, 100% reliable) provided 10 true alarms with 10 SF events. The A67f condition (f = false alarm, 67% reliable) provided 10 true alerts and an additional 5 false alarms. The A67m condition (m = miss, 67% reliable) provided 10 true alerts but failed to alert an additional 5 events (10 true alarms plus 5 misses). During a false alarm, the participant was instructed to ignore the warning after cross-checking with the raw data to confirm the inaccuracy of the alarm. If an automation miss occurred, the participant was instructed that he or she was still responsible for "catching" the SF and correcting it. The final condition was a baseline condition, with no automation aid to assist participant performance.

RESULTS: EXPERIMENT 1

Three planned comparisons were used throughout to assess statistical effects. For each dependent measure, the following were compared: (a) baseline versus the combination of A67f and A67m in a planned comparison (i.e., weights of –1, 0.5, 0.5); (b) baseline versus A100; and (c) A67f versus A67m. Because only three a priori comparisons were implemented to view important differences between particular groups of interest, familywise error rates were not adjusted (see Keppel, 1982, for more details). One participant in the baseline condition was dropped because the data file was corrupted. Note that because of frequent missing data points (e.g., if a target does not come into view on the 3-D 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 the data.

Primary Task: Mission Navigation and CT Inspection

Tracking error and CT reporting. Planned comparisons revealed no main effect for tracking error (all ps > .10) or for CT reporting speed and accuracy. Participants clearly treated mission navigation and CT inspection as the primary task.

Repeats. Planned comparisons revealed that the 67% reliable conditions (mean of A67f and A67m) did not statistically differ from baseline, $t(20) = 1.49, p > .10$. There was also no significant difference between the A100 condition, $t(13) < 1.0$, and baseline. However, the A67m condition generated twice as many repeats as the did A67f condition, $t(14) = 2.52, p = .01$. Thus, miss-prone automation imposed more of a load on memory, which was compensated by the repeat key, relative to false-alarm-prone automation.

Secondary Task: TOO Monitoring

TOO detection rates. Planned comparisons revealed no significant difference between the baseline condition and the 67% reliable conditions, $t(18) < 1.0$, or the A100 condition, $t(12) < 1.0$; however, detection rates were significantly lower in the A67m (miss) condition than in the A67f (false alarm) condition in both the low-workload, $t(12) = 2.25, p < .05$, and high-workload trials, $t(12) = 2.20, p < .05$.

TOO detection times. Because low-workload trials revealed no effects of condition on TOO detection times (all ps > .10), we focused primarily

TABLE 1: An Overview of the Data from Experiment 1

	Baseline	A100	A67f	A67m
Tracking error (MAE in meters)	84.25 (0.81)	83.80 (0.69)	79.32 (4.61)	83.08 (1.03)
Number of repeats (per leg)	3.03 (0.82)	2.25 (0.48)	3.04*** (0.67)	6.5*** (1.20)
CT detection time (s)	2.45 (0.80)	2.41 (0.51)	2.31 (0.31)	3.37 (1.07)
TOO detection rate (%)	58.57 (6.7)	56.57 (1.3)	65.56** (6.1)	41.25** (9.0)
TOO detection time (s)				
High workload	6.03 (1.99)	7.83 (0.96)	13.82* (3.08)	7.7* (2.07)
Low workload	6.04 (0.91)	5.32 (1.0)	5.38 (0.96)	6.59 (3.1)
SF detection rate (%)				
Low load	100.0 (0.0)	100.0 (0.0)	94.46 (4.2)	97.92 (1.4)
High load	95.83 (4.2)	88.0 (7.1)	68.75** (7.8)	92.19** (5.2)
SF detection time (s)				
Low load	2.17 (0.35)	3.00 (0.71)	2.69 (1.19)	3.15 (0.61)
High load	10.75** (3.51)	3.21** (0.52)	11.0 (2.34)	13.75 (2.06)
SF report accuracy (%)	88.36 (3.0)	91.22 (2.2)	96.58 (1.3)	96.67 (1.1)

Note. SE values are in parentheses. MAE = mean absolute error, CT = command target, TOO = target of opportunity, SF = system failure.
$^{*}p < .10.$ $^{**}p < .05.$ $^{***}p < .01.$

on high-workload trials, when participants were concurrently dealing with an SF, and resources were assumed to be scarce. Planned comparisons revealed no statistical difference between the mean of the 67% reliable conditions relative to baseline, $t(11) < 1.0$, or the A100 condition relative to baseline, $t(10) < 1.0$. However, the A67f condition may have generated longer detection times than the A67m condition did, $t(6) = 1.40, p = .10$ (approaching significance).

SF Monitoring

SF detection rates. The main focus of interest in the SF task was during high-workload trials (i.e., concurrent with TOO inspection), when resources were assumed to be scarce, as low-workload trials showed no effects (H1). Planned comparisons revealed that the 67% reliable conditions resulted in poorer detection rates than did the baseline condition, $t(19) = 1.97, p = .06$

(approaching significance); however, these effects were probably attributable to the A67f condition (69%), in which performance was much worse than in the A67m condition (92%), $t(14) = 2.32$, $p < .05$. The A100 condition did not differ statistically from baseline, $t(10) < 1.0$.

SF detection times. As with SF detection rates, the only effects were observed in high-workload trials. Figure 2 presents the overall SF detection times as a function of workload and reveals that performance in the 67% reliable conditions was no better than in the baseline condition, $t(20) < 1.0$, whereas performance in the A100 condition was better than in baseline, $t(11) = 1.96, p < .05$. The A67f and A67m conditions did not differ statistically overall, $t(14) < 1.0$. However, it is interesting to note that the A67m condition resulted in detection times slower than those in the A67f condition on those occasions when the automation failed to notify the participants of an SF, $t(14) = 2.64, p <$
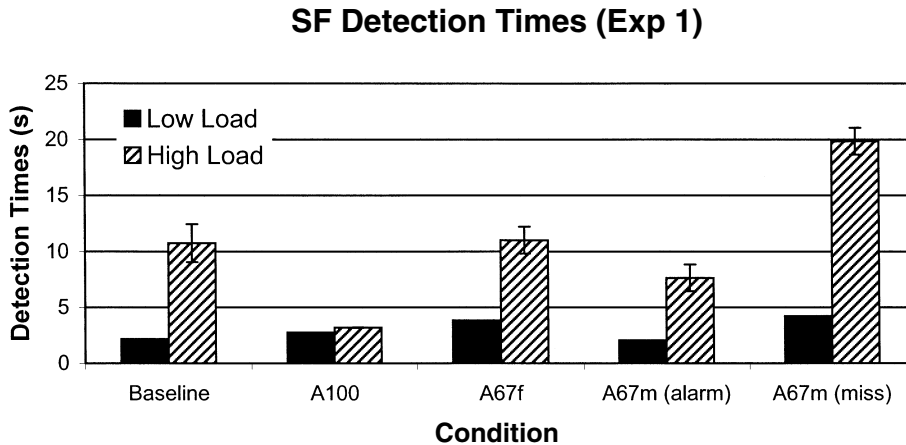
SF Detection Times (Exp 1)

Figure 2. SF detection times across condition and workload. Experiment 1. The A67m condition is divided into two subgroups: a) Automation true alerts (67% of the time), and b) Automation misses (33% of the time). *SE* bars are included.

.01, reflecting a form of complacency, as shown by the bars at the right in Figure 2. Furthermore, there was a tendency for faster response times (RTs), compared with the A67F condition, on those occasions when the alarm sounded.

DISCUSSION: EXPERIMENT 1

Participants were effective at protecting the primary task indices of tracking and CT report accuracy. As hypothesized (H1), automation reliability effects were also seen most strongly in high-workload situations. Perfect automation had a beneficial effect, relative to baseline, on performance in the automated task, but it had no benefit on concurrent task performance, replicating Metzger and Parasuraman (2005) and previous UAV studies (e.g., Dixon et al., 2005). Importantly, imperfect automation (67%) hurt both the automated task and concurrent tasks, even dropping these below baseline in some cases. False alarms and misses yielded qualitatively different kinds of effects related to compliance (H3) and reliance (H2), respectively. False alarms hurt the system-monitoring task by reducing SF detection rates and increasing SF detection times as compared with baseline. This indicates that the operators were less compliant with the autoalerts (reduced compliance). Misses hurt performance in remembering flight instructions and possibly in the target search task, indicating a reduction in reliance. We discuss these effects in more detail fol-

lowing the presentation of converging evidence provided by Experiment 2.

METHODS: EXPERIMENT 2

The procedures of Experiment 2 replicated those of Experiment 1 with the following exceptions: No baseline condition was run. An A80 condition (A = automation, 80% reliable) failed by giving 1 false alarm and 1 miss during each mission (8 true alarms, 1 miss, and 1 false alarm). These 2 automation failures, occurring out of a possible 10 alerted system failures, defined a .80 reliability level (1 – 2/10). An A60f condition (f = false alarm, 60% reliable) was created by imposing 3 automation false alarms and 1 automation miss (4 automation failures) out of the 10 possible system failures. An A60m condition (m = miss, 60% reliable) resulted in 3 misses and 1 false alarm (6 true alarms plus 3 misses and 1 false alarm). Participants were not aware of the precise level of reliability provided by each automation aid; however, in contrast to Experiment 1, depending on the participants' assigned condition, they were told in advance that the automation was either "fairly reliable" or "not very reliable" as well as the bias setting of the alert (i.e., more false alarms or more misses). There were 24 participants (8/group), none of whom participated in Experiment 1. Participants were of the same demographics as those in Experiment 1, including the same proportion of pilots to nonpilots.

RESULTS: EXPERIMENT 2

Because of the between-subjects design and the close temporal proximity of the two experiments, the baseline data for Experiment 1 were used in the data analysis of Experiment 2 as well. Table 2 presents an overview of the data. As with Experiment 1, statistical inference was based on planned contrasts of baseline versus 60% reliability (mean of A60f and A60m), baseline versus A80, and A60f versus A60m.

Mission Completion

Planned comparisons revealed no main effect for tracking error or for CT report accuracy (all ps > .10), findings consistent with Experiment 1. However, planned comparisons did reveal that for CT detection times (i.e., how long it took participants to detect the CT once it entered the 3-D display), performance in the two 60% reliable conditions was worse than baseline, $t(20) = 2.77$, $p < .05$, whereas the A80 condition did not differ from baseline, $t(13) < 1.0$. There was no statistical difference between the A60f and A60m conditions, $t(14) < 1.0$. Compared with baseline, both the 60% reliable conditions, $t(19) = 2.49$, $p < .05$, and the A80 condition, $t(11) = 1.72$, $p = .06$ (approaching significance), generated more repeats. The A60m condition generated significantly more repeats than did the A60f condition, $t(14) = 1.85$, $p < .05$.

TOO Monitoring

TOO detection rates. Planned comparisons revealed that there was no difference between the 60% reliable conditions and baseline, $t(20) = 1.17$, $p > .10$, whereas performance in the A80 condition was better than baseline, $t(13) = 2.15$, $p < .05$.

TABLE 2: An Overview of the Data from Experiment 2

	Baseline	A80	A60f	A60m
Tracking error (MAE in meters)	84.25 (0.81)	84.45 (1.95)	82.75 (5.11)	85.76 (1.92)
Number of repeats (per leg)	3.03** (0.82)	5.57* (1.72)	5.25** (1.65)	8.5** (1.59)
CT detection time (s)	2.45** (0.80)	1.96 (1.07)	4.16** (1.10)	4.11** (1.84)
TOO detection rate (%)	58.57** (6.7)	93.0** (7.4)	87.0 (7.1)	82.0 (7.2)
TOO detection time (s)				
High workload	6.03 (1.99)	8.58 (2.82)	14.72*** (2.63)	11.86*** (5.51)
Low workload	6.04 (0.91)	5.94 (1.28)	6.68 (1.20)	5.89 (1.24)
SF detection rate (%)				
Low load	100.0 (0.0)	100.0 (2.8)	97.0 (2.7)	98.0 (2.7)
High load	95.83 (4.2)	69.0 (19.7)	50.0 (53.0)	75.0 (26.0)
SF detection time (s)				
Low load	2.17 (0.35)	2.08 (0.71)	2.50 (0.19)	3.15 (0.19)
High load	10.75 (3.51)	11.27 (3.31)	19.98** (3.19)	13.62** (3.20)
SF report accuracy (%)	88.36 (3.0)	97.0 (4.8)	98.0 (4.4)	94.0 (5.0)

Note. SE values are in parentheses. MAE = mean absolute error, CT = command target, TOO = target of opportunity, SF = system failure.

*p < .10. **p < .05. ***p < .01.

There was no significant difference between the A60f and the A60m conditions, $t(14) < 1.0$.

TOO detection times. Figure 3 presents TOO detection times as a function of condition and workload. On low-workload trials, there were no effects of condition (all $ps > .10$).

In high-workload trials, planned comparisons revealed that performance in the 60% reliable conditions was worse than baseline, $t(16) = 3.09$, $p < .01$, but there was no difference between the A80 condition and baseline, $t(12) < 1.0$. A comparison of the A60f and A60m conditions revealed no significant difference, $t(11) = 1.04$, $p > .10$, although the trend toward greater decrement with the A60f condition is consistent with that observed in Experiment 1.

SF Monitoring

SF detection rates. There were no statistical effects of condition on SF detection rates (all $ps > .10$); however, the reduced rates in the A60f condition in high workload (50%), as compared with the other conditions (mean = 74%), are consistent with those observed in Experiment 1.

SF detection times. Figure 4 presents the SF detection times as a function of condition and workload. No differences in performance were revealed in the low-workload trials; however, in the high-workload trials, performance in the 60% reliable conditions may have been worse than baseline, $t(20) = 1.89$, $p = .07$ (approaching significance). This difference was attributable primarily to the A60f condition, in which performance was worse than in the A60m condition, $t(14) = 2.16$, $p < .05$.

The A80 condition did not differ from baseline, $t(13) < 1.0$.

In Figure 4, we note that each of the 60% condition means was composed of two different components: responses when an alert correctly sounded (A60f = 13.93 s; A60m = 3.96 s) and those when the alert failed to sound (A60f = 26.05 s; A60m = 23.29 s). These data within the high-workload condition reveal the clear slowing for RT when the alarm "missed" the SF event, indicating that in both conditions participants had relied heavily upon the automation and their detection suffered when it failed: the classic "complacency" effect (Parasuraman et al., 1993). Although this complacency effect was less pronounced in the miss-prone condition, the difference between the two error conditions did not approach significance. Correct alerts were responded to more rapidly with the miss-prone automation than with the false-alarm-prone automation, $t(14) = 2.00$, $p < .05$, reflecting the participants' immediate *compliance* with the auditory alert (Meyer, 2001, 2004) in the former condition, in contrast to the false-alarm-prone condition, in which participants were less likely to interrupt target inspection to deal with the alarms. We also infer that greater compliance in the miss-prone condition was coupled with an ongoing greater awareness of the SF gauges, fostered by a reduced *reliance* on that automation and causing the greater disruption to CT memory recall described previously. The difference between reliance and compliance effects will be explored later in greater detail.
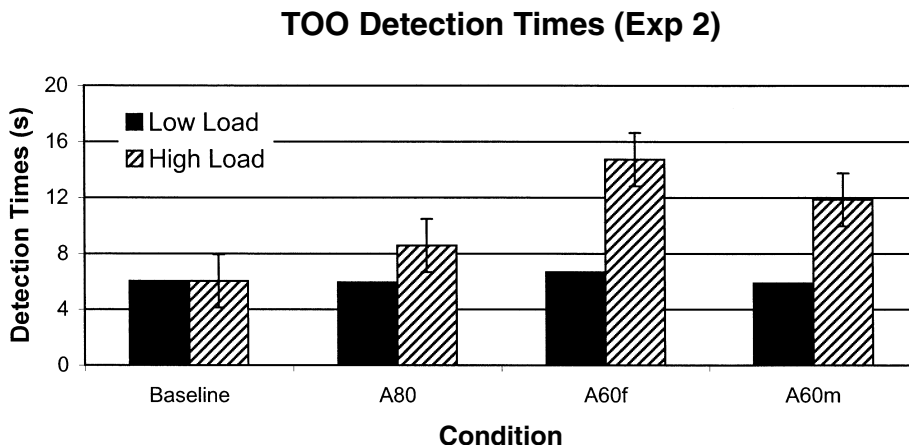
TOO Detection Times (Exp 2)

Figure 3. Experiment 2: TOO detection times across condition and workload. *SE* bars are included.
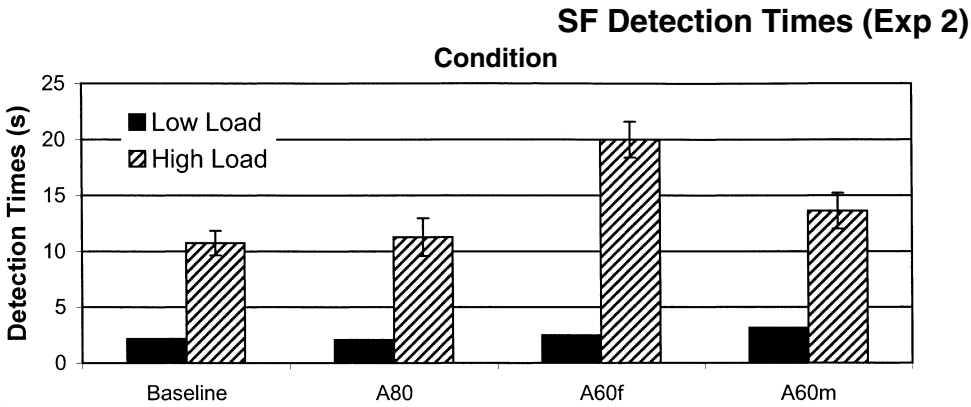
Figure 4. Experiment 2: SF detection times across condition and workload. *SE* bars are included.

DISCUSSION: EXPERIMENT 2

As with Experiment 1, the primary tasks of tracking and CT reporting were fully protected from the effects of degraded reliability, although degraded reliability, particularly that prone to misses, induced more use of the repeat key to compensate for the attention demands of this degradation. The effects on other tasks were primarily seen in high-workload situations. Highly reliable automation did not benefit performance in the automated task relative to baseline, but it had a small benefit to concurrent task performance. Low-reliability automation (60%) hurt both the automated task and concurrent tasks, with different effects for false alarms and misses related to compliance and reliance.

MODELING OF AUTOMATION DEPENDENCE

The current simulation results provide an ideal opportunity to evaluate a computational version of the model of reliance and compliance (Meyer, 2001), the two components of diagnostic automation dependence. Within each condition it is possible to assess measures of reliance and compliance:

Reliance is indexed by (a) the performance on secondary or concurrent tasks. Here TOO accuracy and detection time (during non-SF periods, when reliance was necessary) are examined, as is frequency of use of the memory refresh repeat key (higher reliance → better performance and less

use of the memory repeat). (b) Reliance is also indexed by the time required to respond to an unannounced failure (e.g., RT to an automation "miss": higher reliance → longer RT, reflecting the "complacency effect" with highly reliable automation; Molloy & Parasuraman, 1996). We evaluated this latter measure only under high-workload conditions, in which reliance is most likely to be observed.

Compliance is indexed by the response time and accuracy to an announced system failure (higher compliance → shorter RT), again under high workload.

To the extent that reliance and compliance are components of automation dependency, and that operators are perfectly calibrated to true reliabilities, we predicted that those two vectors of reliance and compliance performance measures should be linearly affected by the independent variables of miss rate (H2) and false alarm rate (H3), respectively. Furthermore, to the extent that it is an independent component, each vector should be unaffected by the other independent variable (H4).

Examination of the data revealed that all four measures of reliance showed a correlation in the expected direction. SF automation miss rate correlates with TOO miss rate, $r = .50$; RT to TOO, $r = .73$; repeats, $r = .76$; and RT to SF misses, $r = -0.97$ – that is, higher miss rate → less reliance → poorer concurrent task performance and faster response to the automation miss (Meyer, 2001; Parasuraman et al., 1993).

The two measures of SF alert compliance were

assessed at high workload, when the participants' attention was heavily engaged in manipulating the 3-D image to inspect targets (and therefore might be more reluctant to leave the image inspection task and switch to the alerted system display). Here again, the correlations were in the expected direction. The correlation of automation FA rate with RT to SF was $r = .37$; with SF miss rate it was $r = .73$ – that is, higher FA rate → less compliance → slower and less accurate response to the SF alerts. Here also, as with one of the TOO reliance measures, a closer model fit was thwarted by an Experiment 1 data point where, for FA = 5 (A67f from Experiment 1), performance was better (more compliance) than one might otherwise predict from the skeptical participant who is mistrustful of a false-alarm-prone system. By way of explanation, we note that in Experiment 1, participants were not prealerted to the high false alarm rate. Hence it would have taken a few trials for the lack of compliance to evolve, thereby diluting the effect.

Hypothesis 4 posits the independence of compliance from miss rate and of reliance from false alarm rate. To assess this, we correlated miss rate to the two indices of compliance. The correlations were $r(3) = .29$, $p = .33$ (SF RT), and $r(3) = -.33$, $p = .17$ (SF miss rate), supporting such a model of independence. The correlations of FA rate to the four indices of reliance were $r(2) = .92$, $p = .08$ (RT to SF miss), $r(3) = -.69$, $p = .18$ (TOO miss rate), $r(3) = -.16$, $p = .14$ (RT to TOO), and $r(3) = -.10$, $p = .27$ (repeats). The former two values, though not significant, suggest that reliance may have been affected by the false alarm rate. High false alarm rates appear to have produced greater reliance upon the automation, although this claim cannot be proven with the current data.

Because all individual correlations were based on a small *N,* we examined Hypotheses 2 through 4 in a different way to increase statistical power. Each variable was standardized and expressed as a proportion of the range between minimum and maximum observed value. These standardized values were inverted where necessary, such that changes in all variables within a vector that were associated with increases in reliance or compliance were of the same sign. The standardized variables within each vector were then pooled. Correlations on the pooled data revealed that miss rate → reliance ($r = .67$, $p < .01$); miss rate →

compliance ($r = .07$, *ns*); FA rate → reliance ($r = -.50$, $p = .06$); FA rate → compliance ($r = .49$, $p = .11$). As we will discuss, this pattern is only partially consistent with the independence hypothesis, because higher false alarm rates appear to have an influence on reliance.

GENERAL DISCUSSION

Prior literature has well established that perfect automation will offer benefits when workload is high, either because the task being automated is challenging (e.g., Maltz & Shinar, 2003) or, as in the current case, because other multitask responsibilities are competing for the operators' limited attentional resources (C. D. Wickens & Dixon, 2005). The current data confirmed this effect, as A100 performance was superior to baseline performance in the RT to system failures only at high workload, supporting H1. Also, there are now ample data showing that people depend on automation even when it is imperfect, and here we found in the A80 condition that benefits were still evident over baseline performance in detecting TOOs, just as such benefits have been observed in other studies (e.g., Maltz & Shinar, 2003; St. John & Manes, 2002; Yaacov et al., 2003).

In the current experiment, we were particularly interested in the manifestations of this dependence when the reliability dropped still further and, in particular, how it was reflected in the two components of dependence, reliance and compliance, articulated by Meyer (2001, 2004). We found first, in support of Hypothesis 1, that dependence costs emerged more markedly under high-workload than under low-workload conditions. This was particularly true for the manifestations of compliance, in which the prolongation of RT to auditory alerts with the false-alert-prone system was observed only while participants were concurrently engaged in TOO and CT image inspection (high workload, Figures 2 and 4), and only in this condition was the decrease in SF detection rate evident (Experiment 1 only).

We also found support for Hypotheses 2 and 3 when examining the independent effects of miss rate on reliance and false alarm rate on compliance, respectively; this has not been previously reported in a multitask experiment. Our approach was through creating the "vector" measures of each construct. Our data revealed a strong effect

of miss rate on reliance ($r = .67$), as participants became less trusting of the automation to alert them if a failure occurred and (a) allocated more attention to monitoring the raw data at the expense of two concurrent tasks (TOO monitoring and CT coordinate memory) but (b) caught the rare automation miss of the system failure more frequently. Correspondingly, we found support, although somewhat weaker (lower correlation, $r = .49$), for the negative effects of high false alert rate on compliance, reflecting the "cry wolf" phenomenon.

Hypothesis 4 concerns independence, which was not explicitly framed as a property of reliance and compliance by Meyer (2004) but has indeed appeared to be an implication of his research. Here, however, the data were mixed. Indeed, miss rate appeared to have little influence on the vector of compliance. The participants' attention was drawn more or less to the alert, independent of the imperfection of that alert when it was silent. Puzzling, however, was the influence of false alert rate on reliance ($r = -.50$), which was just as strong as its effect on compliance ($r = .49$). Upon closer examination of the components of the reliance vector, the direction of this effect (more false alarms → less reliance) was driven heavily by the fact that more false alarms increased the response time to the rare automation-missed system failure. In this regard, it appears that a false-alarm-prone system may leave the operator somewhat less inclined to pay any attention to the entire automated domain, whether it be its alerting signal or the raw data contained within.

As a final observation, we note the general pattern of the current data: Our two lowest levels of reliability clearly inhibited performance below baseline, whereas our higher level of imperfect reliability (.80) showed general improvements. Such findings are consistent with the recent integration of the literature, suggesting that reliability levels of 70% to 75% represent a rough "threshold" of imperfect reliability assistance (C. D. Wickens & Dixon, 2005). Although not all studies show that reliability levels below 70% are worse than having no automation at all (St. John & Manes, 2002), the majority of the studies examined in the literature do seem to indicate that this may be an emerging conclusion. Furthermore, this may have implications for other domains (outside of the UAV arena) that use diagnostic alerts, such as airport luggage screening and air traffic control.

Perhaps the most important implications of the current results go beyond those specific to UAVs and relate to the general implications of the designer's flexibility in setting the alerting threshold in multitask environments. On the one hand, by extending the findings of Maltz and Meyer (2003), these results reveal profoundly different effects on attention allocation and attention switching, the ultimate costs of which must depend on the importance of ongoing tasks and alerting tasks. In this context, attention appears to be driven by the cost of total (human and system) misses versus false alarms. On the other hand, the results provide some promise for the development of computational models of automation effects that can be employed in predicting human-automation interaction.

ACKNOWLEDGMENTS

REFERENCES

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19,* 775–779.

Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology, 13,* 249–268.

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms.* Hillsdale, NJ: Erlbaum.

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors, 47,* 479–487.

Galster, S. M., Bolia, R. S., Roe, M. M., & Parasuraman, R. (2001). Effects of automated cueing on decision implementation in a visual search task. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 321–325). Santa Monica, CA: Human Factors and Ergonomics Society.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics.* New York: Wiley.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Maltz, M., & Meyer, J. (2003). Use of warnings in an attentionally demanding detection task. *Human Factors, 43,* 217–226.

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors, 45,* 281–295.

Masalonis, A. J., & Parasuraman, R. (1999). Trust as a construct for evaluation of automation aids. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 184–188). Santa Monica, CA: Human Factors and Ergonomics Society.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors, 47,* 35–49.

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, 46,* 196–204.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38,* 211–322.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology, 3,* 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230–253.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332–336). Santa Monica, CA: Human Factors and Ergonomics Society.

Sorkin, R. D. (1989). Why are people turning off alarms? *Human Factors Society Bulletin, 32*(4), 3–4.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1,* 49–75.

Swets, J. A., & Pickett, R. M. (1982). *The evaluation of diagnostic systems.* New York: Academic Press.

Wickens, C. D., & Dixon, S. R. (2005). *Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature* (Tech. Rep. AHFD-05-01/MAAD-05-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York: Oxford University Press.

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,* 335–342.

Yeh, M., Merlo, J., Wickens, C. D., & Brandenburg, D. L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors, 45,* 390–407.

Stephen R. Dixon is an assistant professor of psychology at New Mexico State University. He received his Ph.D. in psychology from the University of Illinois in 2006.

Christopher D. Wickens is a retired professor of psychology at the University of Illinois at Urbana-Champaign. He received his Ph.D. in psychology from the University of Michigan in 1974.