Automation-induced monitoring inefficiency: role of display location*

INDRAMANI L. SINGH, ROBERT MOLLOY AND RAJA PARASURAMAN†

Cognitive Science Laboratory, The Catholic University of America, Washington, DC 20064, USA

(Received 23 June 1995 and accepted in revised form 25 July 1996)

Operators can be poor monitors of automation if they are engaged concurrently in other tasks. However, in previous studies of this phenomenon the automated task was always presented in the periphery, away from the primary manual tasks that were centrally displayed. In this study we examined whether centrally locating an automated task would boost monitoring performance during a flight-simulation task consisting of system monitoring, tracking and fuel resource management sub-tasks. Twelve nonpilot subjects were required to perform the tracking and fuel management tasks manually while watching the automated system monitoring task for occasional failures. The automation reliability was constant at 87.5% for six subjects and variable (alternating between 87.5% and 56.25%) for the other six subjects. Each subject completed four 30 min sessions over a period of 2 days. In each automation reliability condition the automation routine was disabled for the last 20 min of the fourth session in order to simulate catastrophic automation failure (0% reliability). Monitoring for automation failure was inefficient when automation reliability was constant but not when it varied over time, replicating previous results. Furthermore, there was no evidence of resource or speed accuracy trade-off between tasks. Thus, automation-induced failures of monitoring cannot be prevented by centrally locating the automated task.

1. Introduction

Automation has become more advanced in recent years. Automation has been implemented to reduce operators’ workload and fatigue, improve safety, and allow faster and more precise control of multiple simultaneous tasks. However, several potential human performance problems have also emerged relating to the user’s interaction with automation technology. These problems include (1) a reduction in the operator’s system awareness, (2) an increase in monitoring workload, and (3) a degradation in manual skills. A reduction in manual control skills appears particularly critical, as degradation in this skill limits the ability of the operator to quickly resume accurate manual control of a process following automation failure (James, McClumpha, Green, Wilson & Belyavin, 1991; Parasuraman, Molloy & Singh, 1993).

Another potential cost of automation that has been noted is the problem of “complacency” (Wiener, 1981; Thackray & Touchstone, 1989; Singh, Molloy & Parasuraman, 1993). Wiener (1981) searched NASA’s aviation safety reporting

* This paper was presented at the 33rd Annual Conference of The Psychonomic Society, held in St. Louis, MO, USA from November 13 to 15, 1992, by the first author.
† Author for correspondence.

17
system (ASRS) database and found over 500 incidents of complacency that could be attributed to crew over-reliance on automated systems. The ASRS Coding Manual defines “complacency” as “self-satisfaction which may result in nonvigilance based on an unjustified assumption of satisfactory system state” (Billings, Lauber, Funkhouser, Lyman & Huff, 1976). Wiener (1981) called for empirical investigations of the phenomenon of complacency so that counter-measures could be developed.

Thackray and Touchstone (1989) designed an experiment to empirically validate the concept of “complacency” on an ATC monitoring task. Participants performed a simulated ATC task either with or without an automated aid message indicating aircraft-to-aircraft conflict situations. The automation failed on two occasions, early and late during a 2 h session. Thackray and Touchstone (1989) hypothesized that participants would be less efficient in detecting ATC conflicts with automated aids than when detecting without automated aids. However, participants detected ATC conflicts equally well in both conditions. Thus, they failed to obtain reliable evidence of complacency which may be related to participants in this study having responsibility for only a single task, monitoring.

Thackray and Touchstone (1989) may have failed to find evidence of poor monitoring of automation because participants performed only one task, a situation rare in the cockpit. As ASRS incident reports suggest (Billings et al., 1976; Mosier, Skitka & Korte 1994) many monitoring failures occur when the pilot is engaged in a multi-task situation. Parasuraman et al. (1993) therefore reasoned that monitoring of automation would be poor only under multiple-task and not single-task conditions. They tested nonpilot participants on a laboratory flight-simulation task consisting of three tasks: two-dimensional compensatory tracking, probability monitoring task of engine status, and fuel management. In the multi-task condition, participants were required to perform the tracking and fuel management tasks manually while an automation routine detected engine malfunctions. The automation routine failed from time to time and participants were required to detect and rest the fault. For two groups, the failure rate (reliability) for the automation was a constant 88% or a constant 62%. For an additional two groups, the reliability varied every 10 min between 62% and 88% with the order counter-balanced. Although participants in the constant failure rate condition normally had a detection rate of over 70% when performing the engine-status task manually, they detected only 33% of the automation failures. This effect was the same for both constant levels of automation reliability. This poor monitoring of automation was also present for the variable automation reliability groups but only for the first 10 min of performance under automation. Therefore, Parasuraman et al. (1993) found monitoring of automation to be poor under multi-task conditions.

Parasuraman et al. (1993) conducted a second study to show how the same automated monitoring task would be performed in a single task environment. Participants in this second study were only required to monitor the engine-status task for automation failures. When engine monitoring was the only task, detection was equally accurate (98%) and about as quick (2.5 s) during manual performance as it was under automated control. This result indicates that previous results finding good monitoring of automation were probably the result of the use of a single task (e.g. Thackray & Touchstone, 1989), a situation rare in the cockpit.

Using a different procedure, Knapp and Vardaman (1991) also examined the
development of complacency in a multi-task environment. In this study participants were required to perform an airspeed maintenance task, an altitude maintenance task, and a simple reaction time task (light detection). Yoked to the simple reaction time task was a second light that came on with the first light but was “automatically” extinguished. Participants believed the second light was a remnant of a previous experiment and would turn off when they turned off the first light. If the yoked light failed to turn off with the first light (automation failure) then participants had to manually turn the light off. During an automation failure, Knapp and Vardaman (1991) measured the time from the extinguishing of the first light to the extinguishing of the second light. They found that participants had significantly slower RTs to the automation failure than to the simple reaction time task, a result that was interpreted as evidence of complacency.

One possible explanation for the poor monitoring of automation may be that participants make insufficient eye fixations towards the automated display. Comstock (1993, pers. comm.), using the MAT battery, collected eye tracking data while participants performed in a multi-task condition. Participants performed the monitoring and fuel tasks manually while an automation routine performed the tracking task. Comstock found significantly fewer gazes by participants at the tracking task when it was automated than when it was manual.

In both the Parasuraman et al. (1993) and the Knapp and Vardaman (1991) studies the automated task to be monitored was not located centrally but to the left of the center of the display. It is possible that participants are poor monitors of automation only if an automated task is presented in the periphery, but not if it is displayed centrally. For example, in the Parasuraman et al. (1993) study, participants may have fixated the system monitoring window less often in the constant-reliability condition than in the variable-reliability condition. The display was such that if participants fixated the tracking or fuel-management windows the center of the monitoring window was about 5° away from fixation. Informal observation of participants using a video camera did not reveal any systematic deviation in the pattern of eye movements in the two automation conditions, but the possibility of small differences in scanning behavior between the two conditions could not be ruled out.

The present study examined whether the inefficiency in monitoring for automation failure noted by Parasuraman et al. (1993) can be overcome by central location of the automated task. If the monitoring task were displayed centrally, it should be fixated relatively frequently, and differences in scanning patterns should have less of an impact. More specifically, centrally locating the automated monitoring task should yield efficient monitoring of automation.

2. Method

2.1. PARTICIPANTS

Twelve volunteers (six males and six females) participated in this study. Participants ranged in age from 18–29 years, were right handed and had normal (20/20) or corrected to normal vision. Each participant was tested for 2 h on two separate days, and was paid $25 for completing the study. None of the participants had prior experience with the flight simulation task.
2.2. FLIGHT SIMULATION TASK

A revised version of the Multi-Attribute Task Battery (MAT) (Comstock & Arnegard, 1992) was used for each session (see Figure 1). The MAT is a multi-task flight simulation package that consists of tracking, monitoring, and fuel management tasks. This battery was used because: (1) the component tasks are similar to actual flight crew activity; (2) the tasks are dynamic, adding to the sense of realism; (3) the tasks are easily modifiable.

The locations for the three component tasks were changed from Parasuraman et al. (1993). The system monitoring task was moved from the upper-left corner to an upper-central position. The tracking task was moved from the upper-central position to the lower-right corner. Finally, the fuel management task was moved from the lower-central position with the pump status window on its left to the lower-left corner with the pump status above in the upper-left corner.

2.3. SYSTEM MONITORING

The system monitoring task consisted of four vertical gauges with moving pointers that represented temperature and pressure levels for each of two aircraft. In addition, there was a red “warning” light and a green “OK” light located above the vertical gauges. Under normal circumstances, the green light was on and the pointers fluctuated around the center line remaining within a plus or minus one
limit range of center. In each 10-min block of the simulation, 16 “system malfunctions” occurred at unpredictable intervals ranging from 13 to 72 s. When a malfunction occurred the pointer moved beyond the one limit range of center, either above or below. During manual sessions, participants were required to detect these malfunctions and reset them by pressing one of four corresponding functions keys which were labeled below each vertical gauge. Participants were required to reset the malfunctions within 10 s or the malfunction reset automatically and scored a miss. Feedback on correct detections was provided by the malfunctioning gauge returning to a center position and remaining there for 1.5 s.

During automated sessions, malfunctions were normally reset automatically, 4 s after the malfunction occurred. When the automation routine detected the malfunction the green OK light was extinguished and the red warning light lit. During this time, the participant’s response keys were disabled to prevent manual input. When the automation routine failed, the green light did not extinguish. Participants responded to this “automation failure” as they responded to the malfunctions during the manual sessions.

If participants responded appropriately to automation failure by pressing the appropriate function key, the response was scored as a correct detection. If participants failed to detect the automation failure within 10 s, the response was scored a miss. Finally, if participants responded when all gauges were normal, the response was scored a false alarm. Therefore, the performance measures for the system monitoring task were the probability of detecting an automation failure, reaction time (RT) for detection and the number of false alarms.

2.4. TRACKING

A first-order, two-dimensional compensatory tracking task with joystick control was presented in one window of the MAT display (see Figure 1). Dashed x- and y-axes were provided for reference. Within the window was a smaller dashed rectangle drawn around the center point of the window. A green circular target symbol representing the deviation of the aircraft from its course fluctuated within the window in the x- and y-directions according to a specified forcing function consisting of a sum of nonharmonic sine waves. The highest frequency of the forcing function was 0.06 Hz in this study. Control inputs were provided by a displacement joystick. The control dynamics were first-order, or velocity control. If no control input was applied, the aircraft symbol drifted away from the center towards the edges of the window. The participant’s task was to keep the aircraft within the central rectangle by applying the appropriate control inputs in the x- and y-directions. For example if the aircraft was to the right of center, a leftward joystick movement would cause the circle to return to the center. Participants were given training in first-order control by demonstrations of the effects of small and large control inputs (in either the x- or y-direction) on the speed of movement of the aircraft.

Operator performance of the tracking task was evaluated by sampling the x and y control inputs at 10 Hz and thus deriving the x and y deviations. The root mean square (RMS) error was then computed for the samples obtained over a 1 s period. In computing the combined horizontal and vertical deviations from the target, vertical deviations were converted (in proportion to the monitor x and y resolution)
to horizontal pixel units before combination with the horizontal deviations. RMS error scores for successive 1s epochs were averaged over a longer time period of performance (e.g., 15 min) to yield a mean RMS error score for a block.

2.3. FUEL MANAGEMENT

This task was meant to simulate the actions needed to manage the fuel system of the aircraft. Figure 1 displays the fuel (resource) management window. The six rectangular regions were tanks which held fuel. The green levels within the tanks represented the amount of fuel in each tank, and these levels increased and decreased as the amount of fuel in a tank changed.

Pumps connected the tanks so that fuel could be transferred from one tank to another in the direction indicated by the corresponding arrow and fuel line. The numbers underneath four of the tanks (Tanks A, B, C, and D) represented the amount of fuel in gallons for each of the tanks. This number was updated every 2 s as the amount of the fuel in the ranks increased or decreased. The maximum capacity for either Tank A or B was 4000 gallons. Tanks C and D contained a maximum of 2000 gallons each. The remaining two supply tanks had an unlimited capacity.

Participants were instructed to maintain the level of fuel in both Tanks A and B at 2500 gallons each. This critical level was indicated graphically by a tick mark in the shaded bar on the side of these two tanks. The numbers under each of these tanks provided another means of feedback for the participant. The shaded region surrounding the tick mark represented acceptable performance. Tanks A and B were depleted of fuel at the rate of 800 gallons per minute. Therefore, in order to maintain the task objective, participants transferred fuel from the lower supply tanks.

The process of transferring fuel was accomplished by activating the pumps. Each pump transferred fuel in one direction, as indicated by the corresponding arrow. These pumps turned on when the corresponding number key was pressed by the participant. Pressing the key a second time turned that particular pump off and so on. The pump status was indicated by the color of the square area on each pump. When that area was black, or lacking in color, the pump was off. A green light in this area indicated that the pump was actively transferring fuel.

The flow rates for each pump were presented in the “Pump Status” window. The first column of numbers represented the pump number, 1 through 8. When a pump was activated, its flow rate was presented next to the pump number in this window. When a pump was off, its flow rate was zero. Pump 1 and 3 transferred fuel at the rate of 800 gallons per minute. Pumps 2, 4, 5, and 6 transferred fuel at the rate of 600 gallons per minute and Pumps 7 and 8 at 400 gallons per minute.

At the onset of each flight simulation, Tanks A and B contained approximately 2500 gallons of fuel each and Tanks C and D contained approximately 1000 gallons of fuel each. All pumps were off at the beginning of the task, leaving all strategic action to the operator’s discretion.

Operator performance on the fuel management task was evaluated in a number of ways. Detailed records of the key presses that the operator makes were kept so that the particular strategy that the operator used (if any) to meet task objectives could
be ascertained. A global measure of task performance was also obtained by computing the mean RMS error in the fuel levels of Tanks A and B (deviation from the required level of 2500 gallons). Fuel levels were sampled and RMS error computed over a 30-s period. RMS error scores for successive periods were also averaged over a longer time period of performance (e.g., 10 min) to yield a mean RMS error score for a block. A second global measure of fuel management performance was the number of pump activations per block.

2.6. DESIGN
Automation reliability was defined as the percentage of system malfunctions correctly detected by the automation routine. The consistency of automation reliability over time was varied as a between-subjects factor (constant or variable reliability) and blocks (or sessions) as a within-subjects factor in a factorial design. Participants were randomly assigned to either a constant or the variable-reliability condition. In the present study, the reliability in the constant condition was 87.5% (14 out of 16 malfunctions detected) while in the variable condition the reliability changed from 87.5% to 56.25% every 10 min.

2.7. PROCEDURE
Each participant completed four 30 min sessions, each consisting of three continuous 10 min blocks, or 12 blocks in all. After receiving general instructions on the flight simulation task and completing a biographical questionnaire, participants were trained on the system monitoring, tracking, and fuel-management tasks. After a three-minute demo of all three tasks, each participant performed all three tasks manually for a 10 min period. Participants were requested to give equal attention to all three tasks. At the end of the manual session, feedback on performance was given for all three tasks. Before starting the automated sessions participants were instructed that the monitoring task would be automated and their primary tasks were the tracking and resource management. However, participants were also told the automation routine may not be 100% accurate. Further, they were told if the automation routine did miss a malfunction they were required to reset the malfunction. Performance feedback on the tracking and fuel-management tasks was provided at the end of each 30 min session.

In each automation-reliability condition the automation routine was disabled for the last 20 min of the fourth sessions (blocks 11 and 12) in order to simulate catastrophic automation failure (0% reliability). Thus, participants were required to intervene manually all the time following the total failure. However, instructions for the last session were the same as for the first three sessions, when participants were required to intervene only occasionally.

3. Results
3.1. MANUAL PRACTICE
The performance of participants on three MAT tasks during manual practice was analysed to ensure that the two automation-reliability groups did not differ in their initial performance capabilities. There were no significant group effects on any of the four dependent measures of performance on the system monitoring, tracking and
fuel management task (see Table 1). However, participants in the variable group were significantly slower at detecting malfunctions than the constant-reliability group \((t = -4.99; p < 0.05)\). This significant difference was not considered problematic since it was in the opposite direction to that predicted for the automated sessions (variable group faster than the constant group).

### 3.2. System Monitoring

Mean probability of detection of automation failure was higher for the variable reliability (mean = 0.80; SD = 0.23) than for the constant reliability group (mean = 0.41; SD = 0.39). A 2 (automation reliability) \(\times\) 12 (blocks) analysis of variance (ANOVA) of the detection rate showed that the effect of automation reliability was significant, \(F(1, 10) = 12.06, p < 0.006\). The variable reliability group showed consistently better performance than the constant reliability group (Figure 2). The performance of the constant reliability group improved in the total automation failure blocks (block 11 & 12) but did not approach the performance of the variable reliability group.

### 3.3. Reaction Time

RTs could not be computed for each of the 12 10-min blocks separately because many participants had 0% detection rates in at least one of the blocks. Hence RTs were computed for each of four sessions of performance by averaging across three successive blocks comprising each session. A 2 (automation reliability) \(\times\) 4 (sessions) ANOVA of the RT data gave no significant effects. Figure 3 shows the mean RTs for the constant-reliability and variable-reliability conditions. The lack of significant effects of automation reliability on RT indicate that the detection rate difference between constant and variable reliability conditions is unlikely to be the result of a speed–accuracy tradeoff.
Figure 2. Effects of automation reliability (constant or variable) on probability of detection of automation failures. ●: Constant; ○: variable.

Figure 3. Effects of automation reliability (constant or variable) on reaction time for detection of automation failure. Key as for Figure 2.
3.4. FALSE ALARMS
The majority of participants in both the constant and variable reliability conditions had no false alarms in many of the 10 min 12 blocks. The mean numbers of false alarms for constant and variable reliability conditions, were 5.83 (SD = 7.70) and 6.33 (SD = 4.08), respectively.

3.5. TRACKING
Tracking performance demonstrated improvement across sessions, as indicated by a significant effects due to sessions, $F(3, 30) = 4.64, p < 0.008$. Both the constant and the variable reliability groups improved across sessions (Figure 4). There was no significant difference in tracking error between reliability groups nor was there a significant reliability by session interaction.

3.6. FUEL MANAGEMENT
The fuel mean RMS error decreased across sessions, $F(3, 30) = 3.51, p < 0.02$. The main effect of automation reliability and interaction with session (Figure 5) were not significant.

4. Discussion
The present results are consistent with the findings of Parasuraman et al. (1993). The probability of detecting an automation failure under multi-task conditions was poor when automation reliability was unchanging over time compared to when it varied from block to block. Moreover, monitoring performance under automation was
inferior to performance of the same task under manual conditions. Contrary to our expectations, centrally locating the monitoring display did not affect the pattern of results appreciably, indicating that the automation “complacency” effect discovered by Parasuraman et al. (1993) is a relatively robust phenomenon. The results also suggest (but do not prove since eye movements were not recorded) that poor monitoring is not necessarily due to reduced eye fixations to display locations in the periphery. Even an automated task presented in foveal vision can be monitored inefficiently under certain conditions.

The magnitude of monitoring inefficiency was almost as large as in the previous study in which the automated task was located peripherally. In that study the mean probability of detection of automation failure was 0.82 in the variable-reliability condition and 0.33 in the constant-reliability condition, a 60% reduction. With central display location, performance in the variable-reliability condition was approximately the same at 0.80, whereas detection probability for the constant-reliability condition was somewhat increased (although not significantly) at 0.41, a 49% reduction. This relatively large effect testifies to the importance of the phenomenon.

Given that monitoring of automation was not increased by centrally locating the display, what caused the poor monitoring of automation? One possible explanation is that participants were not allocating sufficient resources to the monitoring of the automation (Wickens, 1984). However, neither tracking nor fuel-management performance differed as a function of automation reliability, suggesting that differential allocation of processing resources cannot account for the superiority of monitoring performance in the variable-reliability condition. Furthermore, there

**Figure 5.** Effects of automation reliability (constant or variable) on fuel RMS error. Key as for Figure 2.
were no differences in reaction time between the two reliability conditions indicating no speed accuracy trade-off.

Another possibility is that participants were too trusting of the automation. Riley (1994) found trust to be positively correlated with automation usage. Participants, in his study, had to play a game either with or without automated control. When the aid was trustworthy, operators were more likely to use automation. However, when the automation performed at chance level participants lost trust in the automation and turned it off. Lee and Moray (1992) found similar results with a process simulation task.

In the current study, participants may have begun the automated sessions with an equal amount of trust in the automation. However, as the reliability of the automation fluctuated for the variable group their trust may have declined. Therefore, the variability group may have been more skeptical of the automation, and, thus, been more vigilant for automation failures. The constant group’s performance would have stayed low since they never developed a lack of trust for the automation. Further evidence for this comes from the increase in performance participants in the constant reliability group showed following the catastrophic failure of the automation in the final twenty minutes of the experiment. Of course, this theory is based upon inferred levels of trust. Further research, in which trust is empirically measured, needs to be conducted.

4.1. PRACTICAL IMPLICATIONS

The problem of over reliance on automation is known in the aviation industry. James et al. (1991) found that British pilots believe that pilots of automated aircraft rely too heavily on automation. The results of the present study provides empirical evidence for this over-reliance on automation. However, there are two potential criticisms toward the application of these experimental results to the aviation realm.

The first potential criticism concerns the low automation reliability (88%) used in the current automation routine. While such a reliability would be unacceptable in any real system, we used the value because of time constraints on data collection. Using realistic reliability rates (over 99%) would have required numerous sessions in order to collect reliable data. Further, the current results may underestimate the problem of over-reliance on automation. May, Molloy and Parasuraman (1993) examine several reliabilities of automation ranging from 12% to 98%. They found an inverse relationship between automation reliability and automation failure detection performance. Thus, one would expect operators using realistic automation failure rates would be less likely to detect failures than the participants of the current study.

A second potential criticism is that the type of automation problems presented in the current study are unrepresentative of realistic automation errors in aviation. The current study defined automation failures as a failure of the automation to reset a system malfunction. However, it is rare that cockpit automation fails outright in performance or catastrophically. In a survey of ASRS incident reports involving automation. Mosier et al. (1994) reported that the automation was doing exactly what it was supposed to in a majority of these incidents. How then did the incidents occur?

One way automation can perform perfectly and still “surprise” the crew is in the case of sudden mode reverion or a mode error by the crew (Sarter & Woods, 1995).
An example of this in the Bangalore crash of an Indian Airlines A320. In that case, the pilot put the automation in the “open descent” mode prior to landing the plane. In this mode the engines are set to idle and speed is controlled via pitch (in the correct mode thrust controls speed). In order to maintain target airspeed, the crew had to fly the plane well below descent profile. Further, due to the engines being at idle, the pilots were unable to recover the craft when the problem was discovered. While the automation did not fail per se, it did not behave the way the crew expected it to behave.

In both the current study and the Indian Airlines example, the operators over-relied on the automation. Where the differences lies is in the reason the automation failed to help the operator. The automation in Indian Airlines failed the crew in that the crew trusted the automation to follow the correct descent profile. It did not. Likewise, the participants in our study trusted the automation routine to reset all malfunctions. It, also, did not.

One motivating factor in doing this study was to examine the possibility of changing display location as a counter-measure to poor monitoring of automated tasks. We reasoned that if monitoring was only poor for peripheral tasks, then centrally locating an automated task might overcome the problem. However, the present results indicate that at least for the flight tasks examined in this experiment, display location does not influence monitoring of automation. Moreover, even if central display location had been found to be effective, it would still only provide a limited countermeasure, because many tasks are automated in the cockpit and not all can be displayed centrally.

Another possible counter-measure is presenting the automated task in a heads-up display. While the display changes in this study presented the automated task in a foveal position, participants were required to visually fixate on the automated task. One way to force fixation is to superimpose the monitoring task upon a task that requires frequent monitoring, such as the tracking task. Such a study is currently planned.

Finally, automation complacency could be reducible through training. Pilots of the Airbus A320 are being trained to prevent overreliance and overconfidence in an aircrafts capabilities (Stix, 1991). Further, pilots are also being trained to take a more active role in the automated flight deck. French pilots are trained to intermittently take control from the automated system, thereby increasing their awareness of the situation. All of these measures may provide countermeasures to automation-induced complacency.

Research for this article was supported by research grant NAG-17 from the Aeronautics and Space Administration (NASA), Langley Research Center, Hampton, VA. to The Catholic University of America. Alan Pope was the technical monitor. The views presented in this article are those of the authors and not necessarily representative of NASA.

References


operator workload and strategic behavior research. Technical Memorandum 104174. NASA Langley Research Center, Hampton, VA, U.S.A.


Paper accepted for publication by Editor, B. R. Gaines