Int. J. Human-Computer Studies (1999) **51**, 991–1006
Article No. ijhc.1999.0252
Available online at http://www.idealibrary.com on IDEAL

Does automation bias decision-making?†

LINDA J. SKITKA
Department of Psychology, University of Illinois at Chicago, 1007 W. Harrison St., Chicago, IL 60607-7137, USA. email: lskitka@uic.edu.

KATHLEEN L. MOSIER
San Francisco State University

MARK BURDICK
San Jose State University Foundation/NASA Ames Research Center

(Received 30 June 1998 and accepted in revised form 22 February 1999)

Computerized system monitors and decision aids are increasingly common additions to critical decision-making contexts such as intensive care units, nuclear power plants and aircraft cockpits. These aids are introduced with the ubiquitous goal of "reducing human error". The present study compared error rates in a simulated flight task with and without a computer that monitored system states and made decision recommendations. Participants in non-automated settings out-performed their counterparts with a very but not perfectly reliable automated aid on a monitoring task. Participants with an aid made errors of omission (missed events when not explicitly prompted about them by the aid) and commission (did what an automated aid recommended, even when it contradicted their training and other 100% valid and available indicators). Possible causes and consequences of automation bias are discussed

1. Does automation bias decision-making?

Computers and related automated decision aids have been introduced into many work environments with the explicit goal of reducing human error. For example, in response to the fact that many aviation accidents can be attributed to human error (e.g. Diehl, 1991), the aviation industry and federal aviation and safety industries are successfully pushing to increasingly automate flight systems (Weiner, 1989; Billings, 1996). Flight management computers are assuming greater control of flight tasks, such as calculating fuel-efficient paths, navigation, detecting system malfunctions and abnormalities, in addition to flying the plane. Other fields as disparate as nuclear power plants and intensive care units are similarly relying more and more on automated decision aids. Because these aids are generally quite accurate, airplanes fly safely, patient status is accurately monitored and power plants run more efficiently.

†This research was supported by NASA Ames #NCC 2-237, and was conducted at the University of Illinois at Chicago.

1071-5819/99/110991 + 16 $30.00/0

The introduction of computers as system monitors and automated decision aids into human decision-making environments is likely to have an important impact on the decision-making context. Although introduced with the expressed goal of reducing human error, remarkably few evaluation studies of whether automated decision aids in fact lead to an overall reduction in error rates have been conducted. In fact, there is some evidence to suggest that the introduction of automated decision aids does not unilaterally lead to a reduction in human error, but instead often creates opportunities for simply a different class of errors. For example, various problems with automated decision aids have been identified, including mode misunderstandings and mode errors; failures to understand automation behavior; confusion or lack of awareness concerning what automated aids are doing and why; and difficulty associated with tracing the reasoning processes and functioning of automated agents (e.g. Sarter & Woods, 1993; Billings, 1996).

The goal of the present paper is to compare people's performance on the same task with and without an automated aid, and to explore the extent to which people use automated decision aids in biased ways. We suggest that the presence of automated decision aids might reduce one class of errors (those that the automated decision aid has been explicitly programmed to detect and make recommendations about under normal functioning conditions), but that they introduce the possibility of making new kinds of errors. We propose that especially in the context of introducing automated decision aids to explicitly reduce human error, people become primed to use decision aids in biased ways. Rather than necessarily leading to fewer errors, automated decision aids may simply lead to different kinds or classes of errors.

For example, there has been considerable documentation of the fact that people tend to be "cognitive misers". That is, most people will take the road of least cognitive effort, and rather than systematically analyse each decision, will use decision rules of thumb or heuristics (for a review, see Fiske & Taylor, 1994). Automated decision aids may act as one of these decision-making heuristics, and be used as a replacement for more vigilant system monitoring or decision making.

Sharing system monitoring tasks and decision making with a computer may also have psychological parallels to sharing tasks with humans. A considerable body of research indicates that people tend to expend less effort when working collectively than when working individually (see Karau & Williams, 1993 for a review). For example, when asked to make as much noise as they could by cheering and clapping, people produced much more noise when alone than when with others (Latané, Williams & Harkins, 1979). Similarly, individual performance on a lexical decision task was higher when tested alone, as compared to in a group of others (Pratarelli & McIntyre, 1994). In short, people often slack off in the presence of others who are sharing in task responsibilities. Given that people treat computers who share task responsibilities as a "team member", and show many of the same in-group favoritism effects for computers that they show with people (Nass, Fogg & Moon, 1996), it may not be surprising to find that diffusion of responsibility and social loafing effects also emerge in human-computer interaction. To the extent that some tasks are shared with computerized or automated decision aids people may well diffuse responsibility for those tasks to those aids, and feel less compelled to put forth a strong individual effort.

Finally, people may respond to computers and automated decision aids as decision-making authorities. Obedience can be defined as people's willingness to conform to the demands of an authority, even if those demands violate people's sense of what is right (Berwer & Crano, 1994). Considerable evidence points to the fact that people will do harm to others if ordered to do so by an authority figure. For example, 21/22 nurses (more than 95%) administered a dangerous dose of a drug to patients when ordered over the phone to do so by an unfamiliar physician (Hofling, Dalrymple, Graves & Pierce, 1966; cf. Rank & Jocobson, 1977). Given that computers and automated decision aids are introduced into many work environments with the articulated goal of reducing human error, they may well be interpreted to be smarter and more authoritative than their users. To the extent that people view computers and automated decision aids as authorities, they may be more likely to blindly follow their recommendations, even in the face of information that indicates they would be wiser not to.

Based on the possible influence of cognitive laziness, social loafing and diffusion of responsiblity, and possible belief in the relative authority of computers and automated decision aids, we proposed that the presence of automated decision aids was probably associated with two kinds of errors: *commission errors* and *omission errors*. Errors of omission result when decision-makers do not take an appropriate action, despite non-automated indications of problems, because they were not informed of an imminent system failure or problem by an automated decision aid. Errors of commission occur when decision-makers follow automated information or directives, even in the face of more valid or reliable indicators suggesting that the automated aid is not recommending a proper course of action. Commission errors can be the result of not seeking out confirmatory or disconfirmatory information, or discounting other sources of information in the presence of computer-generated cues (Mosier & Skitka, 1996).

There are a variety of anecdotal examples that suggest that people may use automated decision aids in biased ways. For example, the crash of an Eastern Airlines aircraft in 1972 may have been the result of an omission error. During the time immediately prior to crashing, the crew was preoccupied trying to determine why the landing gear indicator did not light when the gears were dropped. The autopilot was set to hold an altitude at 2000 ft, but was accidentally disengaged by a nudge to the control column. The crew did not know the autopilot had disengaged until prompted by Air Traffic Control to check their altitude: by that time they had descended to 30 ft above ground level and it was much too late to make a correction (NYSB Report AAR-86-03 in Billings, 1996). The crew in this example had allocated responsibility for maintaining altitude to the automation, and subsequently neglected to check whether it was operating correctly until prompted by air traffic control.

Similarly, the problems with the 1983 Korean Airlines plane that was shot down by Soviet fighters can also be traced in part to a lack of crew vigilance in monitoring of automated systems. In this case, the crew selected a magnetic heading and followed it throughtout the flight rather than coupling the navigational system's inertial reference system to the autopilot. The flight did not follow its originally planned flight path, but instead maintained the selected magnetic heading until it was shot down. The crew was relying entirely on automation that had been inappropriately set up and never checked their progress manually, allowing the flight to stray well into Soviet airspace ("Analysis of flight data", 1993).

Experimental evidence of automation bias leading to commission errors was provided by a full mission simulation in the NASA Ames Advanced Concepts Flight Simulator

(Mosier, Palmer & Degani, 1992), although sample sizes were not sufficiently large to justify hypothesis testing ($N = 12$ crews). During one phase of flight, crews received contradictory fire indications. An auto-sensed checklist suggested that the crew shut down the #1 engine, but traditional engine parameters indicated that the #2 engine was actually more severely damaged. Seventy-five percent of the crews in the auto-sensing condition shut down the #1 engine, whereas only 25% with a traditional paper checklist did likewise. Analysis of the crews' audiotapes also indicated that crews in the automated condition tended to discuss much less information before coming to a decision to shut down the engine, suggesting that automated cues short circuited a full information search.

Another study strategically set up circumstances to examine the extent to which errors of omission and commission occur. Mosier, Skitka, Heers and Burdick (1998) examined a sample of 25 pilots who flew a simulated flight from Los Angeles to San Francisco airports, and then from San Francisco to Sacramento, after being completely trained in the part-task. The part-task itself had displays very similar to those found in a B747-400, and was therefore familiar to the participants.

Four flight events represented opportunities for omission errors: an altitude clearance misload; a heading change that loaded correctly, but was incorrectly executed by the flight system; a frequency change misload; and a tracking task automation failure (similar to an autopilot failure). Information about the automation failure was available on traditional display gauges or indices, as it would be on a traditional aircraft. In addition to these omission error opportunities, one event represented a commission error opportunity. Pilots were presented with a computer-based message (an EICAS alert) indicating that an engine was on fire. The engine fire message was contradicted by normal engine parameters and the absence of any other indicators (five other indicators would be present during an engine fire, a fact that pilots were reminded about during their training on the part-task). Pilots had to determine whether there really was a fire and decide whether or not to shut down the supposedly affected engine.

Results revealed an omission error rate of 55%. Omission error rates were related to pilot's total flight hours and years of experience, indicating that increased experience decreased the likelihood of catching automation failures. Virtually all participants shut down the engine in response to the false fire message, indicating a 100% commission error rate. These same pilots on the post-experimental questionnaire nonetheless indicated that an EICAS message without other associated cues would not be sufficient to diagnose a fire, and that in the absence of additional indices of fire, it would be safer under these circumstances not to shut down the engine. Further evidence emerged to support the idea that automated directives may bias how people process information: 67% of the pilots had a false memory of at least one additional cue that was not actually present in the engine fire incident; many remembered more than one cue.

Taken together, these results suggest that when people have an automated decision aid available, that they do as it directs. The presence of automated cues appears to diminish the likelihood that decision makers will either put forth the cognitive effort to seek out other diagnostic information or process all available information in cognitively complex ways (cf. Layton, Smith & McCoy, 1994; Parasuraman & Riley, 1997). In the absence of an automated directive, conversely, people often do nothing, regardless of what other system indices imply should be done.

Although the results of Mosier *et al.* (1998) are highly suggestive that errors of omission and commission occur in highly automated environments, because people's behavior was not compared across both automated and non-automated contexts, it remains difficult to judge: (a) how severe of a problem automation bias may be, and (b) whether automation bias is in fact due to *automation*. The present study was designed to address these questions by exploring the extent to which omission errors are unique to automated settings by comparing performance in a flight part-task with and without an automated decision aid.

It was hypothesized that participants with an automated decision aid would show vigilance decrements relative to participants on the same task without an automated decision aid. Possible explanations for this effect were explored (e.g. social loafing/diffusion of responsibility, preceived authority/accuracy of computerized decision aids).

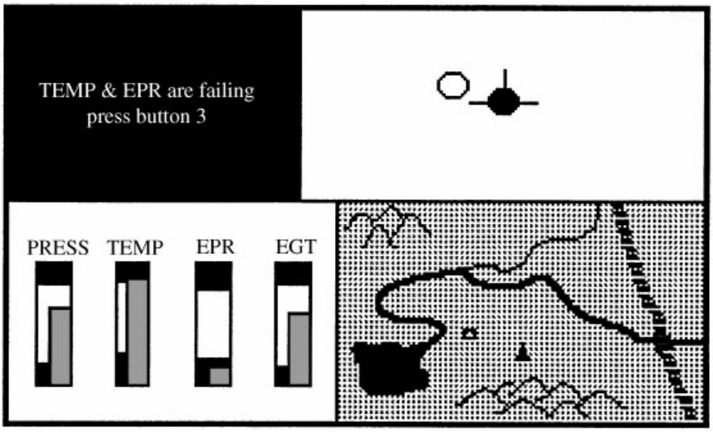
2. Method

2.1. PARTICIPANTS

Eighty undergraduate students paricipated in partial fulfillment of course requirements.

2.2. TASKS

Participants' primary task was to complete eight "flights" or trials using the Workload/PerformANcE Simulation software (W/Panes) developed by NASA Ames Research Center (1989). This program presented participants with a set of tasks designed to simulate the types of monitoring and tracking tasks involved in flying commercial aircraft. Participants were exposed to four quadrants of information using a 486/33 Personal Computer, and 14″ color monitor (see Figure 1: Note—although the figure is black and white, the participant's screen was in color).



TEMP & EPR are failing
press button 3

FIGURE 1. Primary task display.

2.2.1. The tracking task

Participants used a two-axis joystick to keep their own-ship symbol (the circle with a line through it represented in the top right quadrant of Figure 1) aligned with a moving circular target. The target circle moved as a function of the disturbance imposed by a sum of sine's algorithm. Therefore, participants' goal was to keep the target circle centered around the own-ship symbol by following the motion of the target circle with the joystick, compensating for movements away from the center in heading (horizontal) and altitude (vertical). This task ran continuously throughout each of the eight trails, and required the greatest consistent attention from the participant.

2.2.2. Waypoints

In addition to having to maintain their tracking performance, participants were also required to monitor their location on a map (presented in the lower right quadrant of Figure 1). A pink square representing the own-ship traveled from left to right across the map. Red and blue triangles on the map represented "waypoints". Each time their symbol passed one of these waypoints, participants were trained to push a button (presumably to simulate radioing their location to a communications tower).

Specifically, when the bottom-right corner of the pink square touched the top of the triangle, the participant either pushed the Beta frequency button (if the triangle was blue) or the Alpha frequency button (if the triangle was red). Response buttons were labeled with red or blue dots to facilitate correct responses. There were four waypoints on each trial. The layout of the map and the location of the waypoints varied across trials.

2.2.3. Gauge events

The lower-left quadrant displayed four analog slide-rule gauges that were used for a gauge-monitoring task. Each gauge had an upper and lower red zone. When a given gauge went into a red zone (gauge indicators moved continuously, but stayed within the normal range the vast proportion of the time), participants were instructed to press the gauge's corresponding button on their response box; if all three gauges went into a red zone at the same time, participants were trained to push a "gauge reset" button. Gauges were programmed to return to a "safe" value after 10 s in the red zone if no button was pressed (i.e. if the participant missed an event). The order of gauge events and when they occurred were randomized across the eight trials.

2.2.4. Automated monitoring aid

The top-left quadrant was reserved for AMA announcements. The AMA notified participants of events (e.g. three gauges moving into the red zone), and recommended specific courses of a action (e.g. to push the gauge-reset button).

2.3. OVERVIEW

The critical manipulation of the present study was whether participants had the assistance of the automated monitoring aid (AMA) during their session. All the participants in the automated condition had a functional AMA that provided them with messages that prompted them about system events. Participants in the automated condition were told that the AMA was not completely reliable and could give faulty information about an event, or fail to detect some events, but all their other instruments were always 100% accurate. Participants in the non-automated condition were given no information about an AMA, and did not receive any automated prompts about system events or states. Six out of 100 events across eight trials represented omission error opportunities. On these six trials, the AMA failed to prompt participants of a system event. Of critical interest was how many of the participants in the automated condition missed these events relative to participants in the non-automated condition.

2.4. PROCEDURE

Participants were run individually in a cubical with a computer and response box. Each participant was trained on the display and W/Panes tasks, and how to respond to each type of event. After being given complete verbal instructions and addressing any questions, the experimenter retreated to a separate station. At this time, each participant performed two 5-min full-task practice trials; performance feedback was provided at the end of each trial.

After a 3-minute break, participants began the experimental trials. There were eight experimental trials that lasted 8 min and 20 s each. Twelve orders of experimental trials were used to control for any possible event order effects. Across all eight trials, there were 22 waypoint events, 22 three-guage reset events, and 56 single-guage reset events. Participants were given a 5-min break at the end of the fourth trial. At the end of the experimental trials, participants completed a questionnaire that tapped their attitudes about the task, as well as manipulation checks.

2.5. DEPENDENT MEASURES

Six events were of critical interest across conditions. Across all events, participants in the non-automated condition received no prompts from an automated monitoring aid (AMA). Participants in the automated condition received prompts from an AMA on all but six events, which allowed six opportunities for participants to make an omission error. Because events were constant across automated and non-automated conditions, these six events represented the comparison points where relative levels of vigilance could be compared: how many people responded correctly to these events in the automated, as compared to the non-automated, condition?

Two occurrences of each of the three types of omission error opportunities occurred for each participant: (1) no directive to squawk the appropriate frequency when the aircraft symbol passed a waypoint, (2) no directive to press the RESET button when three gauges went into their critical zones simultaneously and (3) no directive to press the corresponding button when a single guage went into its critical zone. Responses to these events were scored as “miss” or error only if the participant failed to respond to that event. Any response, correct or incorrect, indicated awareness of the event, and therefore was not coded as an omission error.

In addition to these primary measures of interest, all participants in the automated condition were also presented with six opportunities to make commission errors. Commission error opportunities were characterized by an AMA directive that was contradicted by the other system indices. Because participants were aware that the other system indices were 100% accurate, and that the AMA was not perfectly reliable, the rational response would be to follow the other indices. The six commission error opportunities, similar to the omission error opportunities, were equally distributed across system events and across all eight trials. A response was scored as a commission error if it was consistent with what the AMA directed. Commission errors are only logically possible within the context of an AMA, so their occurrence could not be compared as a function of automation condition. Instead, the relative extent of commission errors provided descriptive baseline information only.

Other measures included a non-error performance measure, which was operationalized as the total number of correct responses for the 88 non-error events, and tracking performance, measured as the weighted average of the linear-deviations between the tracking circle and the ownship symbol in feet (for altitude error) and degrees (for heading error).

Perceptions of the experiment were tapped with a post-experimental questionnaire. Task difficulty and comfort were assessed with four-items. Respondents were asked to respond on a 1 (strongly disagree) to 7 (strongly agree) scale to the following items: "I found the tasks involved in the study to be difficult", "I was very comfortable with the experimental task", "I felt nervous about my performance", and "It was difficult to concentrate". These items were scaled by averaging across items (after reversing the comfort item to ensure that all items moved in the same direction) and had a Cronbach's alpha of 0.85 indicating that this measure had good internal reliability. Effort was measured by three-items using the same response scale: "I tried to improve my tracking performance on every trial"; "I tried to respond to waypoints faster on every trial"; and "I tried to respond to guage events faster on every trial". These items were scaled by averaging across the items, and the resulting scale had a Cronbach's alpha of 0.86. Two items tapped perceptions that the AMA allowed participants to respond more accurately: "The AMA helped me to respond more accurately to the waypoint task"; and "The AMA helped me respond more accurately to gauge events", Cronbach's alpha = 0.95. Two other items tapped the extent to which participants used the AMA as a means to diffuse responsibility: "The AMA allowed me to off-load responsibility for guage events", and "The AMA allowed me to off-load responsibility for the waypoint task", and had a Cronbach's alpha of 0.96.

Participants were also asked about the relative accuracy of the AMA and other system indices, to estimate the proportion of times that the AMA failed to prompt them about a system event, and to estimate the proportion of times it incorrectly prompted them to do something.

3. Results

3.1. MANIPULATION CHECKS

All of the participants in the automated condition correctly recalled that the gauges and other system indices were 100% accurate, but 3 out of 40 participants in the automated condition reported that the AMA was 100% reliable, when in fact they were instructed (and in reality) it was not. Because these same participants gave estimates greater than 0% for how often, if ever, the AMA failed to notify them of an event, and for how often, if ever, the AMA gave an inappropriate directive, they nonetheless were retained for analysis because they did recall something less than perfect reliability.†

Respondents in the automated condition on the whole underestimated the reliability of the AMA. In reality, the AMA failed to notify participants of an event 6 times across 100 trials, or was 94% reliable in giving notification of system events. Participants on average believed the AMA was only 81.35% reliable in giving notification of system events. The AMA also was 94% reliable in giving correct directives when directives were given; participants' average estimate was that it was 81.89% accurate in doing so.

3.2. DO PEOPLE MAKE MORE ERRORS OF OMISSION IN AUTOMATED THAN NON-AUTOMATED SETTINGS?

Does the presence of automated decision aids lead to decrements in vigilance and higher rates of omission errors relative to when automated aids are not available, even though participants on the whole did not feel the AMA was very reliable? Put simply, the answer is yes. Participants in the automated condition missed more of the comparison events than did participants in the non-automated condition.

The critical comparison for this hypothesis involved the six trials for which neither the automated or non-automated conditions received any prompt about system states. Results revealed that participants in the automated condition missed more of these events ($M = 2.44$ or a 59% accuracy rate) than those in the non-automated condition ($M = 0.18$ or a 97% accuracy rate), $F(1,62) = 44.32$, $p < 0.05$, $\mathrm{MSE} = 2.28$. In sum, strong support was found for the hypothesis that people were less vigilant in environments with automated decision or monitoring aids, relative to those without an automated aid.

3.3. COMMISSION ERRORS IN THE AUTOMATED CONDITION

In addition to exploring the extent to which omission errors were unique to automated contexts rather than a reflection of a generalized tendency toward complacency, the present study also investigated the base rate tendency to make errors of commission. Within the automated condition, there were six opportunities for participants to make an error of commission. It will be recalled that participants understood that the AMA was not 100% reliable, but that their gauges and other indices were always accurate. Therefore, when the AMA suggested doing something that was contradicted by the gauges or their waypoint map, participants should have followed the gauge or map, and not the AMA.

Analysis of participant responses across the six commission error events indicated that on average, participants made 3.92 commission errors across the six error events; an average accuracy rate of 35%. Only one participant made no commission errors; 23.1% of the participants made commission errors on all six events. In short, these results indicated that not only are omission errors a likely occurrence in automated contexts, but also that commission errors are highly probable events as well.

†The pattern of results remained the same regardless of whether these three participants were included or excluded from the analysis.

3.4. OTHER PERFORMANCE MEASURES

On further examination of performance across the automated and non-automated conditions, the responses across all non-error opportunity trials ($N = 88$) were compared. Not surprisingly, on events where the AMA gave correct direction, participants in the automated condition made more correct responses ($M = 83.03$) than those in the non-automated condition who had no such assistance ($M = 71.85$), $F(1, 77) = 26.37$, $p < 0.05$, MSE $= 93.51$.

Interestingly, no significant differences emerged on tracking performance as a function of automation condition, $F(1, 77) = 1.89$, *ns*, MSE $= 12991.01$. Therefore, having an AMA did not free participants' cognitive resources sufficiently to facilitate tracking performance. However, people who delegated responsibility for system monitoring to the AMA and suffered a subsequent increase in commission and omission error rates may have had more cognitive resources to bring to bear to the tracking task. If so, increases in commission and omission error rates should be positively related to tracking performance. To examine this question, tracking performance was correlated with omission, commission, and other event error rates for participants in the automated condition. Interestingly, higher rates of omission errors were negatively, not positively, associated with tracking performance; as omission errors increased, tracking performance decreased ($r = -0.21$, $p < 0.05$). Commission error rates were virtually unrelated to tracking performance ($r = -0.01$, *ns*). Interestingly, however, errors on non-error events were strongly and positively associated with tracking performance in the automated condition, $r = 0.45$, $p < 0.01$, but were completely unassociated with tracking performance in the non-automated condition, $r = 0.02$, *ns*. In sum, there is some support for the idea that having an AMA frees resources for other tasks (increased performance on non-error events), but there is no specific evidence to indicate that more complete diffusion of responsibility to the AMA for monitoring tasks (measured indirectly through omission and commission error rates) leads to increased performance on non-monitoring tasks.

3.5 PERCEIVED WORKLOAD

Although automated aids are generally assumed to reduce workload, there was no support for automation providing any subjective reduction on perceived workload relative to the non-automated condition. Specifically, participants perceived the experiment to be equally difficult, and perceived that they put forth equal amounts of effort ($F < 1$), regardless of condition.

3.6 EXPLORATORY ANALYSES

To explore possible explanations for why people make errors of omission and commission in automated contexts, omission and commission error rates were correlated with perceived task difficulty, effort, believing that automation improves accuracy of responses, and diffusion of responsibility. Only the automated condition sample was used for these analyses. As can be seen in Table 1, analysis of the perceived percentage of events that the AMA missed or gave an inappropriate directive were negatively correlated with both omission and commission error rates. This means that participants who

TABLE 1
Correlates of omission and commission errors in the automated condition ($N = 40$)

Variable	Omission errors	Commission errors
Perceived difficulty	– 0.17	– 0.37*
Perceived effort	0.22	0.05
Belief AMA lead to high accuracy	0.44**	0.59**
Diffusion of responsibility to AMA	0.44**	0.66**
Estimated % of events AMA missed	– 0.32	– 0.28
Estimated % of events the AMA gave an inappropriate directive	– 0.35*	– 0.44*

*$p < 0.05$; **$p < 0.01$.

felt that the automation made few if any errors were more likely to make errors of omission and commission. In a similar vein, participants who rated the AMA as highly accurate and who diffused responsibility to the AMA, made more errors of omission and commission that those who did not. Finally, people who perceived that the task was easy were more likely to make commission errors than those who perceived the part-task to be difficult. Perceived effort, however, was unrelated to error rates.

In addition, we explored whether the tendency to make either omission or commission errors was more likely in early relative to late trials. It is possible that people may make omission and commission errors because they are less comfortable with the experimental task, but as their comfort with the task increases their performance improves. Similarly, it may be that it takes several trials for participants to gauge the actual reliability of the AMA, which in turn could lead them to rely more on the gauges as trials progress. It will be recalled that there were eight trials that were presented to participants in one of 10 different orders. Only one trial was presented to participants in each possible position; this trial included one omission error opportunity, and one commission error opportunity. To test for order effects, omission and commission† error rates were compared as a function of order for this one trial. Results indicated that order had no significant impact on omission or commission error rates (for both analyses, $F < 1$). Similar analyses comparing whether the trial was presented in the first half or second half of the experiment yielded the same result.

To summarize, results indicated that strong belief in the accuracy and infallibility of the AMA, as well as delegation of responsibility to the AMA for system monitoring, led to higher rates of omission and commission errors. Omission and commission error rates did not vary as a function of order.

4. Discussion

Information integration in complex high-technology settings can be the most challenging aspect of work in these contexts. Automated aids and decision support tools

†Only the automated condition was examined to find out whether commission error rates varied as a function of order of presentation.

have become nearly indispensable in airplane cockpits, nuclear power plants, intensive care units, as well as numerous other settings. Moreover, automated decision aids are becoming increasingly common features of even less high-tech work settings as cars are equipped with increasingly sophisticated automation, people begin to rely on spell-checkers to edit their work and so on. Although the presence of automation in most work settings has many benefits—e.g. automated devices can generally process more information, and process that information more efficiently, than can human operators—it remains important to carefully evaluate the consequences of introducing automated aids into complex or even not so complex decision-making environments.

The present study compared decision making in the same context with and without an automated monitoring aid (AMA). Traditional system indices (e.g. gauges) were described as 100% reliable, but the AMA was described as highly but not perfectly reliable. The AMA failed to prompt participants about a system event 6 times (omission error opportunities), and incorrectly prompted a response on another six events (commission error opportunities) across 100 events. Performance across the 88 non-error events indicated that people performed better with an AMA than without one. In short, when the AMA worked properly, people in the automated condition made fewer errors than those in the non-automated condition.

However, when the AMA did not work properly, and despite the presence of a 100% reliable alternative system (e.g. gauges), much higher error rates were observed in the automated than the non-automated condition. Participants in the non-automated condition responded with 97% accuracy on the six omission error events, whereas participants in the automated condition responded with only a 59% accuracy rate on these same events. People with an AMA were therefore more likely to miss events than those without an AMA, if the AMA failed to notify them of the event. Participants in the automated condition were even more likely to make errors of commission, with an accuracy rate of only 35% across six possible commission error opportunities. The failure rate to detect events in automated settings when not prompted by an AMA was consistent with other research using both student and professional samples (e.g. Mosier *et al.*, 1998; Parasuraman, Molloy & Singh, 1993; Parasuraman, Mouloua & Molloy, 1994). Similarly, the commission error rate was consistent with rates observed with student and as well as samples of professional pilots (e.g. Mosier *et al.*, 1998; Mosier & Skitka, 1996).

In sum, when automated monitoring aids operated properly, their presence led to an increase in accuracy and a reduction in errors over not having an aid. However, when the automation yielded a "miss" (i.e. failed to detect an event), the presence of an automated aid led to an increase in errors on vigilance tasks relative to non-automated contexts.

Although these results illustrate both the benefits and possible limitations of automated monitoring aids, the extent to which these results will generalize to aviation or other complex real-life decision-making settings remains an open question. Although the W/Panes task (NASA, 1989) used in the present study was designed to provide a reasonable analogue to workload in avionic settings, W/Panes nonetheless cannot capture the true complexity of glass cockpit environments, or the psychological mindset of experienced pilots who have considerable experience dealing with complex multitasking in highly automated contexts. Existing research does suggest differences between professional and more novice samples. For example, when given a choice, pilots were found to rely on automation more than students did, especially during periods of automation failure (Riley, Lyall & Weiner, 1993; Riley, 1996). What we do not know is if pilots' preference for automation is based on experience that leads them to know that relying on reliable automation over the long haul will lead to fewer errors, or some other factor. Our own research using pilot samples in a mini-advanced concepts flight simulation at NASA Ames revealed omission and commission error rates that very closely parallel those observed in the present study and with other novice samples (Moiser *et al*., 1998). Presuming that these results are ultimately replicated across decision-making context and sample, the good news is that these results provide at least initial support to the assumption that automated monitoring devices will help reduce human error when they operate properly.

However, the results also point to the fact that automated conditions lead to more errors when aids do not explicitly prompt the operator about a system event, or when the aid provides a recommendation that is contradicted by other reliable indices. To what extent should we be concerned about automation bias as a problem, given that automation is usually designed to very stringent reliability standards and indications that people make fewer errors in automated than non-automated settings? The answer to this question depends on the cost associated with errors within a given decision-making domain. The cost of a 2% margin of error may be small on a widget manufacturing line (smaller profit margin), but may be unacceptably high at a nuclear power plant or an intensive care unit.

Many tasks are automated to avoid catastrophic accidents. In cases like these, error—any error—can be extremely costly in terms of human life or the environment. Although using highly reliable automated monitoring aids should decrease the probability of making catastropic errors considerably, there is likely to be high motivation to correct for automation bias in settings where errors can have disastrous consequences.

Moreover, it is important to note that commission errors can also occur even when automation is working perfectly (see also research on complacency effects, e.g. Parasuraman *et al*., 1993). For example, a China Airlines B747-SP, flying at 41 000 ft, lost power in its #4 engine. The autopilot, which was set for pitch guidance and altitude hold, attempted to correct for the loss by holding the left wing down, masking the approaching loss of control of the airplane. The crew did not notice the problem with the #4 engine, which would have been apparent on traditional display indices, and therefore took no action to deal with it. When the pilot disengaged the autopilot, the airplane rolled to the right, yawed and entered a steep descent. Extensive damage occurred during the descent and recovery (NYSB Report AAR-86-03, in Billings, 1996). The autopilot was operating just as it was designed to do, but because the crew failed to monitor the system after delegating control of the aircraft to the autopilot, the crew failed to recognize that the safety of the aircraft was seriously compromised.

The goal in settings where errors can lead to catastrophic consequences will not be to avoid using automated decision aids because they may introduce a new class of errors, but instead to use automated monitoring devices where they seem to be indicated, and then to take steps to try to minimize the probability of errors related to automation bias. Discovering what may help to reduce these effects without eroding the benefits of automated monitoring and decision aids is clearly an important area for future research. Research on adaptive automation (e.g. Carmody & Gluckman, 1993; Parasuraman,

1993) seems especially promising as a means to reduce commission errors and complacency effects, but whether it could effect the tendency to make commission errors remains an unexplored question. Adaptive automation takes over tasks only when operators are overloaded, but gives them back to the human operator during periods of lower workload. This approach represents a creative intervention that maximizes the strengths of automated aids—they can assist people during periods of high workload and free cognitive resources to be more focused on tasks that may not be as appropriate for automated control—but that also takes into account the potential costs of highly automated systems in terms of not only errors of omission and commission, but also potential skill decay.

Other research hints that explicit training about automation bias has at least short-term ameliorative consequences, but explicit instructions to verify automated directives, display prompts to suggest verifying automated directives, or adding a second "crewmember" do not have any impact on omission and commission error rates in student samples (Skitka, Mosier, Burdick & Rosenblatt, in press) or a professional pilot sample (Mosier *et al.*, 1998). Future research should continue to explore intervention and training strategies to guard against automation bias. Given that there seems to be considerable individual variability in omission and commission error rates, exploration of measures of individual differences may also provide useful insight and eventual selection tools.

In sum, the present research was designed to begin a more explicit comparison of decision-making in automated and non-automated contexts. Although future research is needed to further validate these results, the present study indicated that automated aids enhanced performance in a multi-task environment when the aid provided accurate feedback. However, when the aid provided inaccurate feedback (e.g. it missed a system event), participants in the non-automated condition performed much better than participants in the automated condition on these same events. The extent to which this pattern of results represents a concern that needs to be addressed in real-world settings depends to a large degree on the reliability of automated aids in a given context (and the inversely related need for human oversight), and the cost of making an error in that setting.

References

ANALYSIS OF FLIGHT DATA AND CABIN VOICE TAPES FOR KOREAN AIRLINES FLIGHT KE-007. (1993). *Aviation and Space Technology*, **138,** 17.

BILLINGS, C. E. (1996). *Human-centered aircraft automation: principles and guidelines*. Tech, Mom. No. 110381. NASA Ames Research Centre, Moffett Field, CA.

BREWER, M. B. & CRANO, W. D. (1994). *Social Psychology*. NY: West Publishing Co.

CARMODY, M. A., & GLUCKMAN, J. P. (1993). Task specific effects of automation and automation failure on performance, workload, and situation awareness. In R. S. JENSEN & D. NEUMEISTER, Eds. *Proceedings of the 7th International Symposium of Aviation Psychology*, pp. 167–171. Columbus, OH: Department of Aviation, The Ohio State University.

DARLEY, J. M. & GROSS, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, **44,** 20–33.

DIEHL, A. (1991). The effectiveness of training programs for preventing air crew "error". *Proceedings of the 6th International Symposium on Aviation Psychology*, pp. 640–655. Columbus, OH: The Ohio State University.

FISKE, S. T. & TAYLOR, S. E. (1994). *Social Cognition*, 2nd. (edn). New York: McGraw-Hill.

GLICK, P., ZION, C. & NELSON, C. (1988). What mediates sex discrimination in hiring decisions? *Journal of Personality and Social Psychology*, **55,** 178–186.

HAMILTON, D. L. (1981). *Cognitive Processes in Stereotyping and intergroup Behaviour*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

HOFLING, C. K., BROTZMAN, E., DALRYMPLE, S., GRAVES, N. & PIERCE, C. M. (1966). An experimental study in nurse-physician relationships. *Journal of Nervous and Mental Disease*, **143,** 171–180.

KAHNEMAN, D., SLOVIC, P. & TVERSKY, A. (1982). *Judgment under Uncertainty*: *Heuristics and Biases*. New York: Cambridge University Press.

KARAU, S. J. & WILLIAMS, K. D. (1993). Social loafing: a meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, **65,** 681–706.

KAHNEMAN, D., SOLVIC, P. & TVERSKY, A. (1982). *Judgment under Uncertainty*: *Heuristics and Biases*, Cambridge: Cambridge University Press.

LATANÉ, B., WILLIAMS, K. & HARKINS, S. (1979). Many hands make light the work: the causes and consequences of social loafing. *Journal of Personality and Social Psychology*, **37,** 822–832.

LAYTON, C., SMITH, P. J. & MCCOY, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. *Human Factors*, **36,** 94–119.

MESHKATI, N. (1991). Human factors in large-scale technological systems' accidents: three Miles Island, Bhopal, and Chernobyl. *Industrial Crisis Quarterly*, **5,** 133–154.

MESHKATI, N. (1996). Organizational and safety factors in automated oil and gas pipeline systems. In R. PARASURAMAN & M. MOULOUA, Eds. *Automation and Human Performance*: *Theory and Applications*, pp. 427–448. Mahwah, NJ: Lawrence Erlbaum Associates.

MOSIER, K. L. & SKITKA, L. J. (1996). Human decision-makers and automated decision aids: Made for each other? In R. PARASURAMAN & M. MOULOUA, Eds. *Automation and Human Performance*: *Theory and Applications*, pp. 201–220. Mahwah, NJ: Lawrence Erlbaum Associates.

MOSIER, K. L., SKITKA, L. J., DUNBAR, M., BURDICK, M., MCDONNELL, L. & ROSENBLATT, B. (1998). Automation bias and errors: Are terms better than individuals? *Proceedings of the 42nd Annual Meeting of the Human Factors Society*, Chicago, IL.

MOSIER, K. L., SKITKA, L. J., HEERS, S. & BURDICK, M. (1998). Automation bias: decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, **8**(1), 47–63.

NASS, C., FOGG, B. J. & MOON, Y. (1996). Can computers be teammates? *International Journal of Human– Computer Studies*, **45,** 669–678.

PARASURAMAN, R. (1993). Effects of adaptive function allocation on human performance. In D. GARLAND & J. A. WISE, Eds. *Human Factors and Advanced Aviation Technologies*, pp. 147–158. Daytona Beach, FL: Emory-Riddle Aeronautical University Press.

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

PARASURAMAN, R., MOULOUA, M. & MOLLOY, R. (1994). Monitoring automation failures in human-machine systems. In M. MOULOUA & R. PARASURAMAN, Eds. *Human Performance in Automated Systems*: *Current Research and Trends*, pp. 45–49. Hillsdale, NJ: Lawrence Erlbaum Associates.

PARASURAMAN, R. & RILEY, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors*, **39**(2), 230–253.

PRATARELLI, M. E. & MCINTYRE, J. A. (1994). Effects of social loafing on word recognition. *Perceptual and Motor Skills*, **78**(2), 455–464.

PRYOR, T. A. (1983). The HELP system. *Journal of Medical Systems*, **7,** 87–101.

RANK, S. G. & JACOBSON, C. K. (1997). Hospital nurses' compliance with medication overdose orders: a failure to replicate. *Journal of Health and Social Behaviour*, **18,** 188–193.

RILEY, V. (1996). Operator reliance on automation: theory and data. In R. PARASURAMAN & M. MOULOUA, Eds. *Automation and Human Performance*: *Theory and Applications*, pp. 19–36. Mahwah, NJ: Lawrence Erlbaum Associates. Inc.

RILEY, V., LYALL, E. & WEINER, E. (1993). *Analytic methods for flight-deck automation design and evaluation, phase two report*: *pilot use of automation*. FAA contractor Report, Honeywell Technology Centre, Minneapolis, MN.

SARTER, N. R. & WOODS, D. D. (1993). *Cognitive engineering in aerospace application*: *Pilot interaction with cockpit automation*, NASA Contractor Report 177617. NASA Ames Research Centre, Moffett Field, CA.

SKITKA, L. J., MOSIER, K. L., BURDICK, M. & ROSENBLATT, B. (1998). *Automation bias and errors*: *are teams better than individuals*? in press, International Journal of Aviation Psychology.

WEINER, E. (1989). *Human factors of advanced technology* ("*glass cockpit*") *aircraft*. NASA Contractor Report #177528, Moffett Field: CA.

WICKENES, C. D. & FLACH, J. M. (1988). Information processing. In E. L. WEINER & D. C. NAGEL, Eds. *Human Factors in Aviation*, pp. 111–156. San Diego, CA: Academic Press.

Paper accepted for publication by Associate Editor, Prof A. Sutcliffe