

Unmanned Aerial Vehicle Flight Control: False Alarms Versus Misses

Stephen R. Dixon, Christopher D. Wickens, & Devron Chang
University of Illinois Aviation Human Factors Division
Savoy, Illinois

Thirty-two undergraduate pilots from the University of Illinois School of Aviation performed simulated military reconnaissance missions with an unmanned aerial vehicle (UAV). Pilots were required to: a) navigate the UAV through a series of mission legs, b) search for possible targets of opportunity, and c) monitor system health. They performed the missions under three types of auditory auto-alert aids (a 100% reliable system, a 67% reliable system with automation false alarms, and a 67% reliable system with automation misses), as well as a non-automated baseline condition. Results indicate that while reliable automation can benefit performance relative to baseline in the automated task, the unreliable automation aids reduced performance to that of baseline or worse. The automation false alarms and misses harmed performance in qualitatively different ways, with false alarm prone automation appearing to cause more damage than miss prone automation to both the automated task and the concurrent target search task.

INTRODUCTION

There are several advantages of unmanned aerial vehicles (UAV) over manned systems, including portability, cost efficiency, more radical flight control, longer missions, and safety (Draper & Ruff, 2000). Previously impossible civilian and military missions can now be conducted with UAVs (Gawron, 1998), with no danger to the pilots, who can control the UAV from a remote location (Mouloua et al, 2003).

The Army currently uses a two-operator team to fly its Hunters and Shadows. One operator is responsible for piloting the aircraft, while the other handles image inspection and systems monitoring. In order to increase efficiency, the Army would like to reduce the number of operators, while increasing the number of UAVs in service. To accomplish this goal, the ratio of pilot/UAV must be reduced from the current 2:1 to 1:1 or even 1:2. One concern with this concept is that of workload; that is, having one operator handle the responsibilities of two previous operators could have a detrimental effect on workload, such that the mission requirements may be jeopardized. A single operator would be responsible for aviating, navigating, monitoring system parameters, and target search and inspection with no help from an assistant operator. This inevitably leads to greater mental and physical workload demands on the pilot.

One solution for reducing the workload demands on the pilot is to incorporate automation aids which assist the pilot in handling concurrent tasks, and indeed Dixon, Wickens, and Chang (2003) observed a decrease in UAV operation workload (increase in concurrent task performance) associated with an automated alerting system. These automation aids, however, are not always perfectly reliable and can lead to different states of overtrust, undertrust, or calibrated trust (Parasuraman & Riley, 1997). The fear is that automation aids will be provided, with little concern as to their effects on trust and concurrent performance when the automation fails. It is

important to understand the consequences of these automation failures before they are incorporated into the design; otherwise, we may find that much time and resources have been wasted on a system that provides little or no benefit to the operator.

Automation detection/diagnostic aids may often be imperfect (Wickens, 2004), and designers must then choose where to set the “threshold” or response criterion on such aids; that is, to trade-off the relative frequency of automation misses versus false alerts. Meyer (2001, in press) and his colleagues (Cotte et al, 2001; Maltz & Shinar, 2003), have argued that the trade-off produces two different cognitive states. False-alarm prone automation reduces *compliance*, representing the so called “cry wolf” effect (Bliss, 2003; Breznetz, 1983), whereby all alarms (including true ones) may be responded to late, or possibly not at all. Miss-prone automation reduces *reliance* whereby spare visual attention, available for other tasks when there is perfect automation of the alerted task, must now be re-allocated to monitor the raw data of the alerted domain, at the expense of those concurrent tasks.

While some prior research has examined the two cognitive states (Cotte et al, 2001, Maltz & Shinar, 2003, Meyer, 2001), Dixon and Wickens (2004) appears to be the only study that has done so with a careful calibration of false alarm and miss rate within the multi-task context necessary to examine the dual task (attentional) consequences of the two cognitive states.

The current study examines both the level of reliability and the comparison between false alarms and misses induced by varying the threshold of a system monitoring alert system within the UAV. Pilots conducted UAV reconnaissance missions under four conditions: a) baseline, b) perfectly reliable auditory auto-alerts, c) 67% reliable auto-alerts with false alarms, and d) 67% reliable auto-alerts with misses. We hypothesized 1) that the lower reliability levels would result in

performance dropping to that of baseline, 2) that false alarms and misses would adversely affect both the automated task and concurrent tasks, but will do so in qualitatively different ways.

METHODS

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

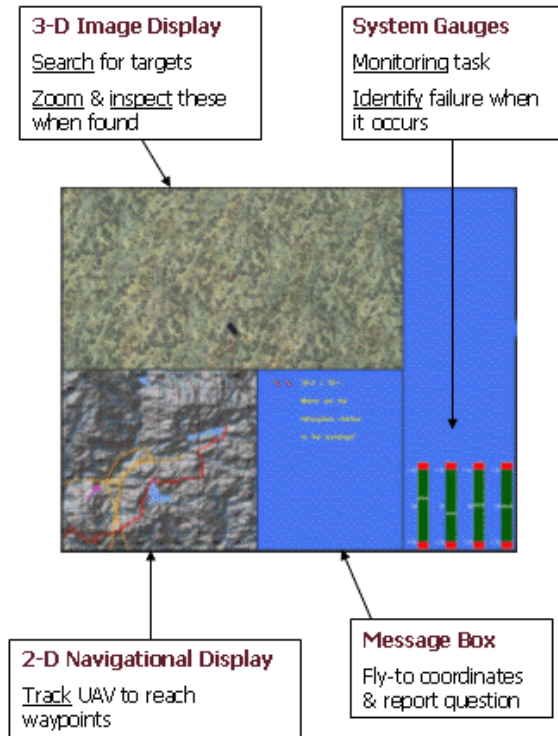


Figure 1. A UAV display with explanations for different visual areas.

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 flight instructions via the Message Box, including fly-to coordinates and a report question pertaining to the next command target (CT). These instructions were present for 15 seconds; in case the pilot forgot the instructions, pressing a repeat key refreshed the flight instructions for an additional 15 seconds.

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?*). This challenging CT report demanded motor, visual and cognitive resources (Gugerty & Brooks, 2001). Along each mission leg, pilots were also responsible for detecting and

reporting low-salience targets of opportunity (TOO), a task similar to the CT report, except that the TOOs were much smaller (1-2 degrees of visual angle) and camouflaged. TOOs could occur during simple tracking (low workload) or during a pilot response to a system failure as described below (high workload).

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

Automation aids, in the form of auditory auto-alerts during SFs, were provided for three of the four conditions. The A100 condition (A = automation, 100% reliable) never failed to alert pilots of SFs. The A67f condition (f = false alarm, 67% reliable) failed by producing 5 false alarms in addition to 10 true SFs. The A67m condition (m = miss) failed to notify pilots on 5 of 15 SFs. 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; they were simply told that the automation was either “perfectly reliable” or “not perfectly reliable.” They were not told whether the automation would produce false alarms or misses. Ratings of trust were given by each pilot at the end of the mission.

RESULTS

Mission completion

An ANOVA revealed no main effect for tracking error [$F(3, 27) < 1.0, p > .10$], revealing no differences between conditions. This finding is consistent with all previous findings in our UAV studies (e.g. Dixon, Wickens & Chang, 2003), and provides evidence that pilots treated the tracking task as primary, protecting that task from interference regardless of the flight conditions. An ANOVA on the use of the repeat key revealed a main effect of condition [$F(3, 27) = 4.64, p = .01$], indicating that the A67m condition produced more repeats than the other three conditions [all $p < .05$]. Apparently, pilots were so involved in trying to detect SFs which the auto-alert system missed that they reallocated resources they might otherwise have used for memory recall.

TOO monitoring

An ANOVA conducted on TOO detection accuracy revealed no effect of condition, [$F(3, 25) < 1.0, p > .10$], suggesting that neither type of automation failures, nor even perfectly reliable automation, had any effect on surveillance accuracy within the 3D image window. A second ANOVA on TOO detection RT shown in Figure 2, revealed a main effect of load

[F(1,15)=6.60, p=.02], suggesting that the high workload situations (concurrent with a system failure) were more detrimental to performance than the low workload situations (simple tracking and search). However, these results should be analyzed in the context of the load x condition interaction [F(3, 15) = 5.99, p = .01], which indicates that the A67f condition suffered more at high workload than the other conditions [all p < .05]. TOO report accuracy was not affected by condition [F(3, 25) < 1.0].

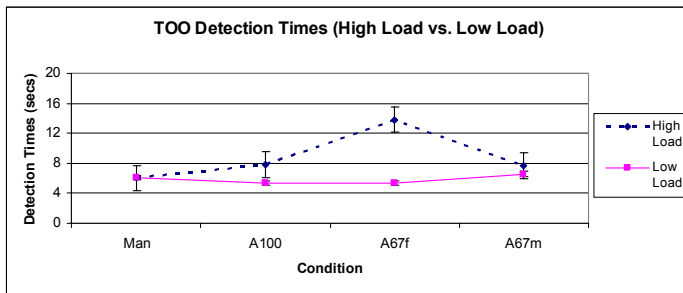


Figure 2. TOO detection times across condition and workload. SE bars are included.

System Failures

Figure 3 presents the overall SF detection accuracy as a function of workload (i.e., whether or not the SF appeared while the pilot was engaged in manipulating the 3D image window to report on a target). The ANOVA revealed a main effect of condition [F(3, 25) = 4.49, p = .01], and load [F(1, 25) = 11.21, p < .01]; however, a marginally significant interaction between load and condition [F(3, 25) = 2.28, p = .10] suggests that some conditions suffered more than others during the high workload trials. Further planned comparisons at high workload revealed that pilots detected fewer SFs in the false alarm condition than the other conditions [all p < .05].

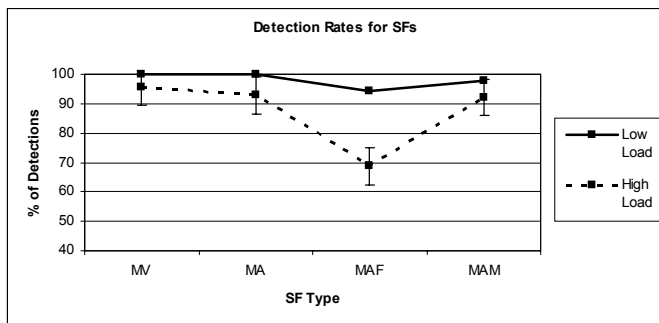


Figure 3. SF detection accuracy across condition and workload. SE bars included.

Figure 4 presents the overall SF detection times as a function of workload. An ANOVA reveals a main effect of load [F(1, 25) = 24.60, p < .001], as well as a marginally significant effect of condition [F(3, 25) = 2.59, p = .07]; however, these results must be taken in context of the marginally significant interaction between load and condition [F(3, 25) = 2.66, p =

.07], which suggests that the condition effect was only evident in the high workload trials.

Further comparisons revealed that the perfectly reliable condition facilitated faster detection times than the other conditions [all p < .05], and that the 67% reliable automation failed to provide any benefit above the baseline manual condition; in particular, the A67m condition resulted in slower detection times than baseline [p < .01] on those occasions when the automation failed to notify the pilots of a SF. We asked if the long RT was a result of “complacency” in response to the first failure encountered (Yeh et al, 2003). However this did not appear to be the case, since equally long RTs were manifest across all trials in which the SF alert failed to sound.

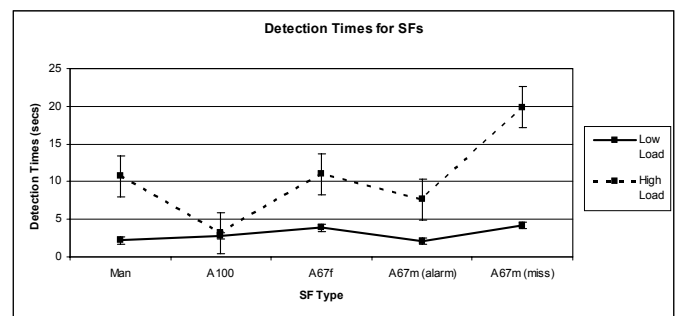


Figure 4. SF detection times across condition and workload. For this SF performance measure, the A67m condition is divided into two subgroups: a) Alarm (67% of the time) – when the auto-alert detected a SF, and b) Miss (33% of the time) – when the auto-alert missed the SF and failed to notify the pilot. SE bars are included.

SF report accuracy was measured collectively by what percentage of times pilots correctly determined which SF gauge had gone “out of bounds” and also by whether or not they entered the proper ownership coordinates where the SFs occurred. An analysis of variance revealed a main effect of condition [F(3, 27) = 3.63, p < .05], with all three auto-alert conditions supporting more accurate SF reports than baseline [all p < .05].

Pilot ratings of trust

At the end of the missions that introduced automation failures, pilots were asked to give subjective ratings of trust. Pilots accurately assessed the A100 condition to be perfectly reliable, but underestimated the true reliability levels of the A67f (38%) and the A67m (20%) conditions. This is in contrast to the more accurate pilot ratings of trust in unreliable automation reported by Dixon & Wickens (2004), with the difference being that in the current study, subjects were given much less prior knowledge of the reliability levels than in the previous study.

DISCUSSION

The imperfect automation employed here at 67% reliability generally supported performance that was as low, if not

sometimes lower, than the non-automated manual (baseline) condition. However we note that nearly all of the disruptions caused by these imperfections occurred in high workload circumstances, associated with the operator dealing with either a system failure or with a ground target identification. These are indeed circumstances in which automation dependence would be expected to be greatest, so that the costs of its failures would also be greatest.

According to the reliance-compliance dichotomy (Meyer, 2001; in press), we had expected qualitatively different effects of miss-prone versus false-alarm prone automation, on both the automated task and concurrent tasks. Such differences were observed, but only partially in the manner predicted in the context proposed by Meyer (2001; see also Dixon & Wickens, 2004).

Regarding automation false alarms, the data were clear that high false alarm rates reduced compliance, as operators missed more true failures, reflecting the “cry wolf effect”. Surprisingly however, we did not observe higher concurrent task performance here (compared to miss-prone automation), as we might have predicted had pilots, knowing that the alert would always sound whenever there was a failure, deployed full visual attention to the concurrent tasks. In fact, TOO RTs were prolonged by such false-alarm prone automation, as though pilots, receiving an automation alert believed to be imperfect, would spend extra time carefully examining the SF gauges to ascertain whether they really were or were not out of tolerance. Such examination was unnecessary in the miss condition, since any alert was instantly known to be true.

Regarding automation misses, the data were consistent with predictions in showing that one of the concurrent tasks – memory for target information – was more disrupted by miss-prone automation. This prediction is based on the notion that automation reliance avails more visual attention for concurrent tasks, and so a reduction in such reliance would reduce concurrent visual attention. However, there was little evidence that the other task (TOO monitoring) suffered with miss-prone automation, as would be predicted had pilots re-directed attention away from the TOO image window to visually monitor the raw data in the SF display with miss-prone automation. Apparently, they did not really do so, and continued to rely on the auto-alerts; therefore, their response times to the three SFs which automation missed remained long. Hence, reliance was not driven as much by automation miss rates as we had anticipated, although compliance was certainly driven by automation false alarm rates.

In summary, it appears from the data that FA-prone automation may be more disruptive overall, in that it harmed performance in both the automated task (SF detection rates) as well as the concurrent target inspection task (TOO detection times), in ways that the miss-prone automation did not. While miss-prone automation adversely affected performance in the automated task only during automation “misses” (i.e. longer SF detection times when the automation failed to alert the operator), the FA-prone automation affected performance in

the automated task even when the automation was correct (i.e. fewer SF detections even when the automation correctly alerted the operator). The miss-prone automation did cause more repeat requests for flight instructions, an effect not seen with the FA-prone automation; however, this penalty has little practical value, as pilots can easily refresh visual data during real-world missions without jeopardizing the mission.

The results from this study are generally consistent with those found by Dixon and Wickens (2004), in that both forms of automation failure (FA and misses), harmed overall performance in qualitatively different ways. In both studies, imperfect automation led to performance equal to, or worse than, baseline performance.

Practical implications of this study suggest that while automation can benefit performance relative to non-automated systems, the automation must be highly reliable in order to have any beneficial effects. Imperfect automation at reliability levels below 70% appears to be unproductive and costly, both in the resources used to develop it, as well as in the associated penalties of using it.

ACKNOWLEDGMENTS

This research was sponsored by a subcontract #ARMY MAAD 6021.000-01 from Microanalysis and Design, as part of the Army Human Engineering Laboratory Robotics CTA, contracted to General Dynamics. David Dahn was the scientific/technical monitor. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Federated Laboratory. The authors also wish to acknowledge the support of Ron Carbonari and Jonathan Sivier (in developing the UAV simulation), of Bobby Bernard and Mark Juntenen for assisting with data collection, and of Dr. Michael Barnes of the Army Research Lab at Ft. Huachuca, Arizona for assisting in interviewing UAV pilots within the E CO, 305th Military Intelligence Battalion to carry out the cognitive task analysis that underlies the simulation developed.

REFERENCES

- 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.
- 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. & Wickens, C.D. (2004). Automation reliability in unmanned aerial vehicle flight control. *In Proceedings of the 5th Human Performance, Situation Awareness and Automation Technology Annual Meeting*.

- Draper, M. H., & Ruff, H.A. (2000). Multi-sensory displays and visualization techniques supporting the control of unmanned air vehicles. *IEEE International Conference on Robotics and Automation*, San Francisco, California.
- 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.
- Gawron, V.J. (1998). Human factors issues in the development, evaluation and operations of uninhabited air vehicles. *Proceedings of the Association for Unmanned Vehicle Systems International (AUVSI)*, Huntsville, AL, 431-438.
- Gugerty, L., & Brooks, J. (2001). Seeing where you are heading: Integrating environmental and egocentric reference frames in cardinal directions judgments. *Journal of Experiment Psychology: Applied*, 7(3), 251-266.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*(46)1, 50-80.
- Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors*, 45, 281-295.
- 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.
- Mouloua, M., Gilson, R., & Hancock, P. (2003). Human Centered Design of Unmanned Aerial Vehicles. *Ergonomics and Design, Winter*, 6-11.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.
- 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.
- Wickens, C. D. (2004). Imperfect automation and CDTI alerting. Implications from literature and systems analysis for display design. *Aviation Space and Environmental Medicine*. 75, #4 section II. B-138.
- Yeh, M., Merlo, J., Wickens, C.D., Brandenburg, D.L., & Merlo, J. (2003). Head up versus head down.: The costs of imprecision, unreliability and visual clutter on cue effectiveness for display signaling. *Human Factors* 45(3), 390-407.