Automation:
Decision Aid or Decision Maker?

Linda J. Skitka, Ph. D.

University of Illinois at Chicago

Final Report

NASA Ames Cooperative Research Agreement NCC 2-986

Technical Monitor: Mary Connors

Department of Psychology, MC 285
1007 W. Harrison St
Chicago, IL 60607-7137

E mail Lskitka@uic.edu
Phone (312) 996-4464
Fax (312) 413-4122

Table of Contents

List of Tables and Figures

Automation:

Decision Aid or Decision Maker?

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., Deihl, 1991), the aviation industry and federal aviation and safety agencies have successfully pushed to increasingly automate flight systems (Billings, 1991; Weiner, 1989). Flight management systems computers are assuming greater control of flight tasks, such as calculating fuel efficient paths, navigation, detecting system anomalies, in addition to flying the plane. Other fields as disparate as nuclear power plants and even medical diagnostics are similarly becoming more and more automated. Because automated aids are generally accurate, airplanes fly safely, medical diagnoses are made correctly, and power plants run more efficiently. However indiscriminate or inappropriate reliance on automation will result in errors, just as inappropriate use of other decision making heuristics can result in errors. In short, while designed to explicitly reduce human error, introduction of automated decision aids may have the unanticipated consequence of not really eliminating human errors, but instead, creating new types of, and different opportunities for, human errors.

For the last four years, we (my research group at the University of Illinois, in close collaboration with Kathleen Mosier, NASA Ames Research Center/San Jose State University Foundation and more recently of San Francisco State University) have been engaged in a programmatic effort to understand the phenomena we have been referring to as "automation bias." If automation biases human judgment, it may lead to at least two major types of human errors: Omission and commission errors. Omission errors happen when people fail to notice a problem because an automated aid fails to detect it. Commission errors happen when people follow an automated directive or warning, at the expense of verifying it against other available information, or in spite of contradicting indications from other available sources of information.

This document reports on five studies conducted to gain further insight into the dynamics of automation bias, and to further explore possible solutions. Study 1 investigates the question of whether omission errors are more than lack of vigilance effects that we could observe even in non-automated settings. To investigate this question, the presence or absence of an automated monitoring aid was manipulated, and performance on the same omission error events in the automated context were compared with performance in the non-automated context. If people miss more of these events in the automated than non-automated domains, then there is greater credence to the notion that omission errors are in fact something unique to automated contexts.

The second study included on this report investigated whether the presence of a second crew-member helped prevent automation bias effects. In addition, this study also experimented with training manipulations and whether participants were given a prompt to verify automated directives to explore whether these variables might be able to ameliorate automation bias.

The remaining studies were scenario studies that examined what happened when automated and non-automated information was made equally salient to participants. These studies allowed us to examine whether commission errors happen because people fail to seek out additional confirming or disconfirming non-automated information before acting (a "short-circuited analysis" explanation) or if people notice information from non-automated sources, but discount it in favor of information from automated sources. In addition to

exploring these important process questions, the scenario studies also investigated whether an "action bias," or preference to do something rather than nothing, might also help account for commission error effects.

Study 1:

Are People Less Vigilant In Automated Than Non-Automated Settings?

The first study examined the relative rate of omission errors across automated and non-automated contexts using a part task. It has been a working assumption of our research that automation can lead to vigilance decrements because operators are turning over decision making responsibility in many cases to the automation. To more directly examine whether this assumption is based on sound footing, the present study compared the rate of omission errors on the same events as a function of whether an automated monitoring aid was available or not.

Method

Participants

80 undergraduate students participated in partial fulfillment of course requirements.

Tasks

Participants' primary task was to complete 8 "flights" or trials using the Workload/PerformANcE Simulation software (W/Panes) developed by NASA Ames Research Center (1989). This program presents 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).

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 ownship 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 through out each of the 8 trials, and required the greatest consistent attention from the participant.

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 lay-out of the map and the location of the waypoints varied across the four trials.

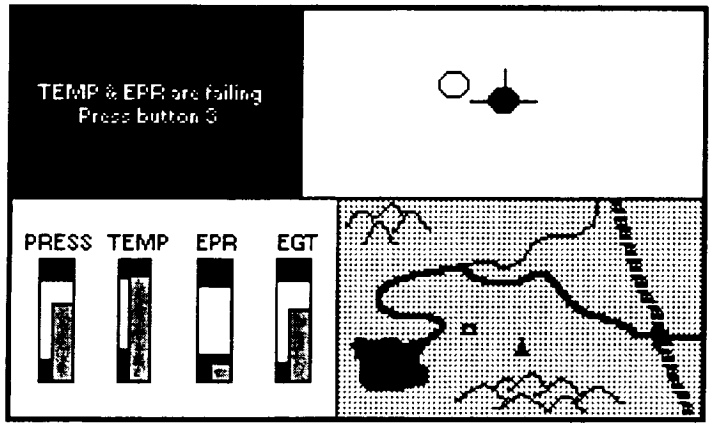


Figure 1. Primary task display.

Gauge Events. The lower-left quadrant displayed four analog slide-rule gauges that were used for a gauge monitoring task. Each gauge had upper and lower red. 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 to trained to push a "gauge reset" button. Gauges were programmed to return to a "safe" value after 10 seconds 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.

Automated Monitoring Aid. The top left quadrant was reserved for AMA announcements. The AMA notified participants of events, and recommended specific courses of action.

Overview

The critical manipulation of the present study involved 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 8 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.

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 five-minute full-task practice trials; performance feedback was provided at the end of each trial.

After a three-minute break, participants began the experimental trials. There were 8 experimental trials that lasted 8 minutes and 20 seconds each. Twelve orders of experimental trials were used to control for any possible event order effects. Across all 8 trials, there were 22 waypoint events, 22 three-gauge reset events, and 56 single-gauge reset events. Participants were given a five-minute 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.

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 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 gauge 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 6 opportunities to make commission errors. Commission error opportunities were characterized by an AMA directive that was contra-indicated by the other system indices. Because participants were aware that the other system indices were 100% accurate, and 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 8 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.

Other measures included a non-error performance measure, which was operationalized as the total number of correct responses for the 88 non-error events; an overall performance measure collapsing across error and 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 4-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 taking their average (after reversing the comfort item to ensure that they all moved in the same direction) and had a Cronbach's alpha of .85. Effort was measured by 3-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 gauge events faster on every trial." These items were scaled by taking their average, and had a Cronbach's alpha of .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 help me respond more accurately to gauge events," Cronbach's alpha = .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 gauge events," and "The AMA allowed me to off-load responsibility for the waypoint task," and had a Cronbach's alpha of .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.

Results

Manipulation Checks.

All of the participants in the automated condition correctly recalled that the gauges and other system indices were 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 since 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.

Do people make more errors of omission in 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 they were very reliable? The critical comparison was examination of 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 six events ($M = 2.44$) than those in the non-automated condition ($M = 0.18$), $F(1, 62) = 44.32$, $p < .05$, $MSE = 2.28$. In short, strong support was found for the hypothesis that people were less vigilant in environments with automated decision or monitoring aids, relative to those that do not.

In further examination of performance across the automated and non-automated conditions, the responses across all non-error opportunity trials 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, ($M = 71.85$), $F(1, 77) = 26.37$, $p < .05$, $MSE = 93.51$. In addition, overall performance on the 100 trials was assessed, opertationalized as the total number of correct responses. Participants in the automated condition performed better overall ($M = 88.67$) than participants in the non-automated condition ($M = 83.68$), $F(1, 77) = 4.99$, $p < .05$, $MSE = 98.61$. Interestingly, no significant differences emerged on tracking performance as a function of automation condition, $F(1, 77) = 1.89$, ns, $MSE = 12{,}991.01$. Therefore, having an AMA did not free participants' cognitive resources sufficiently to facilitate tracking performance.

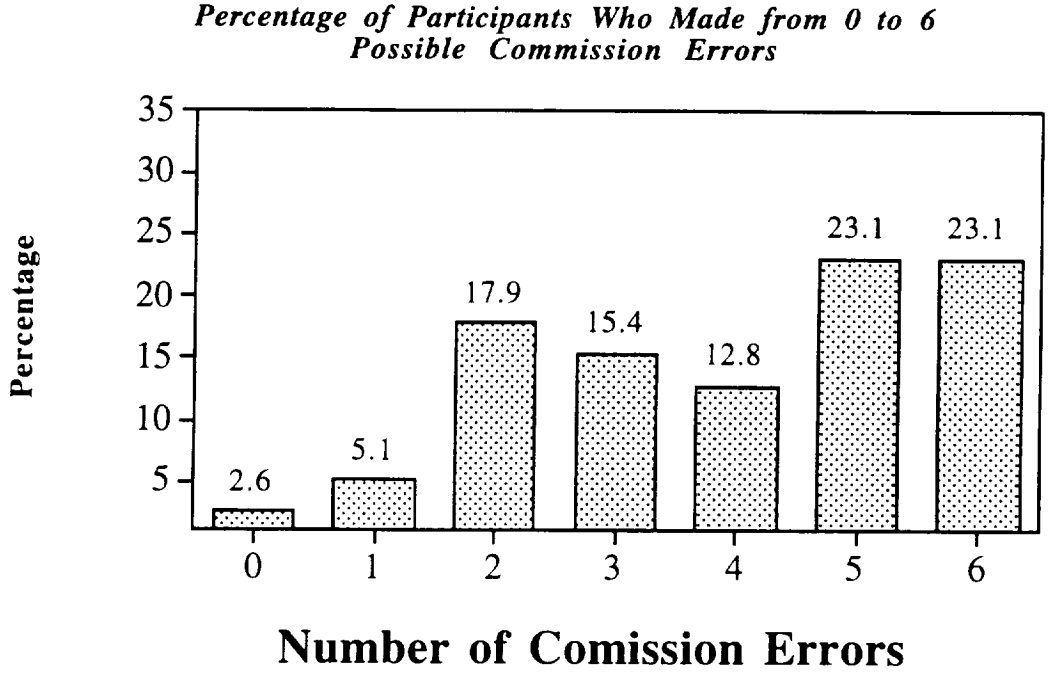
Although automated aids are generally assumed to reduce work load, there was no support for automation providing any subjective reduction on perceived work load relative to the non-automated condition. Specifically, participants perceived the experiment to be equally difficult, and perceived that they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

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 contra-indicated 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 6 error events. Only one participant made no commission errors; 23.1% of the participants made commission errors on all 6 events (see Figure 2 for more details). 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.

Figure 2.



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. As can be seen in Table 1, higher levels of omission errors were associated with the belief that following the AMA would lead to making more accurate responses, and the notion that the AMA was there to assume responsibility for monitoring the waypoint and gauge tasks. Interestingly, the higher the perceived system reliability, measured by the percentage of occasions participants estimated that the AMA either missed an event or failed to give an appropriate directive, was associated with lower rates omission and commission errors. Higher rates of commission errors were associated with the beliefs that the task was not very difficult, that following AMA directives would lead to more accurate responses, and that the AMA was assuming responsibility for these tasks.

In sum, these correlations paint an interesting and somewhat contradictory picture. Participants who felt that the AMA was the most reliable, were the least likely to make errors of omission and commission. Higher rates of error were also highly correlated the belief that the AMA helped these same participants to respond more accurately, and that the presence of the AMA allowed them to off-load responsibility for the waypoint and gauge events.

Table 1.

Correlates of Omission and Commission Errors in the Automated Condition (N = 39).

Variables	*Omission Errors*	*Commission Errors*
Perceived difficulty	-.17	-.37*
Perceived effort	.22	.05
Belief AMA leads to higher accuracy	.44**	.59**
Diffusion of responsibility to AMA	.44**	.66**
Estimated % of events the AMA missed	-.32	-.28
Estimated % of events the AMA gave an inappropriate directive	-.35*	-.44**

* $p < .05$; ** $p < .01$.

Discussion

This study clarified that automation bias is something unique to automated decision making contexts, and is not the result of a general tendency toward complacency. By comparing performance on exactly the same events on the same tasks with and without an automated decision aid, we were able to determine that at least the omission error part of automation bias is due to the unique context created by having an automated decision aid, and is not a phenomena that would occur even if people were not in an automated context. However, this study also revealed that having an automated decision aid did lead to modestly improved performance across all non-error events. Participants in the non-automated condition responded with 83.68% accuracy, whereas participants in the automated condition responded with 88.67% accuracy, across all events. Automated decision aids clearly led to better overall performance when they were accurate. People performed almost exactly at the level of reliability as the automation (which across events was 88% reliable). However, also clear, is that the presence of less than 100% accurate automated decision aids creates a context in which new kinds of errors in decision making can occur. Participants in the non-automated condition responded with 97% accuracy on the six "error" events, whereas participants in the automated condition had only a 65% accuracy rate when confronted with those same six events. In short, the presence of an AMA can lead to vigilance decrements that can lead to errors in decision making.

Study 2: Automation Bias in One- versus Two- Person Crews

A considerable amount of recent research has identified two classes of errors, or automation bias, that commonly emerge in highly automated decision environments: (1) Omission errors, defined as the failure to respond to a system event because an automated monitoring device fails to detect it, and (2) Commission errors, defined as when people follow an automated directive or recommendation, at the expense of verifying it against other available information, or in spite of contra-indications from other sources of information (e.g., Mosier, Skitka, Heers & Burdick, 1997, 1998; Mosier, Skitka & Korte, 1994; Skitka, Mosier & Burdick, 1996). Virtually all of the research to date, however, has examined the emergence of automation bias in the context of a single decision-maker working in isolation.

In most highly automated work settings, more than one person is available to assist with system monitoring. An open question has been whether automation bias occurs only when people are working in isolation, or also occurs when there are co-workers present.

At first glance, it would seem likely that omission and commission errors would be less likely in a two- than one-person crew. When two people are monitoring system events, it would seem to double the chances that they would detect a system anomaly, even if it were not detected by an automated decision aid. Moreover, doubling the number of people would also seem to enhance the probability that at least one of them will notice if the automated decision aid gives a recommendation that is inconsistent with other system indices.

Other research suggests that even if the second person does not increase the odds of detecting more events, that his or her mere presence may have an impact on a given decision maker's behavior. Social facilitation is defined as improvements in performance produced by the mere presence of others, whether these others are an audience or co-actors (Allport, 1920; Triplett, 1898). For example, Allport (1920) asked participants to write down as many word associations as they could think of for different words. Using the same participants in an alone versus with-others condition, he found that 93% of the participants could generate more alternative meanings in the presence of others. Similar effects emerged with animal studies involving feeding behavior and mazes; animals performed these tasks faster in the presence of other animals than when alone (Chen, 1937; Gates & Allee, 1933; Ross & Ross, 1949). Although most research reveals facilitating effects for the presence of others on measures of performance, some research has revealed that the presence of others can also lead to decreased performance, especially on unfamiliar or complex tasks (e.g., Pessin, 1933).

Zajonc offered a solution to this seeming inconsistency with the Drive Theory of Social Facilitation (Zajonc, 1965). According to drive theory, the presence of others has a nondirectional effect on people's behavior. The nondirectional component implies that the presence of others does not influence what type of behavior people engage in (e.g., performance enhancing or debilitating behavior), only that people's motivation, or drive, to behave in a particular way will be enhanced in the presence of others. Situational cues direct what people do, and social facilitation intensifies this response. Social facilitation may therefore lead to an increase or decrease in performance, depending on what the dominant response is in that social context. Research has been generally supportive of drive theory predictions. Individuals are more likely to emit dominant responses in the presence of others than when alone, and performance is either enhanced or impaired depending on the match of the dominant response to the performance being measured (see Geen, 1989; Geen & Gange, 1977 for reviews).

Assuming that attending to system states and responding correctly to system events is a dominant response in the kinds of decision making domains we have been studying, we should expect that the presence of a second crew member should increase people's motivation to maintain vigilance.

However, there are several contraindications to the notion that increased numbers will lead to increased performance. The presence of other people is often found to be distracting to people's performance (Sanders & Baron, 1975). In addition, there is considerable evidence that increasing the number of people responsible for a task leads to social loafing, or the tendency of individuals to exert less effort when participating as a member of a group than when alone (Ingham, Levinger, Graves, & Peckham, 1974; Williams, Harkins, & Latané, 1981; see Karau & Williams, 1993; Karau, Williams & Kipling, 1995 for reviews). Most directly relevant to the goals of the present study was research that indicated that social loafing is not restricted to simple motor tasks, but also applies to cognitive tasks. Harkins and Szymanski (1989) found that people working in three-person groups generated only 75% as many uses for a common object as they did when working alone. These same groups also made more than twice as many errors on a vigilance task of detecting brief flashes on a computer screen than they did when working alone.

The present study was designed to investigate whether the presence of another person either reduced, increased, or had no impact on, the tendency of people to make errors of omission and commission in automated contexts. Although a tandem study investigating this issue was conducted at NASA Ames in conjunction with this experiment, the NASA Ames simulation used only two-person crews, without equal one-person crew control group. Therefore, the primary purpose of the present study was to explicitly compare how performance compared across one- and two- person crews

In addition to exploring the impact of number of crew members, the present study also investigated the efficacy of a variety of possible interventions. Specifically, three training conditions were included in the present design. It was thought that commission errors could be reduced simply by training participants to verify automated directives against other sources of system information. In our previous studies, participants had been trained that they could verify automated directives, but the instructions did not suggest that they must verify automated directives. Therefore, in addition to a control condition using our usual training instructions, another condition trained participants that they must verify automated directives. Finally, it remained an open question whether training participants specifically about the tendency of people to make errors of commission and omission might be necessary to guard against their occurrence. A third training condition was included that explicitly informed participants about these errors, and how they could be prevented (e.g., through monitoring system states even when not prompted to do so by the automation, and by verifying automated directives).

In addition to exploring these variables, we also explored whether a system display change might help guard against especially commission errors. Half the participants received the display prompt "Verify" with each automated directive; the other half of the participants did not.

Method

Participants

One-hundred-and-forty-four students received partial course credit for their participation in the study, yielding 48 two-person crews, and 48 one-person crews.

Overview

Participants' primary task was to complete 4 "flights" or trials using the Workload/PerformANcE Simulation software (W/Panes) developed by NASA Ames Research Center (1989). Participants performed these tasks under conditions that varied as a function of: (a) "crew", that is whether they worked alone or with another person (two levels), (b) one of three levels of training (training that emphasized they could verify automated directives; training that emphasized that they must verify automated directives; or training that included instruction about errors people tend to make in automated contexts, and how they can be avoided), and (c) whether participants received a prompt to verify automated directives each time they received a directive, or were not prompted to verify. In sum, the study represented a crew (2) by training (3) by prompt to verify (2) three-way between subjects experimental design. The dependent variables of interest were the number of omission and commission errors participants made across these conditions. An "Automated Monitoring Aid" or AMA, failed to detect and announce 6 events that required responses, creating 6 opportunities for participants to make omission errors (i.e., failing to detect an event if not explicitly prompted about it by the AMA). Similarly, the AMA gave an inappropriate directive 6 times (e.g., indicating that a gauge was in a red zone, when in fact it was not), providing 6 opportunities for commission errors (i.e., following an AMA directive even when other indices indicated that the directive was not appropriate).

Procedure.

Participants were recruited to participate either singly or in pairs. In all three training conditions, participants were instructed on the W/Panes tasks. For details on the W/Panes task and participant training, see the description for the automated condition in Study 1. Like Study 1, the AMA was described as highly, but not perfectly, reliable. Participants were told that the AMA was quite reliable, but was not always 100% accurate in detecting when they needed to respond to waypoints or to gauge events, and that the AMA might fail to notify them of critical events, or could give inappropriate action directives. It was emphasized that the participant's job was to perform correctly and accurately, and that the information presented to them by their gauges and maps would always 100% accurate.

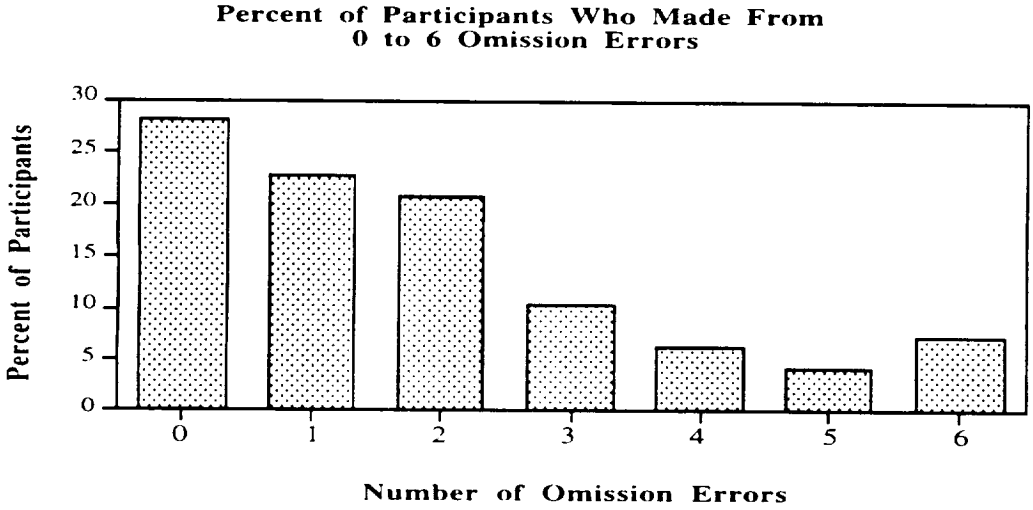
In the normal training condition, participants were told that they could check the appropriate indicators to verify the directives provided by the AMA. In the Verification training condition, participants were told they must verify all directives issued by the AMA by checking the various primary indicators (e.g., the map or the gauges). In the automation bias training condition, participants were told that there was a tendency in decision makers to over-rely on automated decision aids, when they have them, and that these kinds of decision aids can lead to two specific kinds of errors. Omission errors were explained as failures to respond to events, when not explicitly prompted about the event by the AMA. Commission errors were explained as resulting from failing to verify AMA directives, and the tendency to follow AMA recommendations even when the gauges or waypoints indicated that the AMA is not correct. Like the verification training condition, participants in the automation bias condition were instructed to verify all directives.

Participants were first verbally instructed and given a demonstration of how to use the W/Panes software and the response box and joystick. They were then given a 5 minute practice session, before beginning the experimental trials. When they began their experimental sessions, half the of participants received AMA directives that included a verification prompt. Specifically, the word "Verify" was added to any directives they received from the AMA. The other half of the participants received no prompt to verify the AMA directives. In the one-person crew, the "pilot" alone was responsible for all the W/Panes tasks. In two-person crews, one participant was randomly assigned to the pilot role, and the other participant was assigned the role of co-pilot. Pilots and co-pilots were both instructed on all tasks, and were told that they were both responsible for monitoring W/Panes system events. The pilot was responsible for actually doing the tracking task, and for making all button pushes in response to events. To provide a modest workload for the co-pilot in addition to his or her monitoring responsibilities, co-pilots were given a set of 75 2-digit multiplication problems, that they were told were being used to simulate the non-flight responsibilities of co-pilots (e.g., navigational and communication). In sum, the co-pilots were trained that their responsibilities were to assist the pilot with monitoring W/Panes for system events, and to also complete as many math problems as they could accurately do.

The four experimental trials were blocked into two sets to control for possible order effects. Half of the participants completed the AB block of trials first, and the CD block second. The other half of the participants completed the CD block first, and the AB block second.

Results

Figure 3.



Omission Errors

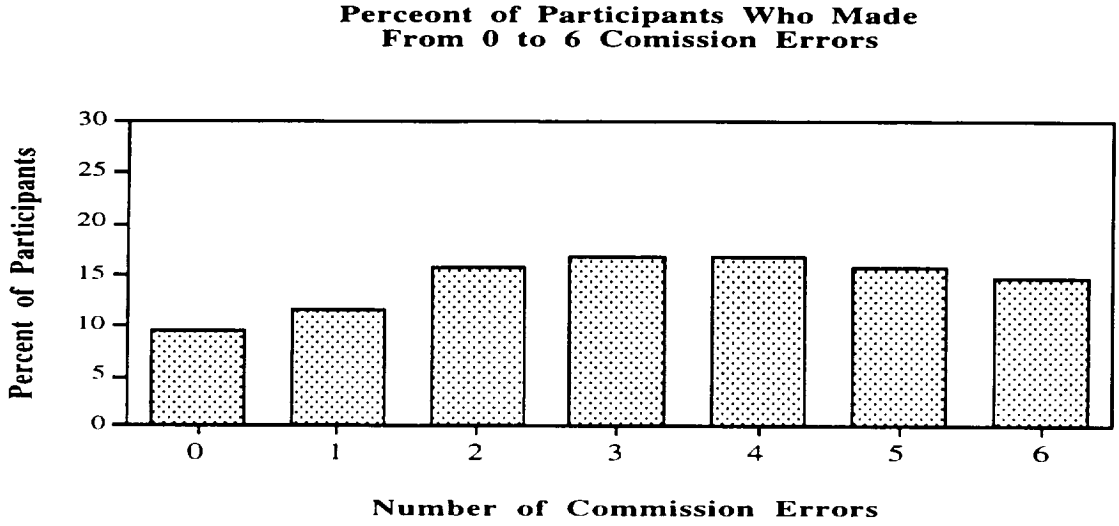
Fifty-one percent of the participants made one or more omission errors (see Figure 3). On average, participants made 1.85 omission errors out of a total of six possible errors (SD = 1.81), regardless of experimental condition. To what extent did having a second crew member, training, or prompts to verify, influence the number of omission errors participants made on the W/Panes task? A 2 (Crew) X 3 (Training) X 2 (Prompt to Verify)

X Trial Order (2) between subjects analysis of variance (ANOVA) revealed no significant effects of these variables on the number of omission errors.

Commission Errors

On average, participants made 3.25 commission errors out of a possible 6 (SD = 1.88). Although less than 30% of participants made 3 or more omission errors, almost 80% made two or more commission errors (see Figure 4).

Figure 4.



An examination of the number of commission errors as a function crew, prompts to verify, training, and trial order indicated that only training affected the number of commission errors participants made, F (2, 72) = 12.80, $p < .05$. Tukey tests indicated that the group that was explicitly trained about omission and commission errors made fewer commission errors (M = 2.59, SD = 1.72), than either the normal training group (M = 3.84, SD = 1.61), or the must verify group (M = 3.31, SD = 2.12). In sum, training about the problem of omission and commission errors helped reduce the number of commission errors made, but explicit training to verify automated directives, having a second crew member to help monitor system events, and being prompted to verify automated directives had no impact on the number of commission errors people made.

Discussion

The present study explored the extent to automation bias remains a pervasive problem even in the context of a two-person crew. Although previous studies have demonstrated, and repeatedly replicated, a base-rate tendency for people to miss more events in automated than non-automated settings, and to tend to follow automated directives even when they are inconsistent with other indices (even when those other indices are 100% reliable), there was some question whether these results would generalize to work settings where decision-makers had a human back up. According to the results of this study, as well as the mini-ACFS study conducted with pilots in a tandem design, automation bias remains a problem even in these contexts.

In addition to exploring the extent to which automation presented a problem even when people were not working alone, the present study also explored a number of other possible interventions that were predicted to be most effective in reducing the number of commission errors made. The present study indicated that training participants that they must verify automated directives did not create a reduction in the number of commission errors, nor did including a display prompt to verify directives. Increasing explicit awareness of the tendency of people to make omission and commission errors, however, did decrease the number of commission errors that people made, but had no effect on the rate of omission errors.

Study 3:

Exploring Competing Accounts for Commission Errors

To what extent can we establish that automation bias, is in fact, a bias? Several of our W/Panes studies have had conditions in which participants understood that the gauges, waypoints, etc. represented on screen were always 100% accurate, whereas the automated decision aid was not. Even in these contexts, we found ample evidence that people rely perhaps too heavily on the automation to detect and announce system events, and followed automated directives that were inconsistent with the information they had available on screen. Although these results suggest that some form of bias is operating, to what extent can we attribute these results solely to the automation? Results of the study described earlier in this report comparing vigilance and performance in automated and non-automated settings, found that people were more likely to miss more of the same events in automated than non-automated settings. This finding supports the notion that it is something about automation that leads people to be less vigilant.

Just how pervasive is this effect? A series of studies were designed to see if we would find commission error biases even when people were presented with overt information about the reliability of automated information relative to other system indices on a decision by decision basis. The impetus for this effort was other models of research that study biases in decision making (e.g., Kahneman & Tversky, 1982, as adapted by Epstein, Lipson, Holstein & Huh, 1992). These researchers presented people with the following scenarios:

> Mrs. Crane and Mrs. Tees were scheduled to leave the airport at the same time, but on different flights. Each of them drove the same distance to the airport, was caught in a traffic jam, and arrived at the airport 30 minutes after the scheduled departure of their flights. Mrs. Crane was told at her gate that her flight left on time. Mrs. Tees was told at her gate that her flight was delayed and had left just three minutes ago. They both had dawdled ten minutes before leaving home.
>
> Who do you think felt her dawdling was more foolish, Mrs. Crane or Mrs. Tees?
>
> Who do you think felt worse about missing her flight, Mrs. Crane or Mrs. Tees?

From a purely logical and normative point of view, Mrs. Crane and Mrs. Tees should feel equally badly about missing the plane-- both missed their flight with the exact same consequence. However, many people respond to this vignette in a biased fashion-- they

report that Mrs. Tees will feel worse than Mrs. Crane (an example of counter-factual reasoning).

Similar vignette evidence could be provided for automation bias if, given precise percentages outlining the reliability of both traditional and automated indices, people still go with automated directives when other, equally salient information, is more reliable. In each of three scenario studies, we manipulated automated system reliability (e.g., 99% , 80% and 50% reliable), traditional system indicator reliability (e.g., gauges that range in reliability, 99%, 80% and 50%), and potential cost or risk of making an error (high or low- - e.g., a wrong decision could lead to loss of life, versus the plane being delayed an hour). In other words, the core design represented a 3 X 3 X 2 within subjects experimental design that was replicated across three decision making scenarios: (a) an aircraft scenario; (b) an automobile scenario; and (c) a nuclear power plant scenario[1].

The scenario studies also allow us to explore the psychology underlying commission and omission errors more closely, and under more controlled circumstances. The W/Panes studies were highly controlled, but there remains several competing accounts for the psychological dynamics of automation bias. What is quite clear is that errors of omission involve vigilance decrements in automated settings. People turn over primary responsibility to the automation to monitor for events that require a response. However, commission errors are open to several alternative explanations. Specifically, what remains to be seen is whether people fail to check information available from other sources, or if they check it and discount it in favor of automated cues. Using a scenario approach to examine the psychology underlying commission errors allows us to make sure that people know what the cue is from other sources, and are even reminded of the relative reliability of the source of information. Under these circumstances, will we still see a tendency to follow automated directives, at the expense of other cues? If a systematic preference for automated cues emerges even in this setting, evidence will be provided for the discounting hypothesis. If no systematic preference for following automated cues emerges under conditions that hold the salience of automated and non-automated cues constant, results will be more consistent with a "short circuited analysis" account of why people make commission errors. The latter account would suggest that the psychological dynamics of both omission and commission errors are similar, that is, both at their root are vigilance problems. The former account would suggest that the psychological dynamic that produces commission errors is fundamentally different than that of omission errors, and that the presence of automated information biases how people interpret non-automated information.

In addition, the dynamics of automation bias (at least how it has been studied to date) make it difficult to determine whether commission errors in fact represent a preference for taking action, or doing something rather than nothing, rather than believing in the veracity of automated directives. Commission errors are defined as doing what the automation tells one to, even when the recommended action is inappropriate or other (more reliable) system indices contradict that information. In the research we have conducted to date, however, automated directives, consistently recommend taking an action, while the gauges generally imply no action, or a different action. In short, the "action orientation" of the automated monitoring aids may be what causes the commission errors, not the fact that the source is automated. To explore whether people prefer taking action to inaction under conditions of uncertainty, the "action orientation" of the automation and other system indices was strategically manipulated in the present set of studies. The first scenario study, the airplane scenario, action was held constant-- both the computer and gauges recommended action (just not the same action). The second and third scenario studies not only manipulated the

[1] The nuclear scenario compared 2 levels of reliability (90% versus 80%), rather than 3, after it was clear that reliability had very stable effects in the previous two versions of the study.

relative reliability of automated and gauge information, and made it equally salient, but also manipulated whether the automation, or the gauges, recommended action versus inaction . Are people equally likely to do what automated decision aids recommend, when that recommendation implies nothing should be done? How does that compare to reactions to gauge recommendations that vary in action orientation?

Study 1 used an airplane scenario that held action orientation constant. Modeling the scenario after the Mosier, Palmer and Degani (1992) incident used in full-flight simulation, the scenario involved a decision about which one of two engines was the most damaged, and needed to be shut down. Automated monitoring devices indicated that one engine was severely damaged, whereas traditional system indices suggested that the other engine was the source of the problem. In short, Study 1 investigates whether we will observe automation bias when action orientation is held constant, and participants are explicitly aware of the relative reliability of both the automated and traditional sources of information.

THE AIRPLANE SCENARIO STUDY

Method

Participants

Seventy students received partial course credit for their participation in the study.

Procedure

Participants were provided with the following background information before making judgments:

> Imagine you are a pilot of a common commercial aircraft. One aspect of your job is to monitor the aircraft for any potential problems, and to make decisions when irregularities arise. You have a variety of sources of information to assist you with decisions, including an on-board computer called a Flight Management System (FMS) that helps monitor all the systems of the aircraft such as the engines, the electrical system, fuel, etc. In addition to the computer, you also have more traditional sources of information to help you do your job and to make sure everything is operating properly. For example, you have gauges that indicate how fast you are flying, your fuel levels, status of your engines, etc.
>
> Given this background information, consider the following situation:
>
> You are flying along when all of a sudden you hear a big bang, the airplane starts handling with difficulty, and you can smell smoke.
>
> The Flight Management System indicates that there is a problem. On the computer screen, a message is flashing to indicate that the #1 engine is on fire, and the computer recommends turning the engine off.

You check your engine gauges. In contrast to the message from the computer, they indicate that your # 1 engine is normal, but that your #2 engine is failing. You have only two engines. Although it is important to turn off an engine that is on fire, turning off the wrong engine is very risky; you may not be able to make it to the nearest airport where you can land.

After going through this background information carefully with participants, they were given a check to ensure they fully understood what the computer and gauges were recommending be done in the above scenario (e.g., What is the computer recommending be done in this scenario? Turn off the #1 Engine, or Turn off the # 2 Engine?).

After passing these checks, participants then were presented with 18 choices that varied in terms of the stated reliability of the FMS, gauges, and the relative risk associated with making a mistake (e.g., whether they were close to an airport, or far from an airport). For example:

What if:

The FMS was known to be correct 50% of the time, the gauges were correct 80% of the time, and the nearest airport is quite far away?

What would you do?

_____ Turn off the #1 Engine _____ Turn off the # 2 Engine

The scenarios were based on a completely crossed design of the described reliability of the FMS (i.e., it was either 90%, 80% or 50% reliable), the reliability of the system gauges (i.e., they were either 90%, 80% or 50% reliable), and the risk associated with turning off the wrong engine (the nearest airport was quite close by, or quite far away). All variables were manipulated within subject. Participants' task was to make a decision to turn off either the #1 or #2 engine. These responses were scored +1 if participants chose to follow computerized recommendations, and scored -1 if they chose to follow gauge recommendations.

Finally, participants completed two sets of personality scales before being debriefed and thanked for their participation: The Revised NEO Personality Inventory (NEO PI-R, Costa & McCrae, 1992), and the Need for Cognitive Closure scale (Webster & Kruglanski, 1994).

The NEO PI-R was developed to operationalize the five-factor theory of personality, that proposes that all major personality traits and constructs can be organized as a function of five major trait constellations: Neuroticism (emotional stability or adjustment, versus emotional distress, negativism, disruptive emotion); Extroversion (as the name implies, extroversion versus introversion); Openness (openness to experiences; imaginativeness, attentiveness to inner feelings versus conventionalism, preference for the status quo and the familiar); Agreeableness (altruistic, inclined to say "yes" to others' ideas versus antagonistic and inclined to say "no"--more self-protective); and Conscientiousness (will to achieve, persistence, resistance to temptation versus laxadaisical, hedonistic, disorganized).

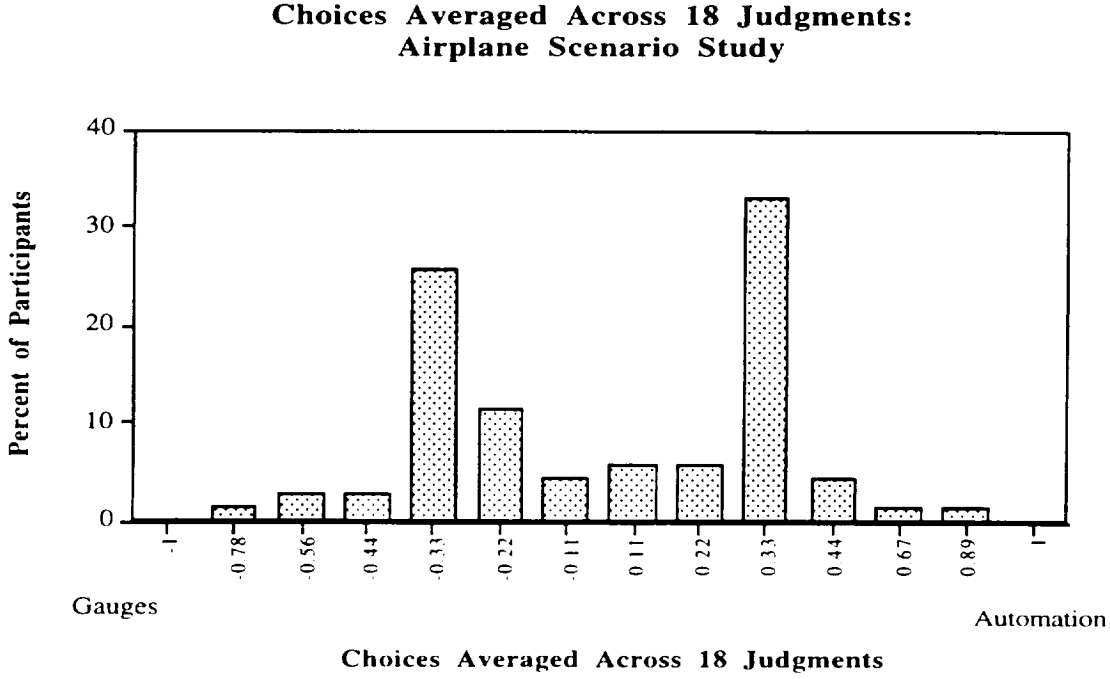
The Need for Cognitive Closure Scale, in addition to a total score, also has five factors including preference for order, preference for predictability, decisiveness, discomfort with ambiguity, and close-mindedness. Overall, the Need for Cognitive Closure Scale was

designed to tap into the need or desire for "*an* answer on a given topic, *any* answer,...compared to confusion or ambiguity on a topic" (Kruglanski, 1990, p. 337). In short, this measure is designed to tap people's proclivity to come to quick decisions versus those who are reluctant to come to a quick conclusion to decision making problems.

Results

The results are organized by first presenting descriptive information about the base rate tendency toward automation bias, possible personality correlates of a tendency to prefer automated information to gauge information, and then analysis of the role of reliability and risk in how people made their choices in this setting.

Figure 5.



Descriptive analyses.

If people are unaffected by the source of information, because all other variables are completely counterbalanced, we should observe a mean of zero across all choices (choices to follow gauge recommendations, scored -1, will cancel out choices to follow automated directives, scored +1, if there is no systematic bias toward either gauge or computerized recommendations). Indeed, the mean across all choices was $\underline{M} = 0.01$ ($\underline{SD} = .35$), which was statistically equivalent to zero. In short, on the aggregate there was no evidence to suggest that people are systematically biased in favor of automated information over gauge information. This finding suggests that the dynamics of commission errors observed in the W/Panes studies is not that people register non-automated information and discount it, but instead, that they fail to check other sources of information before deciding to act.

Despite the fact that on average people showed no overall tendency toward automation bias in this setting, an examination of the distribution revealed that some people do show systematic preferences for automated information, whereas others show a systematic

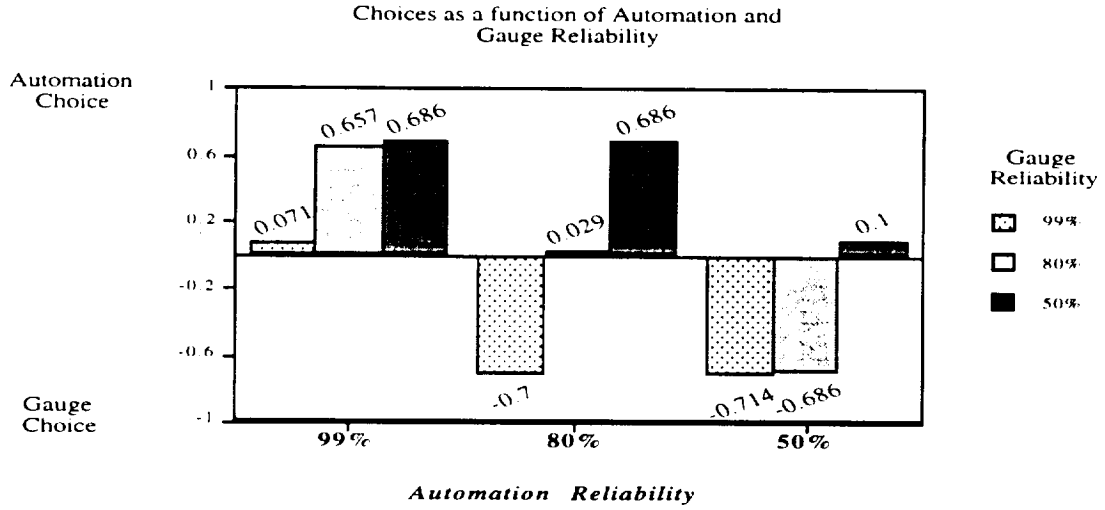
preference for gauge information. In fact, modal responses across judgments was not 0, but instead a bi-modal distribution emerged (-.33, and .33, see Figure 5).

The fact that there is such a high degree of variance in average responses, despite a balanced design, suggests that there may be important individual differences to consider in the extent to which people pay attention to automated versus traditional sources of information, even in contexts where the relative salience of information is held constant. Correlational analysis of the NEO personality constructs, and the need for closure subscale items revealed only one significant correlation: People who made choices more consistent with following automated directives on average were more closed-minded, r (70) = .26, p < .05.

More detailed analysis.

Analysis of the automation reliability (99%, 80% or 50% reliable) by gauge reliability (99%,80%, or 50% reliable) by risk (high or low) within subjects analysis of variance (ANOVA) on the dependent variable of choice (to shut down the #1 versus # 2 engine) revealed main effects for automation reliability, F (2, 138) = 87.79, MSE = 0.24, p < .001 and gauge reliability, F (2, 138) = 76.79, p < .001, MSE = .30, but no significant main effect for risk, F (1, 69) < .01. At the level of main effects, people's choices reflected the reliability of the information. Specifically, as the automation reliability increased from 50% to 80% to 90%, average responses became increasingly closer to +1, and as the gauge reliability increased from 50% to 80% to 90%, average responses became increasingly closer to -1. In addition to these main effects, the interaction of automation reliability and gauge reliability on choice was significant, F (4, 276) = 10.74, p < .001, MSE = 0.25.

Figure 6 .



The Automation by Gauge Reliability Interaction.

As can be seen in Figure 6, when participants were explicitly aware of the true reliability of automated versus non-automated recommendations, they responded in very rational

ways. Specifically, when the reliability of the automation and gauges were equal, participants were equally likely to chose to do what the automation or what the gauges recommended. Although the means in these cases were all positive (indicating a slight preference for automated recommendations, all other things being equal), they were not significantly different from zero. When the gauges and automation had different levels of reliability, the vast proportion of respondents made choices that used a "maximize reliability" rule, regardless of whether the source was automated or not, and also irrespective of the relative risk involved.

Discussion

Results of the airplane scenario study suggested when people were confronted with explicit information about the reliability of computerized versus gauge information, that they responded in very rational ways. Overall, there was not a general trend toward automation bias, but the distribution of results suggested that some people were systematically inclined to follow gauge recommendations, whereas others were systematically inclined to follow automated directives. Closed-mindedness was the only personality variable measured that was associated with the preference for gauge versus automated sources. People who were more close-minded preferred automated to gauge information.

More focused analysis examined how people responded when the reliability of automated and gauge information varied, and whether risk had any role in determining people's preference for following gauge versus automated directives. Results indicated that people responded in very rational ways in this decision making context. Under both high and low risk, people opted to go with the recommendation that had the greatest probability of being right. When both sources of information had an equal probability of being correct, results further revealed a lack of systematic bias, in that people were equally likely to follow either the computerized or gauge recommendations, rather than being systematically more likely to follow one source over another (i.e., means were effectively zero).

In sum, when gauge recommendations were presented on par, and with equal salience, to computerized recommendations, automation bias did not emerge. Instead, people responded as a function of the reliability of the source. These results were more consistent with a short-circuited analysis interpretation of commission errors, than a discounting hypothesis because people did not show any preference for following automation recommendations over gauge recommendations, when both sources of information were made equally salient. Support for the discounting hypothesis would require that people discounted, or devalued, the gauge information in favor of the information provided by the automation, even when both sources of information were made equally salient.

An important point to note, however, is that "action" was held constant in this study. Specifically, both the computer and gauges recommended taking some form of action. The choice for participants was which of the two competing actions they should choose? One competing hypothesis for why people have made commission errors in some of our other research (i.e., doing what an automated monitoring aid recommended, even when the recommendation conflicted with more valid and reliable system indices), is that the computer typically recommended taking some kind of <u>action</u>, whereas the traditional system indices always implied that the correct decision was <u>inaction</u>. Rather than representing a systematic automation bias, this interpretation of our prior results suggests that they may have been reflecting a systematic <u>bias toward taking action</u> under uncertainty. The results of the airplane scenario study indicated that when action is held constant, we do not observe decisions that were biased in favor of automated aids. Two additional scenario

studies were developed to explore whether we similarly find no evidence of bias when varying whether the automated aid, or the gauges, recommended action or inaction.

THE CAR SCENARIO STUDY

Method

Participants

Sixty-nine students, none of whom participated in other versions of the scenario study, received partial course credit for their participation in the present study.

Procedure

Participants were provided with the following background information before making judgments:

> Imagine you have just purchased the latest in luxury automobiles that came with an on-board computer, called an Auto Management System (AMS). The AMS helps to monitor all the systems on the automobile such as the oil pressure, engine temperature, fuel level, etc. In addition to the computer, you also have all the traditional sources of information to help you monitor the state of your new car. For example, you have a gauge for the oil pressure, the engine temperature, the fuel gauge, and so on.

Half of the participants (those in the Computer Action condition) then received the following instructions:

> Given this background information, consider the following situation:
>
> You are driving on a highway through the city, heading home late one evening, when the Auto Management System indicates that there is a problem. A computerized voice repeats that the engine is overheating because there is no oil pressure. To avoid permanent engine damage, the computer recommends that the automobile should be immediately turned off, and that you do not drive it again until you get it serviced.
>
> You check your engine gauges. In contrast to the message from the computer, they indicate that both the engine temperature and oil pressure are normal, and there are no problems with the engine.

The other half of the participants (those in the Computer Inaction condition) were provided with instructions that said that the gauges indicated the problem, but that the AMS did not. Specifically:

> You are driving on a highway through the city, heading home late one evening, when you check your engine gauges. They indicate that both the engine temperature is high, and that there is a problem with the oil pressure. Based on the owner's manual, you know that these are bad signs, and that the automobile should be

immediately turned off, and that you should not drive it again until you can get it serviced.

However, the Auto Management System indicates that there is no problem. Even pushing the button to have it check all engine functions reveals no problems with either the engine temperature or the oil pressure.

After going through this background information carefully with participants, they were given a check to ensure they fully understood what the computer and gauges were recommending be done in the above scenario (e.g., What is the computer recommending be done in this scenario? Pull over and stop the engine, or Continue driving home?).

After passing these checks, participants then were presented with 18 choices that varied in terms of the stated reliability of the AMS, the reliability of the gauges, and the relative risk (e.g, you see a gas station light at the next exit. However it is very late, and you are in what is known to be an unsafe part of town; or you are in a familiar and safe neighborhood). For example:

What if:

The AMS was known to be correct 50% of the time, the gauges were correct 80% of the time, and you are in what is known to be an unsafe part of town ?

What would you do?

_____ Pull over, and turn off the car _____ Keep driving home

The scenarios completely crossed the described reliability of the AMS (i.e., it was either 90%, 80% or 50% reliable), the reliability of the system gauges (i.e., they were either 90%, 80% or 50% reliable), and the risk associated with immediately pulling over (i.e., whether they were in a dangerous and unfamiliar neighborhood, or a safe and familiar neighborhood). These responses were scored +1 if participants chose to follow computerized recommendations, and scored -1 if they chose to follow gauge recommendations.

In other words, the experiment consisted of a 2 (Action orientation: Computer or gauges) X 3 (Computer reliability) X 3 (Gauge reliability) X 2 (Risk) mixed factorial design. Action varied between subjects, whereas the remaining variables represented within subject manipulations.

Finally, participants completed two sets of personality scales before being debriefed and thanked for their participation, the Revised NEO Personality Inventory (NEO PI-R, Costa & McCrae, 1992), and the Need for Cognitive Closure scale (Webster & Kruglanski, 1994) (see the description in the method section of the airplane scenario study for more detail about these scales).

Results

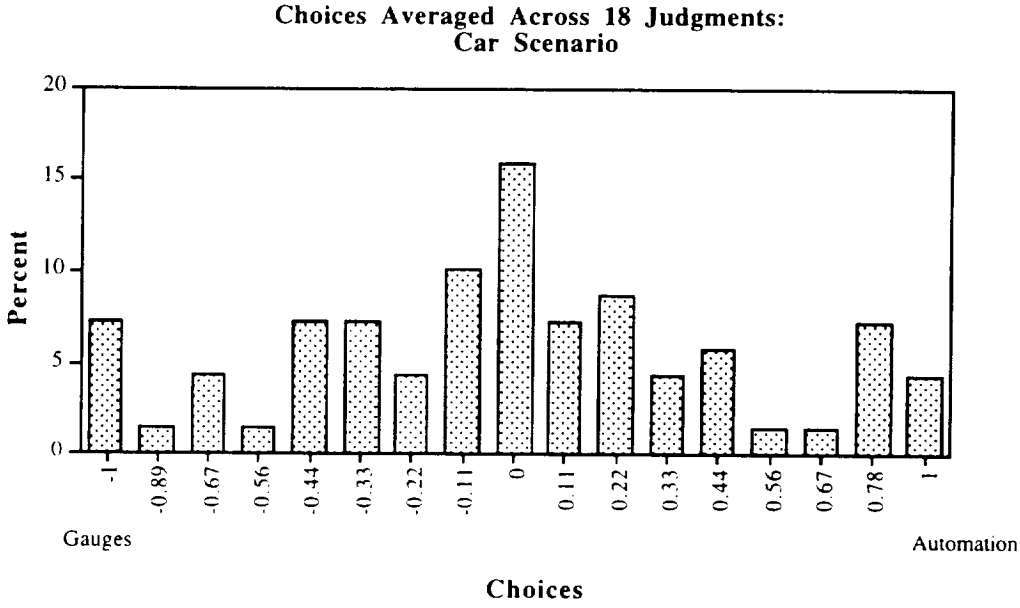
Again, results were organized first in terms of descriptive analysis, personality correlations, and then more detailed analysis about preferences under differing automation and gauge reliability, risk, and action orientation conditions.

Descriptive analyses.

If people are unaffected by the source of information in the car scenarios, because all other variables were completely counterbalanced, we should observe a mean of zero across all choices (where choices to follow gauge recommendations, scored -1, will cancel out choices to follow automated directives, scored +1). The mean across all choices in the car scenario was M = -0.01 (SD = .51) was statistically equivalent to zero. In short, on the aggregate there was no evidence to suggest that people were systematically biased in favor of automated information over gauge information in this scenario. This finding replicated that found in the airplane scenario study, and suggests that the dynamic of commission errors observed in the W/Panes studies is not that people register non-automated information and discount it, but instead, that they fail to check other sources of information before deciding act.

Examination of individual averages across the 18 trials revealed a much more normal distribution, and more people scoring at or around zero, than the distribution that was observed in the airplane scenario (see Figure 7).

Figure 7.



People more likely to make choices consistent with automated directives had greater intolerance of ambiguity than people more likely to make choices consistent with gauge directives, r (69) = .26, p < .05; no other correlations approached significance.

More detailed analysis.

Analysis of the action orientation (computer or gauges) X automation reliability (90%, 80%, or 50%) X gauge reliability (90%, 80%, or 50%) X risk (low, high) design revealed significant effects on choice in the car scenario. Significant main effects emerged for which source recommended action, F (1, 67) = 14.35, p < .001, MSE = 3.86, that

indicated that people preferred to follow gauge recommendations when gauges recommended taking action, and preferred automated recommendations when automation recommended taking action. Significant main effects also emerged for automation reliability, F (2, 134) = 42.33, $p < .001$, MSE = 0.75, and gauge reliability, F (2, 134) = 30.13, $p < .001$, MSE = 0.61. Participants were more likely to chose to follow more than less reliable automation, and more than less reliable gauges.

In addition to these main effects, the action by risk interaction was significant, F (1,67) = 13.76, $p < .001$, MSE = 3.33. Analysis of simple effects indicated that when participants were presented with indications of car trouble in a bad neighborhood, they exhibited no systematic preference for immediately pulling over, as compared to continuing to drive, F (1, 67) < 1, nor did the mean preferences differ significantly from zero (what would be expected on the basis of chance choices). However, an action bias did emerge when participants learned of possible engine problems in a safe neighborhood. When making choices under low risk (a safe neighborhood), participants were more likely to stop the car, regardless of whether the reliability of the source recommending doing so was less than the source suggesting nothing was wrong, F (1, 67) = 7.95, $p < .01$, MSE = 3.33 (see Table 2).

Table 2.

Choices under conditions that vary source of problem identification and risk associated with taking action in the car scenario.

Recommendations: **Do Something/Do Nothing**	Safe to take action (good neighborhood)	Less safe to take action (bad neighborhood)
Computer/Gauges	.31	-.05
Gauge/Computer	-.56	.09

Note: Positive averages reflected that more participants chose to do follow computer recommendations, whereas negative averages reflected that more participants chose to follow gauge recommendations. Means close to zero reflected no systematic preference for gauge versus computerized recommendations. Scores could range from +1 (always following computerized recommendations) to -1 (always following gauge recommendations).

Specifically, when the computerized monitoring aid recommended stopping and turning off the engine, but the gauges showed no indication of problems, people were likely to follow this recommendation if they were in a safe neighborhood. In contrast, when the gauges recommended taking action, and the computer revealed no problems under low risk, people were more likely to follow the gauge's implied action, than the computer's. In short, evidence emerged supporting the idea that people are more inclined to discount reliability information in order to take action, at least under some conditions (e.g., it is safe to do so).

In addition to finding an action by risk interaction, the source of recommended action (the computer or the gauges) also interacted with automation and gauge reliability to predict participants' choices, F (4, 268) = 3.20, $p < .05$, MSE = 0.38.

A careful examination of Table 3 reveals how the variable "action" (whether the computer or gauge recommended taking an action to address a problem with the car) qualified the automation by gauge interaction. Eight out of 9 comparisons across action supported the hypothesis that people prefer to do something rather than nothing, irrespective of the relative reliability of the source that recommended that action should be taken. The only cell that was unaffected by whether the computer or the gauges were inferring an action should be taken was the low computer/ high gauge reliability combination: People were equally inclined to follow gauge recommendations under this condition, regardless of whether the gauge was recommending to do something, or nothing.

Table 3.

Choices in the Car Scenario as a function of Whether the Computer Recommended Action or Inaction, Automation Reliability, and Gauge Reliability.

Computer/Gauge Reliability	Computer Identified a Problem Requiring Action	Gauges Identified a Problem Requiring Action
High/High	.319a	-.318b
High/Moderate	.383a	-.091b
High/Low	.596a	.182b
Moderate/High	-.128a	-.545b
Moderate/Moderate	.298a	-.273b
Moderate/Low	.447a	-.136b
Low/High	-.574a	-.591a
Low/Moderate	-.340a	-.727b
Low/Low	.170a	-.409b

Note: Positive means reflected that on average, participants chose to do what the computer recommended; negative means reflected that on average, participants chose to do what the gauges recommended. Means close to zero indicate no clear preference. Scores could range from +1 (always following computerized recommendations) to -1 (always following gauge recommendations). Underlined scores were not significantly different from zero, the average score expected by chance. Means with non-common subscripts across the action variable are significantly different at $p < .01$

Discussion

Results replicated the basic finding of the airplane scenario, that is, when both automated and non-automated information was presented in an equally salient manner, no preference for automated information over non-automated information emerged.

The major finding of the car scenario study was that action orientation and risk did influence the extent to which people paid attention to reliability information. A preference for action clearly had its strongest impact when the reliability of the automation and gauges were equal, but it nonetheless had a significant impact on 5 out of the 6 choices in which the automation and gauge reliability were different. In short, a certain proportion of people preferred taking action to inaction, even when taking action was a less rational choice from a reliability point of view (see Table 3).

The extent to which people preferred to take action was also moderated by the risk associated with that choice . Risk was manipulated in this study by having different levels of potential harm associated with taking an action, while holding constant the risk associated with doing nothing. For half the choices, participants were told that it was safe to act, that is, they were in a safe and familiar neighborhood. For the other half of their choices, it was riskier for participants to choose to act on problems identified by either their gauges or the automation, because they were in an unfamiliar and dangerous part of town. The cost of inaction was constant: There was some probability of engine damage if they did not stop. People were more inclined to stop the car (take action), even if the reliability of the information suggesting that they do so was low, if they were in a safe than bad neighborhood.

The nuclear power scenario was developed to conceptually replicate the action by risk context presented to participants in the car scenario. However in this case, risks varied as a function of taking inaction, while the risks associated with action were held constant.

NUCLEAR POWER SCENARIO

To conceptually replicate the findings that risk and action interact to determine when people choose to act even on the basis of less reliable information, the nuclear power scenario study was constructed so that there were risks associated with failing to act, rather than with acting. Specifically, for half the choices, participants were told that not shutting down the reactor if it was overheating would have the consequence of releasing radioactivity that could be harmful to the surrounding populace. For the other half of the choices, there were no health consequences of failing to shut down the reactor .

In the car scenario, people were <u>less</u> likely to act when they could avoid risks by inaction. Will people in the nuclear power scenario be <u>more</u> likely to act, because they can avoid risks by taking action?

Method

Participants

Sixty-one students participated in partial fulfillment of course credits. None of the nuclear power scenario participants were included in the samples used in other versions of the scenario studies.

Procedure

Participants were provided with the following background information before making judgments:

> Imagine you are a shift supervisor at a nuclear power plant. One part of your job is to monitor the reactor core for any potential problems, and to make decisions when any irregularities arise. You have a variety of resources to assist you in making decisions. The power plant has a computer system, called a Reactor Management System (RMS), that helps monitor the state of the plant such as the reactor core temperature, coolant levels and flow, etc. In addition to the computer, you also have more traditional sources of information to help you do your job and to make sure that everything is operating properly. For example, you have gauges that indicate the reactor core temperature, that monitor coolant levels and flow, etc.

Half of the participants (those in the Computer Action condition) then received the following instructions:

> Given this background information, consider the following situation:
>
> You are monitoring your station when the Reactor Management System sounds an alarm and a message starts flashing on your computer monitor. On the screen is a message that the reactor core is overheating due to a coolant flow failure. The computer recommends an emergency shut-down of the reactor.
>
> You check your reactor core temperature and coolant flow gauges. In contrast to the message from the computer, they indicate that there is nothing wrong with either the reactor core temperature or the coolant flow.
>
> An emergency shut-down of the reactor dumps thousands of gallons of water into the core, stops energy production, and can cut off electricity to thousands of residents for hours.

The other half of the participants (those in the Computer Inaction condition) were provided with instructions that said that the gauges indicated the problem, but that the RMS did not. Specifically:

> Given this background information, consider the following situation:

You are monitoring your station when you check your reactor core temperature and coolant gauges. They indicate that there is something wrong with the reactor core temperature and the coolant flow. Based on your training, you know that these are indications that the reactor core must be overheating because of a coolant flow failure. Under these conditions, you are trained to do an emergency shut-down of the reactor.

In contrast to the information from your gauges, however, the Reactor Management System has not sounded an alarm or recommended shutting down the system. A manual request for the computer to re-check all the reactor systems yields the computerized message that all systems are fine.

An emergency shut-down of the reactor dumps thousands of gallons of water into the core, stops energy production, and can cut off electricity to thousands of residents for hours.

After going through this background information carefully with participants, they were given a check to ensure they fully understood what the computer and gauges were recommending be done in the above scenario (e.g., What is the computer recommending be done in this scenario? An emergency shut-down of the reactor, or not to shut the reactor down?

After passing these checks, participants then were presented with 18 choices that varied in terms of the stated reliability of the RMS, gauges, and the relative risk of not taking action if the reactor core was in fact over heating (no health hazard, or could create a health hazard). For example:

What if:

The RMS was known to be correct 80% of the time, the gauges were correct 90% of the time, and if in fact the reactor core is overheating, there is a health hazard because radioactivity could be released to affect the surrounding populace?

What would you do?

_____ Shut down the reactor _____ Not shut down the reactor

The scenarios completely crossed the described reliability of the RMS (i.e., it was either 90% or 80% reliable), the reliability of the system gauges (i.e., they were either 90% or 80% reliable), and the risk associated with not taking action if indeed the reactor was overheating (health hazard, or no threat of a health hazard). Half the participants made these judgments under conditions where the computer was recommending that the reactor be shut down, whereas the other half of the participants made judgments under conditions where the gauges were recommending that the reactor be shut down. Responses were scored +1 if participants chose to follow computerized recommendations (either to shut down or not shut down, depending on condition), and scored -1 if they chose to follow gauge recommendations.

Finally, participants completed two sets of personality scales before being debriefed and thanked for their participation: The Revised NEO Personality Inventory (NEO PI-R, Costa & McCrae, 1992), and the Need for Cognitive Closure scale (Webster & Kruglanski, 1994; see the first scenario study method for more detail).

Results

Descriptive analyses

Similar to the airplane and car scenarios, if people are unaffected by the source of information, because all other variables are completely counterbalanced, we should observe a mean of zero across all choices (choices to follow gauge recommendations, scored -1, will cancel out choices to follow automated directives, scored +1). Indeed, the mean across all choices in the nuclear power scenario, like that airplane and car scenarios, was essentially zero ($M = 0.04$, $SD = .63$. Interestingly, however, the distribution was quite flat, with equal numbers of participants across the range from -1 to +1.

The only significant personality correlate of preferences for automation over gauge recommendations was that people who were more likely to follow automated directives were higher in a preference for order, $r\ (61) = .27$, $p < .05$.

More detailed analyses

Analysis of the automation reliability (2: High or moderate) by gauge reliability (2: High or moderate) by risk (high, low) by action source (automation or gauges) design yielded a significant main effort for action, $F\ (1, 59) = 30.87$, $p < .001$, MSE = 2.13, automation reliability , $F\ (1, 59) = 16.97$, $p < .001$, MSE = 0.72, and gauge reliability, $F\ (1, 59) = 13.01$, $p < .01$, MSE = 0.78. Participants were more likely to follow automation system recommendations when it suggested taking action ($M = .36$), and to follow the gauge recommendations when they suggested taking action ($M = -.38$). Similarly, holding other features constant, people were more likely to follow more reliable automation than less reliable automation, and more reliable gauges than less reliable gauges. The main effect for risk was not significant, $F < 1$.

In addition to these main effects, the results of the nuclear power scenario also revealed a significant action by risk interaction, $F\ (1, 59) = 6.24$, $p < .01$, MSE = 1.05 (see Table 4). When there were high risks associated with doing nothing (and being wrong), people were more likely to act, regardless of the reliability of the source recommending action, $F\ (1, 59) = 27.33$, $p < .001$, MSE = 1.05. A similar trend for action also emerged under low risk, but the effect was considerably smaller, $F\ (1, 59) = 7.05$, $p < .01$, MSE = 1.05.

No other interactive effects were significant.

Table 4.

Choices under conditions that vary source of problem identification and risk associated with taking action in the nuclear power scenario.

Recommendations: Do Something/Do Nothing	Less safe to do nothing if source detecting problem is right	Safe to do nothing if source detecting problem is right
[redacted]	.51	.20
[redacted]	-.46	-.31

Note: Positive averages reflected that more participants chose to follow computer recommendations, whereas negative averages reflected that more participants chose to follow gauge recommendations. Means close to zero reflected no systematic preference for gauge versus computerized recommendations. Scores could range from +1 (always following computerized recommendations) to -1 (always following gauge recommendations).

Discussion

The results of the nuclear power scenario study conceptually replicated the basic findings of the airplane and car scenario studies, in that no systematic preference for automated information was observed. In all three scenario studies, information from non-automated sources was presented on equivalent terms to information for automated sources; participants did not have to independently seek this information out. Under these conditions, automation bias effects do not emerge.

In addition to replicating the finding that people are equally inclined to base their decisions on automated versus non-automated sources of information, the nuclear power scenario study also conceptually replicated the action by risk interaction observed in the car scenario study. Risk did not influence decisions made in the airplane scenario study, when both sources of information recommended taking (competing) actions. However, a bias in favor of taking action emerged in both the car and nuclear power scenarios, when only one source recommended action, and the other source recommended doing nothing. In contrast to the car scenario study, the nuclear power scenario study varied the risk associated with doing nothing, and held the risk with doing something constant. When risk varied as a function of inaction, rather than action, people again acted to minimize the potential hazard associated with their choices. Although there was a preference for action under both conditions of manipulated risk, the effect was much stronger under high risk associated with inaction. Participants were most likely to prefer action (shutting down the reactor) when there was a higher risk associated with not taking action if the source detecting the problem was right. It is important to note that the risk by

action interaction effects are not qualified by the relative reliability of the source recommending action.

Overall Discussion of the Scenario Studies

The W/Panes studies left open the question of the psychological dynamics of automation bias, and particularly why people make commission errors. Specifically, do people fail to check other indices (e.g. gauges) before following automated directives, or do they check the gauges, but discount the validity of non-automated information? The scenario studies yield results that are much more consistent with a "short-circuited analysis" interpretation of the W/Panes results, rather than a discounting explanation. In the airplane scenario study, participants were presented with conflicting recommendations for taking action from an automated decision aid, and traditional system indices (e.g., gauges). Under these conditions, people made very rational choices; they chose to base their actions on the most reliable source of information, irrespective of the relative risk associated with failing to make the "right" choice. No evidence of automation bias emerged. Results indicated that overall, when the gauge information was made equally salient and overt information about the relative reliability of automated and non-automated information was available, that no systematic bias for preferring to follow automated directives emerged.

In addition to raising questions about whether people are inclined to make errors of commission, even when made explicitly aware of other valid sources of information, the W/Panes studies also hinted that there may be a tendency to prefer taking action to inaction under uncertainty. The car and nuclear power scenario studies examined this possibility by manipulating which source indicated a problem (and by implication, which source recommended an action be taken by the participant): either the computer aid, or the gauges. No evidence of automation bias emerged in either setting, but some findings supported the notion that people under some conditions are biased toward taking action over inaction when the costs of taking action are low, or the costs of inaction [even a low probability of the cost actually occurring] are high.

Overall, the scenario studies revealed that people's preference for action over inaction under conditions of uncertainty can bias judgment in contexts where they have multiple sources of information to inform choices. When action is held constant, as it was in the airplane scenario study (all sources of information recommended taking action, and the question was which action people would take), people responded in very rational ways. Regardless of the level of risk, people relied on the most reliable source of information to determine which action they took.

However, an action bias emerged when one source of information recommended doing something, and the other source of information recommended doing nothing, or maintaining the status quo. People in general preferred taking action to inaction in these settings. In the car scenario, personal risk associated with taking an action (pulling over the car) was also manipulated, while the risk associated with doing nothing (driving home) remained constant across choices. When taking an action involved low levels of personal risk, people's choices confirmed an action bias. That is, they preferred to follow the recommendation of the source (either the gauges or the computer) that was suggesting they should do something, rather than do nothing, if there was low risk associated with doing so (e.g., they were in a safe neighborhood). When there was higher personal risk associated with taking an action, the bias toward action disappeared, and people followed the recommendation of the most reliable source.

In the nuclear power scenario, inaction (not shutting down the plant) varied in whether there was an associated risk, while the risk associated with action (shutting down the plant)

was held constant. Specifically, not shutting down the reactor if indeed it was overheating was either associated with no health risks, or the risk of radioactivity being released to the surrounding community. The cost of shutting down the plant was constant across choices: It would cost the company money, and some people would be without power. The results conceptually replicated those observed in the car scenario study. People showed high evidence of an action bias when the potential costs of inaction were high. An action bias also emerged under lower risk, but the effect was weaker than under high risk. — they knew the risk, even though unstated.

In sum, at least in contexts where the salience of automated versus gauge information is held constant, there is little evidence of a systematic tendency to prefer automated information to gauge information. People prefer to follow the most reliable source, but are also particularly adverse to potential risks. When the consequence of not taking action is severe, people are more likely act than not act, even if the source recommending action is less reliable than the one that indicates no action is necessary. People are also more likely to act if the probability of personal risk associated with taking action is low, even if the source recommending action is less reliable than the one that indicates no action is necessary.

Results from the scenario studies taken together indicate that the form of information that people get from the person-machine interface can have dramatic consequences on their subsequent behavior. How people make decisions in human factors contexts is far less rationally based than many people may suppose. People may choose to take action, simply because there is little risk in doing so, even if inaction is a more appropriate response. More importantly, when the choice to not act is associated with salient and costly counterfactuals (e.g., even though all indices suggest that inaction is the right choice, if I am wrong, something horrible will happen), people are likely to follow low reliability information and act anyway. In short, action bias under uncertainty is a reality we need to consider in the design of human-machine systems. Future research should investigate whether the action bias results found in the scenario studies can be replicated in more externally valid contexts, for example in a W/Panes study, mini-ACFS study, or full flight simulation. Also, need to look at action slight increase rather than decrease risk.

Unlike "real" decision making contexts, the relative salience and availability of information, regardless of source, was carefully controlled in the three scenario studies. The fact that commission errors, or automation bias, did not emerge in a context where cues from other sources were placed in equal attentional competition, and that this finding was replicated across three different decision making settings, informs us that the psychological dynamics of automation bias can be characterized better by a short-circuited analysis account, than a discounting account, of the effect. When all the information is made easily available, participants did not discount or depreciate the value of information from non-automated sources. These findings suggest that in more traditional decision making contexts where non-automated cues may not be equal in salience to cues from automated sources, people are failing to seek out either confirming or disconfirming evidence before making a choice to act. Overall, these results suggest that the psychological dynamics leading to omission and commission errors are more similar than dissimilar-- at the core of the problem is decreased situational awareness and vigilance in automated as compared to non-automated, decision making contexts. — or perhaps the automation is more reliable.

Interestingly, no replicable findings of individual differences emerged in the tendency to prefer to follow automated directives across scenarios. However there are clear individual differences in a systematic preference for either automated or traditional sources of information, even when the reliability of these sources is made equally salient and available. Given that automation bias effects were minimized in the present set of studies because the relative salience of automated and non-automated cues was controlled, the

ability to also detect correlates of a tendency toward automation bias was probably also quite low. Future research should continue to attempt to identify whether there are identifiable sources of these differential preferences.

Overall Conclusions

Information integration in complex high technology settings can be the most challenging aspect of work in those settings. Automated aids and decision support tools have become nearly indispensable in airplane cockpits, nuclear power plants, intensive care units, and so on. Although explicitly designed to guard against potential human error, the presence of automated monitoring devices and decision aids have important psychological consequences for how people perform their tasks and make decisions. 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 much more efficiently, than can human operators-- the possible costs associated with the presence of automated decisions aids has to be weighed in and understood if we are to use automation wisely. 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., Billings, 1996; Sarter & Woods, 1993). The research presented here also provides compelling evidence that automated decision aids can bias how people use, but not necessarily how they process, other kinds of information.

Our results have indicated that people tend to delegate primary responsibility for system monitoring to automated aids, and routinely fail to notice events that the aid does not detect. These results occur even when participants are made explicitly aware that the automation is not 100% reliable, and when participants have been trained that they are primarily responsible for making all decisions.

In addition to failing to detect events that they are not explicitly prompted about by an automated aid (omission errors), people also tend to follow automated directives that are either patently wrong, or that are inconsistent with other system indices. For example, although participants might be well-trained that the correct response to a three-gauge reset event is to push the reset button (and they have successfully performed this task on numerous trials), they are prone to follow an automated directive to push a different button instead (a commission error). Given that automated decision aids are designed to be highly reliable, how big of a problem is automation bias? Similarly, how big of a problem is the tendency of people to prefer to take action, compounded by the fact that people tend to follow automated directives (that usually recommend some form of action)?

Even when automated devices are functioning perfectly, they can not be programmed for every possible contingency. Automated aids can not take into account context or all aspects of the situation that might render their programmed response inappropriate. For example, the AMA directive to pull over the car to check out a potential engine problem does not take into account whether one is in a safe or bad neighborhood before making a recommendation. Automated decision aids will never be 100% reliable and valid mostly because they can not be programmed for every possible contingency. Therefore, we can be confident that there will be a certain base rate of errors made due to the fact that people tend to assign responsibility for system monitoring to their automated decision aids, and are failing to incorporate other relevant cues into their decision making calculus.

Although automated decision aids will never be 100% valid, they are nonetheless going to stay, if not increase in presence, in high technology decision making settings. Even our

own data is suggestive that there are modest improvements in overall performance in decision making contexts with an automated decision aid, relative to the same context without such an aid. Therefore the goal of future research will be to discover; (1) how to design automated systems in such a way that they encourage, if not demand, that operators seek out multiple sources of information before they can make a decision, (2) how to design decision making contexts, like the airplane cockpit, so that automated decision aids are not so salient and psychologically "loud" that they overwhelm the ability of people to notice and integrate other relevant information into their decision making calculus , (3) continued investigation of training interventions (explicit training about the tendency of people to make errors of omission and commission showed some preliminary encouraging results), (4) and despite the low correlations found so far, continued investigation of whether there are any systematic and measurable individual differences in the tendency to make biased decisions in highly automated decision making contexts.

The most encouraging news that the current set of studies has to offer is that the presence of automated information does not literally bias how people process other information. In other words, information from sources like gauges are not automatically discounted in the context of information from automated sources. If in fact we had found evidence that a discounting process accounted for the effects of automation bias, strategies for potential intervention would be very, very difficult if not impossible to construct. Given that there does not appear to be a major sense among people that automated information is inherently better than information from other sources (as demonstrated in the scenario studies), there is a more reasonable hope that appropriate interventions can be designed. Successful interventions should be focused at helping to ensure that people notice and take into account a broader array of information (.i.e., increase situational awareness) before coming to premature closure on a decision. Given that automation bias, both errors of omission and commission, are at their root vigilance and diffusion of responsibility issues, there is a much greater chance that through design changes and training interventions that their effects can be minimized. Assuming we can develop strategies to encourage information processors to be more vigilant, or to present non-automated information in as salient a manner as automated information, we can be reasonably confident that operators can and will make more rational and safe decisions.

References

Allport, F. H. (1920). The influence of the group upon association and thought. Experimental Psychology, 3, 159-182.

Billings, C. E. (1991). Human-centered aircraft automation: A concept and guidelines. (Tech. Mem. No. 103885). Moffett Field, CA: NASA Ames Research Center.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message and cues in persuasion. Journal of Personality and Social Psychology, 39, 752-766.

Chen, S. C. (1937). Social modification of the activity of ants in nest-building. Physiological Zoology, 10, 420-436.

Costa, P. T. & McCrae, R. R. (1992). NEO PI-R Professional Manual. Odessa, FL: Psychological Assessment Resources, Inc.

Cvetkovich, G. (1978). Cognitive accommodation, language, and social responsibility. Social Psychology Quarterly, 41, 149-155.

Diehl, A. (1991). The effectiveness of training programs for preventing air crew "error." In Proceedings of the Sixth International Symposium on Aviation Psychology (pp. 640 -655. Columbus, OH: The Ohio State University.

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. Journal of Personality and Social Psychology, 62, 328-339.

Fiske, S. T. & Taylor, S. E. (1994). Social Cognition (2nd Ed.), New York: MacGraw Hill.

Gates, M. F., & Allee, W. C. (1933). Conditioned behavior of isolated and groups cockroaches on a simple maze. Journal of Comparative Psychology, 15, 331-358.

Geen, R. G. (1989). Alternative conceptions of social facilitation. In P. B. Paulus (Ed.) The psychology of group influence (2nd Ed.), pp. 16-31. Hillsdale, NJ: Lawrence Erlbaum.

Geen, R. G. & Gange, J. J. (1977). Drive theory of social facilitation: Twelve years of theory and research. Psychological Bulletin, 84, 1267-1288.

Hagafors, R. & Brehmer, B. (1983). Does having to justify one's decisions change the nature of the decision process? Organizational Behavior and Human Performance, 31, 223-232.

Harkins, S. G., & Szymanski, K. (1989). Social loafing and group evaluation. Journal of Personality and Social Psychology, 56, 934 - 941.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. Journal of Experimental Social Psychology, 10, 371-384.

Johnson, E. J., Payne, J. W., Schkade, D.A., Bettman, J. R. (1991). Monitoring information processing and decisions: The Mouselab system. Philadelphia: University of Pennsylvania, The Wharton School.

Kahneman, D. Slovic, P. & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. NY: Cambridge University Press.

Kruglanski, A. W. & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. Journal of Experimental Social Psychology, 14, 448-468.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. Journal of Personality and Social Psychology, 65(4), 681-706.

Karau, S. J., & Williams, K. D. (1995). Social loafing: Research findings, implications, and future directions. Current Directions in Psychological Science, 4(5), 134-140.

March, J. G. & Simon, G. A. (1958). Organizations. New York: Wiley.

Mayseless, O., & Kruglanski, A. W. (1985). What makes you so sure? Effects of epistemic motivations on judgmental confidence. Organizational Behavior and Human Decision Processes, 39, 162-183.

McAllister, P. W., Mitchell, T. R., & Beach, L. R. 91979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility. Organizational Behavior and Human Performance, 24, 228-244.

Milgram, S. (1974). Obedience to authority. New York: Harper.

Mosier, K. L., Palmer, E. A., & Degani, A. (1992). Electronic checklists: Implications for decision making. In Proceedings of the Human Factors Society 36th Annual Meeting (pp. 7 - 11). Santa Monica, CA: Human Factors Society.

Mosier, K. L., Skitka, L. J. Burdick, M. & Heers, S. (1996). Automation bias, accountability, and verification behaviors. Paper to be presented at the Human Factors and Ergonomic Systems conference, Philadelphia, PA.

Mosier, K. L., Skitka, L. J., Heers, S. & Burdick, M. D. (1997). Patterns in the use of cockpit automation. In M. Mouloua & J. Koonce (Eds.), Human-automation interaction: Research and practice. Hillsdale, NJ: Lawrence Erlbaum Assoc., Inc. (pp. 167-173).

Mosier, K. L., Skitka, L. J., Heers, S. & Burdick, M. D. (1998). Automation bias: Decision making and performance in high-tech cockpits. International Journal of Aviation Psychology, 8(1), 47 - 63.

Mosier, K. L., Skitka, L. J., & Korte, K. J. (1994). Cognitive and social psychological issues in flight crew/automation interaction. In M. Mouloua and R. Parasuraman (Eds.), Human performance in automated systems: Current research and trends. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. (pp. 191-197).