Transitioning to Future Air Traffic Management: Effects of Imperfect Automation on Controller Attention and Performance

Ericka Rovira and **Raja Parasuraman,** George Mason University, Fairfax, Virginia

Objective: This study examined whether benefits of conflict probe automation would occur in a future air traffic scenario in which air traffic service providers (ATSPs) are not directly responsible for freely maneuvering aircraft but are controlling other nonequipped aircraft (mixed-equipage environment). The objective was to examine how the type of automation imperfection (miss vs. false alarm) affects ATSP performance and attention allocation.

Background: Research has shown that the type of automation imperfection leads to differential human performance costs.

Method: Participating in four 30-min scenarios were 12 full-performance-level ATSPs. Dependent variables included conflict detection and resolution performance, eye movements, and subjective ratings of trust and self confidence.

Results: ATSPs detected conflicts faster and more accurately with reliable automation, as compared with manual performance. When the conflict probe automation was unreliable, conflict detection performance declined with both miss (25% conflicts detected) and false alarm automation (50% conflicts detected).

Conclusion: When the primary task of conflict detection was automated, even highly reliable yet imperfect automation (miss or false alarm) resulted in serious negative effects on operator performance.

Application: The further in advance that conflict probe automation predicts a conflict, the greater the uncertainty of prediction; thus, designers should provide users with feedback on the state of the automation or other tools that allow for inspection and analysis of the data underlying the conflict probe algorithm.

Keywords: trust, NextGen, automation, eye movement, air traffic management, air traffic control

Address correspondence to Ericka Rovira, U.S. Military Academy, Behavioral Sciences and Leadership, 267C Thayer Hall/Engineering Psychology, MADN-BS&L/646 Swift Rd., West Point, NY 10996; Ericka.Rovira@usma.edu.

HUMAN FACTORS
Vol. 52, No. 3, June 2010, pp. 411–425.
DOI: 10.1177/0018720810375692.
Copyright © 2010, Human Factors and Ergonomics Society.

INTRODUCTION

Background

The National Airspace System (NAS) is under increasing pressure to handle a steady growth in air travel, which is straining air traffic control (ATC) services. One solution is to partition traffic sectors to handle increased demand, but this solution necessitates additional communication and coordination between pilots and air traffic service providers (ATSPs) and among ATSPs. Creating smaller sectors also makes it difficult for ATSPs to build up an understanding of the scheme of traffic flow (necessary for safe separation) because aircraft spend less time in the sector. Continued division of sectors is therefore not a viable option. These and other considerations have led to several proposals for new air traffic management concepts, including Free Flight (FF), Distributed Air Ground Traffic Management (DAG-TM), and most recently, Next Generation Air Transportation System (NextGen).

The goal of FF, as stated by the Radio and Technical Commission for Aeronautics (RTCA; 1995), is to allow aircraft under instrument flight rules the ability to choose, in real time, optimum routes, speeds, and altitudes in a manner similar to the flexibility now given only to aircraft operating under visual flight rules (Parasuraman, Hilburn, & Hoekstra, 2001). DAG-TM has been defined as the use of distributed decision making by flight deck crews, ATSPs, and airline operational control facilities to enable user preferences and increase system capacity while meeting air traffic management (ATM) safety requirements (National Aeronautics and Space Administration, 1999). Finally, and most recently, NextGen has been proposed with the goals of expanding capacity, ensuring safety, protecting the environment, ensuring national defense, and securing the nation for all aerospace transportation via networked enabled information access (Joint Planning and Development Office, 2007).

All of these proposals are not necessarily seen as competing; generally, they are ideas developed through the years incorporating research findings, technological capabilities, and most sobering, the events of 9/11. However, there is disagreement as to how increased capacity may be attained. A proposed solution is to give aircraft separation responsibility to the pilots. However, this leaves the resulting issue of having ATSPs act as monitors needing to intervene in instances in which failures may occur (Wickens, Mavor, & McGee, 1997). Therefore, greater use of automation to support ATSPs in an increasingly dense air space is necessary (Wickens, Mavor, Parasuraman, & McGee, 1998).

Transition to Future ATM

The transition into future ATM will be gradual, so that initially only some aircraft will have the equipment to participate fully in all aspects of the new system, whereas other less-well-equipped aircraft will not. This is the *mixed-equipage* issue. If pilots have full responsibility for their own separation from other aircraft, then ATSPs would be relieved of monitoring freely maneuvering aircraft and could provide better service to managed aircraft in an increasingly dense airspace.

To date, there is little supporting evidence that ATSPs can appropriately disengage their attention from the freely maneuvering aircraft in an increasingly dense airspace. Three recent studies investigated ATSP performance in a mixed-equipage environment where some of the separation responsibility was delegated to pilots with the onboard technology to self-separate. Corker, Fleming, and Lane (1999) examined the effects of different proportions of mixed equipage (100% managed, 80% managed, 20% managed) on ATSP detection of airborne self-separations and mental workload. They found that with 80% of the aircraft managed, mental workload was comparable to a condition in which all aircraft were managed as in current ATM. However, with 20% of the aircraft managed, mental workload significantly *increased* compared with the 80% and 100% managed conditions. The authors speculated that some form of decision support automation could ameliorate the counterintuitive increase in workload posed by the unmanaged aircraft for ATSPs (e.g., presentation of pilot intent information or pilot conformance monitoring).

Another study (Metzger, Rovira, & Parasuraman, 2003) investigated how mixed equipage affects ATSP performance and mental workload with different proportions of managed and unmanaged aircraft. Although some evidence of negative consequences for ATSP performance in a mix of managed and unmanaged traffic was obtained, it was also found that support tools could compensate for these effects. Additionally, mental workload was reduced only slightly with high proportions of unmanaged traffic, suggesting that aircraft providing their own separation assurance might not reduce ATSP workload as much as expected.

Most recently, Prevot, Homola, Mercer, Mainini, and Cabrall (2009) completed three-part task studies investigating trajectory-based conflict detection automation with high traffic densities in a mixed-equipage environment. Their main finding was a mental workload reduction for ATSPs when conflict detection responsibility was assigned to the automation. This finding suggests that reliable automated conflict detection would enable the handling of significantly more aircraft than today. Additionally, it was found that higher traffic densities require higher levels of automation, and mixed-equipage operations are feasible to the extent that the higher the traffic density of equipped aircraft, the lower the number of unequipped aircraft ATSPs can manage in the same airspace. These findings indicate that ATSPs must be supported by automated tools in future ATM concepts so that conflict detection performance can be maintained.

Automation Support

Metzger and Parasuraman (2005) showed that a "look-ahead" conflict probe that highlighted aircraft projected to be in conflict improved ATSP conflict detection performance. However, because conflict prediction is associated with an inherent uncertainty, automated aids cannot be perfect. Metzger and Parasuraman found that the benefit of the conflict probe automation was offset by impairment in performance when the automation was unreliable. Although automated systems are usually highly reliable, they are not

always 100% reliable. This fact does not necessarily imply that there is a problem with the computational algorithm underlying the automation. Automation imperfections may arise for a variety of reasons; for instance, the algorithm may not be context sensitive, the system may have been compromised, or environmental conditions may affect sensor precision. Consequently, when human operators interact with imperfect systems, human performance may be adversely affected.

Recently, issues related to how different types of automation imperfection affect operator performance have begun to be investigated (Dixon, Wickens, & McCarley, 2007; Maltz & Shinar, 2003; Meyer, 2001, 2004; Parasuraman & Wickens, 2008). In the context of signal detection theory, automation may be imperfect by providing a miss or a false alarm (Green & Swets, 1988). Depending on the type of automation imperfection (miss or false alarm), operator reliance on or compliance with an automated system will be affected (Meyer, 2001).

In a highly reliable yet imperfect automated diagnostic system, the operator's trust may not be well calibrated to the reliability of the system. Consequently, if an automated alert is not provided, signifying normal operations when in fact nonnormal operations are occurring, the human operator may become overreliant on the automation. Operators may then allocate attention to concurrent manual tasks because they rely on the automation to correctly alert them of any impending hazard. If the automation fails to announce a problem in the form of a miss, the operator should become less reliant and pay closer attention to the raw data (e.g., radar display), resulting in better detection performance. Hence, miss-prone automation will reduce reliance, whereas infrequent misses will encourage overreliance (Parasuraman, Molloy, & Singh, 1993).

Simply increasing the number of misses in a real-world system would be disastrous, particularly for systems in which the detection of an event may result in a life-or-death situation. False alarms are therefore traditionally considered more acceptable. Compliance, on the other hand, describes operator behavior when the automated alert is provided, regardless of whether it is true or false. Operators who are compliant will allocate attention to the alert in an effort to initiate appropriate responses. Compliance is typically influenced by false alarm–prone automation. In this instance, operators' compliance with alarms will be degraded, resulting in either a delayed response or no response to the automated alert (Breznitz, 1983).

Evidence of reliance and compliance has been found in single- (Meyer, 2004) and multitask domains (Dixon & Wickens, 2006). Specifically, automation misses are correlated with poorer performance on concurrent tasks and false alarms with poorer performance on the automated task (Dixon & Wickens 2006; Parasuraman & Wickens, 2008).

As suggested by Parasuraman et al. (1993), an underlying factor in operator overreliance is suboptimal sharing of attentional resources between manual and automated tasks (see also Moray & Inagaki, 2000). The assumption is that operators' high trust in the automation results in a reduction in attention allocation to the automated system and to the "raw" data feeding the automation in comparison with manual conditions. In some instances, the raw data are readily available in a display, whereas in other displays, they may be buried under a few layers and the operator needs to dig to find them.

Dixon and Wickens (2006) found that perfectly reliable automation reduced visual attention to a system monitoring task compared with manual conditions when the raw data were readily available in the main display. Furthermore, they found that with miss-prone automation, operators allocated more visual attention to the system monitoring task, as compared with false alarm–prone automation. They also found that scan response times were as fast in the miss-prone automation condition as in the perfectly reliable conditions; however, they were as slow in the false alarm condition as in the manual condition. These findings support the reliance-compliance distinction wherein miss-prone automation resulted in increased visual attention to the system monitoring task at a cost to the other tasks, and false alarm-prone automation resulted in slowed alert-driven shifts of visual attention to the system monitoring task.

Does the reliance-compliance distinction apply in real systems? In 2006, the National Transportation

Safety Board reported that a series of aviation accidents were a result of a number of missed alerts potentially attributed to the high false alarm rate in two alerting systems used by ATSPs, the Minimum Safe Altitude Warning (MSAW) and Conflict Alert (NTSB, 2006). Wickens et al. (2009) examined naturalistic traffic data from en route ATC facilities. They found that the greater the false alarm rate in a center, the less ATSPs tended to respond; however, there was no relationship between false alarm rate and loss of separation rate.

These studies point to certain vulnerabilities in ATSP performance in advanced ATM concepts, particularly under high traffic load. Automation support can alleviate the problem, but automation imperfection can further hinder effective performance, depending on whether automation failures are misses or false alarms.

Research Questions and Hypotheses

The present study examined the effects of reliable and imperfect automation on controller attention and performance. The study differed from previous work in the following ways. First, because most studies of future ATM concepts have examined scenarios in which *all* aircraft operate under the future concept, we investigated whether automation benefits would occur in a mixed-equipage environment where ATSPs were no longer responsible for detecting self-separations for freely maneuvering aircraft. We predicted that ATSPs would perform better by detecting more conflicts, detecting conflicts sooner, and providing timely resolutions with reliable automation as compared with the manual condition.

Additionally, although eye movements do not provide precise information about the locus of cognitive processing in all situations, attention and eye movements remain closely related (Irwin, 2004). Therefore, given the link between eye movements and visual attention, we predicted that with automation, ATSPs would exhibit fewer fixations to the radar display compared with manual control. We make this prediction because the automation would highlight potential conflicts, leaving ATSPs free to focus on secondary tasks.

Second, although previous studies have found differential costs associated with types of automation imperfection—miss and false alarms—when operators are supported with automation on secondary tasks (Dixon & Wickens, 2006), we investigated types of automation imperfection when the *primary* task, conflict detection, was automated. In this study, ATSPs were provided with an automated tool—a "look-ahead" conflict probe—that highlighted aircraft projected to be in conflict 6 min in the future. Our general hypotheses were that with miss automation, (a) eye movements to the primary task would increase at a cost to the other current tasks, whereas (b) conflict detection performance would improve because the operator would pay closer attention to the radar display. However, with false alarm automation, it was expected that (a) there would be reduced visual attention to the primary task, and (b) performance would reflect either a delayed response or no response to the automated conflict probe.

METHOD

Participants

For this study, 12 full-performance ATSPs (all male) ages 27 to 49 years (M = 41, SD = 5.10) volunteered and were paid $30 per hour for their participation. Of these, 7 ATSPs were from the Washington, D.C., Air Route Traffic Control Center (ARTCC), 3 were from the Washington area combined Tower and Terminal Radar Approach Control (TRACON) facility, 1 worked at Washington National Tower, and last, 1 was from International Airport Dulles TRACON facilities. Participants' average work experience, including military and civilian positions, was M = 17.5 years, SD = 3.63. En route and other types of ATSPs did not differ in age, $F(1, 10) = .34$, $p > .05$, or experience, $F(1, 10) = .18$, $p > .05$. Visual inspection of the data shows no difference in performance between the tower controller and other participants.

Apparatus

Applied Science Lab (ASL) Eye Tracker 5000. An ASL Model 501 eye tracking system with head-mounted optics was used to detect eye movements at a sampling rate of 60 Hz. The head-mounted eye tracker was used in conjunction with the Ascension Flock of Birds™ to measure the participant's eye line of gaze with

Figure 1. Air traffic control simulator setup.

respect to his head position. This setup allowed participants the freedom to move their heads without loss of data.

Eye position was determined by the corneal subtraction method (ASL, 1999). ASL provides an analysis program in which the fixation parameters were set to be the mean *x*- and *y*-coordinates measured for a minimum of 100 ms during which the eye did not move more than 1° of visual angle both vertically and horizontally. A dwell was defined as a series of contiguous fixations within a defined area of interest.

ATC simulator. A PC-based medium-fidelity ATC simulator (Masalonis et al., 1997) was used to simulate en route airspace. The program was written in C code for a Pentium-II processor with two 21-in. monitors attached. A trackball was used to go between the two monitors. A picture of the complete setup is shown in Figure 1.

The ATC simulator consisted of a radar display and a data link display shown on two adjacent monitors. The radar displayed the 50-mile radius sector consisting of all waypoints, jet routes, aircraft targets, and data blocks. It simulated U.S. airspace; however, names of all waypoints were changed, and the sectors were rotated. This was done in an effort to avoid any previous training effects in the event ATSPs had expertise working a particular region.

The automated conflict probe tool is also shown in Figure 2. It highlighted aircraft projected to be in conflict within the next 6 min. On the radar display, 6 min before a conflict would occur, red bubbles appeared around the aircraft involved in the conflict and would remain on until the initial point of loss of separation or until the ATSPs performed resolution responsibilities. Participants were instructed that the aid was highly reliable but not 100% reliable. The tool was an "intelligent" conflict detection aid to the extent that it used an algorithm based not only on current speed and altitude but also on any scheduled heading changes contained in the aircraft's flight plan. Note that if ATSPs failed to detect a conflict, regardless of condition, they were provided with feedback; the same red bubbles as in the conflict probe would appear around the two aircraft in conflict for the duration of the loss of separation.

The data link display was presented on the leftmost monitor and consisted of three components shown in separate windows: a list of flights, a communication module, and an electronic flight strip window. The simulation did not allow for voice communication. Therefore, the data link

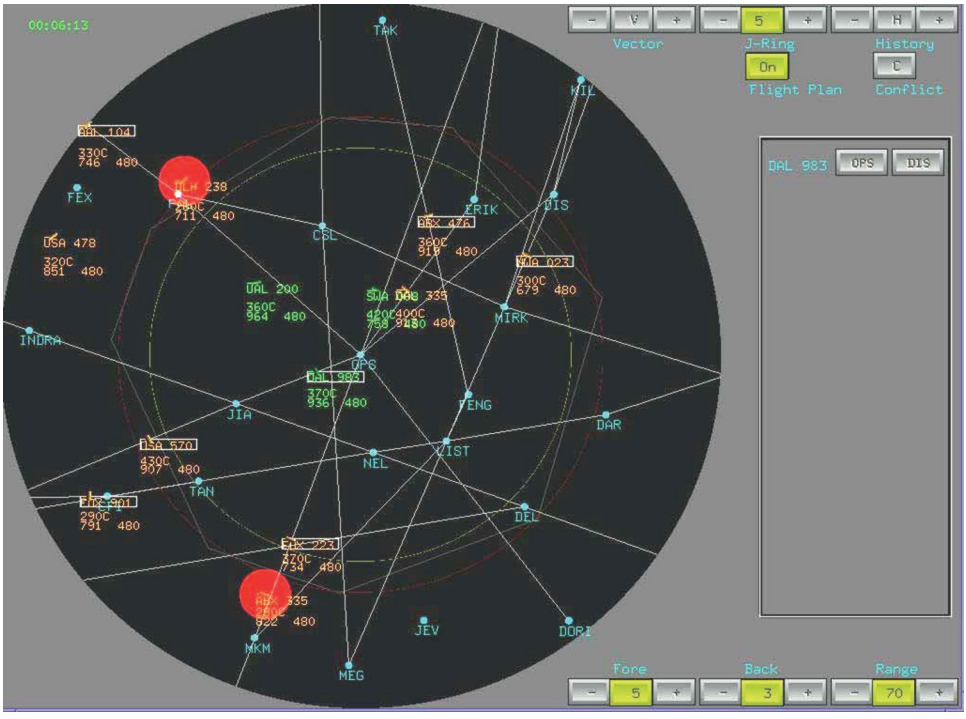

Figure 2. Primary visual display.

display was a means of communication between the pilot and the air traffic controller and substituted for all voice communication. Communication included information regarding when an aircraft entered and exited the sector as well as pilot intent information. A picture of the data link display is shown in Figure 3.

Design

This study was a single factor (automation support) within-subjects design with four levels: (a) manual, (b) reliable automation, (c) miss automation, and (d) false alarm automation. Half of the participants received the manual condition first, whereas for the other half, it was received last. When automation was provided, the reliable condition was presented before the unreliable condition. This procedure was done intentionally so that ATSPs received sufficient experience with reliable automation in an effort to build their trust in the support tool. The ordering of the automation failure conditions was such that half of the participants performed the miss automation condition before the false alarm automation condition, and the order was reversed for the other half of the participants. This resulted in a double crossover design, whereby the first crossover was manual versus automation conditions, and the second crossover was the order of the two automation failure conditions.

A single scenario was used for the manual and reliable automation conditions. However, the sector boundaries, jet routes, and traffic patterns were rotated and aircraft flight names and waypoints were changed so that it would be unrecognizable to ATSPs. The resulting scenarios for the manual and reliable conditions were essentially the same, ensuring that differences in ATSP performance were attributable to the manipulation of automation support type and not to specific features of a particular scenario, such as conflict geometry (Castaño & Parasuraman, 1999). A second scenario was developed for the miss and the false alarm scenarios; similarly, it was rotated and names were changed. This method allowed for a direct comparison of conflict detection performance between the miss and false alarm conditions. Table 1 depicts the three conflicts that occur per condition, where x denotes the occurrence of a conflict for the

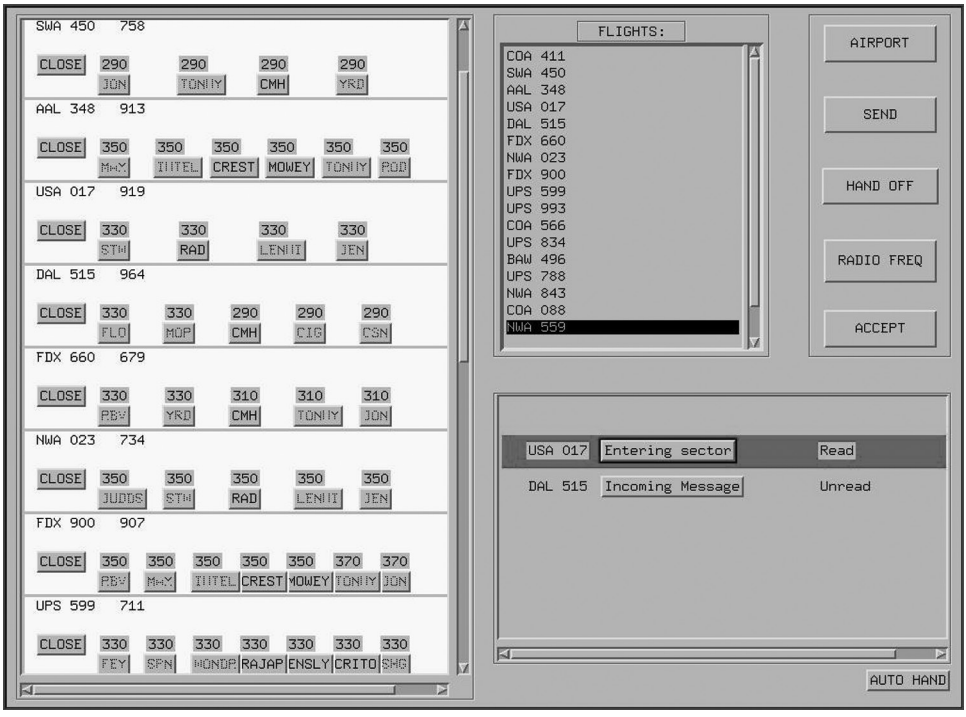

Figure 3. Data link display.

TABLE 1: Experimental Design

	Conflict 1	Conflict 2	Conflict 3
CONDITION			
Manual	x	x	x
Reliable Automation	x	x	x
Miss Automation	x	No Alert	x
False Alarm Automation	x	False Alert & True Alert	x

manual condition and a reliably cued conflict for reliable automation, miss automation, and false alarm automation.

Task Procedures and Dependent Variables

ATSPs received 1 hr of instruction and training before performing the four 30-min scenarios. During this time, they were given instruction on all tasks and the automation support tool; additionally, they were allowed to practice the tasks both manually and with the automation. Sector density for each scenario was 24 aircraft. The proportion of mixed equipage was such that 80% of the aircraft were fully managed and 20% were freely maneuvering. The fully managed aircraft had yellow boxes around the data blocks to draw ATSPs' attention to these aircraft. All freely maneuvering aircraft were above 36,000 feet, and all fully managed aircraft were below 36,000 feet.

There were a total of six potential conflicts (three self-separations and three actual conflicts); however, ATSPs were instructed that they were not responsible for detecting conflicts between self-separating aircraft. These aircraft had the onboard technology to safely separate themselves from the other aircraft in the sector; therefore, there were no conflict probes for the self-separations. The three actual conflicts were

all between managed aircraft. All potential conflicts were randomly distributed during the 30-min scenario.

A reliably cued conflict occurred both before and after each automation failure to encourage trust in the automation. Therefore, for each automation failure condition, there was one miss event in the miss automation condition and one false alarm in the false alarm automation condition. Furthermore, when the false alarm was presented to the participants, an additional reliably cued conflict was also present. The reliably cued conflict was necessary to directly compare conflict detection performance between the miss automation and the false alarm automation; otherwise, if the controllers responded to the false alarm by no action, we would be unable to confirm whether they appropriately responded or failed to detect the conflict.

The detection of conflicts served as a measure of primary task performance. Advance notification time was defined as the time period between a potential loss of separation and the time the ATSP reported the detection of a potential conflict. The greater the value, the earlier the detection took place and the better the conflict detection performance. The separation standards for the experiment were 5 nautical miles horizontally and 1,000 feet vertically. Aircraft moved 8 nautical miles per minute.

Participants could provide either an altitude or a lateral change to resolve the conflict by using the trackball to select the aircraft and then the keyboard to enter a new altitude or new heading using the keys located in the upper right corner of the radar screen to initiate and complete the process. The type of resolution implemented (altitude or lateral) and the timeliness of resolutions were recorded. After a resolution was provided, that aircraft symbol was removed and would no longer be seen on the radar display. One drawback is that ATSPs were unable to see their resolution develop and consequently did not know whether they had resolved the conflict or how successfully they had done so. The aircraft symbol was removed immediately after the initiation of a resolution in an effort to maintain a degree of experimental control in case the ATSPs provided a resolution that would create another conflict in addition to the ones planned by the experimenter.

Participants were additionally instructed to accept all aircraft that entered their sector. For each incoming aircraft, ATSPs received notification on the data link display. They would then read the message "Entering Sector" and click Accept to take over responsibility for that aircraft. Participants could then click on the call sign of the aircraft in the list of flights, and the flight strip would open. ATSPs were instructed to hand off an aircraft to the next sector just before the aircraft reached the sector limits as designated by a ring within the octagon of the sector.

Participants performed an embedded secondary task. ATSPs needed the flight strip to monitor the flight plan of aircraft. As soon as an aircraft passed one of the waypoints on the radar display, ATSPs needed to click on the corresponding waypoint on the electronic flight strip. Participants were instructed to perform the primary tasks (conflict detection and resolution, accepting incoming aircraft, and handing off aircraft) to the best of their ability and as much as possible perform the secondary task. Accuracy and timeliness were calculated for accepting incoming aircraft, updating flight strips, and handing off aircraft.

Subjective ratings of trust in the automation (conflict probes) and operator self-confidence to perform without the automation were obtained on a Likert-type rating scale (ranging from 0 to 100) modeled after scales used by Lee and Moray (1992, 1994) and were administered after the three automation conditions. We additionally obtained separate trust and self-confidence ratings for freely maneuvering aircraft after all four scenarios.

Last, ocular activity was collected in all conditions. On the basis of the hypothesis that miss automation would lead to increased visual attention (eye movements) to the automated task and false alarm automation would result in reduced eye movements to the automated task, we investigated differences in visual attention following an automation failure; hereafter, we refer to this data as postimperfection (i.e., the last 16 min of a scenario). Eye movements up until the time period when a loss of separation occurred and the automation was imperfect (also referred to as preimperfection) were excluded.

Eye movements were first analyzed by defining five areas of interest (AOIs): (a) radar display,

TABLE 2: Conflict Detection Performance for Conflicts 1 to 3 (in percentages)

	Condition			
Conflict	Manual (n = 12)	Reliable (n = 12)	Miss (n = 12)	False Alarm (n = 12)
Conflict 1	91.67 (8.33)	100 (0)	91.67 (8.33)	100 (0)
Conflict 2	75 (13.06)	100 (0)	25 (13.06)	50 (15.08)
Conflict 3	25 (13.06)	75 (13.06)	83.33 (11.24)	91.67 (8.33)

Note. Means shown with standard errors in parentheses.

(b) conflict detection and resolution, (c) flights list, (d) electronic flight strips, and (e) communication window. The ASL analysis program was used to match fixations and dwells to these areas. The number and duration of fixations and dwells to each area were computed.

Second, we developed an algorithm to compute the number of fixations to a specific moving target (i.e., an aircraft). This algorithm allowed for the investigation of the number of fixations to conflict aircraft. We evaluated ATSPs' attention allocation to the aircraft pairs in conflict by computing the sum of fixations to each conflict pair.

RESULTS

According to the hypotheses, three planned orthogonal contrasts were used to analyze whether performance was affected by (a) automation versus manual control, (b) reliable versus imperfect automation support, and (c) miss versus false alarm automation. Eye movement data were analyzed after the automation failure in an effort to avoid washing out performance differences attributable to averaging pre- and postimperfection data. Conflict detection performance was analyzed at three different data points: Conflict 1 (reliable automation for all conditions), Conflict 2 (imperfect automation for miss and false alarm conditions), and Conflict 3 (reliable automation for all conditions).

The F value for the omnibus analysis (i.e., one-way repeated-measures ANOVA with four levels of the independent variable) is presented first, followed by the F values for the three planned orthogonal contrasts for all dependent variables as appropriate. Because of space limitations and little variability in the data, performance for accepting incoming aircraft, handing off aircraft, updating flight strips, lateral and altitude resolutions, and subjective ratings of mental workload will not be reported.

Primary Task Performance

Detection of conflicts. There were no significant effects for the first, pre-automation failure conflict. Conflict detection accuracy was greater than 90% for all four conditions. Table 2 reflects little variability in the data.

For the second conflict, ATSPs detected conflicts better with reliable (M = 100%, SE = 0%) versus imperfect automation (M = 37.5%, SE = 14.07%), $F(1, 33) = 19.64$, $p < .01$. There was a trend for better performance with false alarm automation (M = 50%, SE = 15.08%) as compared with miss automation (M = 25%, SE = 13.06%), but this trend was not significant, $F(1, 33) = 2.35$, $p = .13$. There were no other significant results. Table 2 shows data for Conflict 2 in the four conditions.

For the third conflict, post–automation failure, ATSPs detected conflicts better with (M = 83.33%, SE = 10.88%) than without automation support (M = 25%, SE = 13.06%), $F(1, 33) = 20.13$, $p < .01$. There were no other significant effects. Table 2 shows data for Conflict 3 in the four conditions.

Advanced notification time for conflicts. For the first conflict in each condition, ATSPs detected the conflict on average 296.07 s (SE = 15.42 s) prior to the impending loss of separation with automation support as compared with manual conditions (M = 218.28 s, SE = 15.42 s), $F(1, 33) = 12.71$, $p < .01$. ATSPs detected the conflict earlier under imperfect (M = 350.78 s, SE = 36.38 s) versus reliable automation conditions (M = 186.66 s, SE = 13.49 s), $F(1, 33) = 46.18$, $p < .01$, attributable to either more

practice (all participants completed the reliable scenario prior to either of the imperfect conditions) or the type of conflict (the latter is further addressed in the Discussion section). There were no other significant findings for conflict one.

For the second conflict, in 50% or more of the cells in the miss and false alarm conditions, no advance notification time was available because of the high number of missed events. Thus, there were too few data points to carry out meaningful inferential statistics for the three contrasts. The same was true for the third conflict.

Subjective Ratings

Trust and self-confidence. ATSPs were asked to assess their trust and self-confidence on the automation support tool. Because ATSPs could rate their trust and self-confidence only on the basis of using the automation support tool, a single-factor ANOVA with three levels (reliable automation, miss automation, and false alarm automation) was performed. The effect of automation condition was not significant for amount of trust, $F(2, 22) = .55$, $p > .05$. There was a significant effect of automation condition on ATSPs' self-confidence, $F(2, 22) = 4.24$, $p = .028$. ATSPs rated reliable automation ($M = 69.16$, $SE = 7.9$) as having improved their performance more than did imperfect automation ($M = 61.28$, $SE = 8.02$).

We were additionally interested if ATSPs' perceptions of the freely maneuvering aircraft, experienced in all scenarios, would be altered by their experience with the varying types of automation support tools. Regardless of the type of automation support (manual, reliable automation, miss automation, or false alarm automation), ATSPs' ratings of trust in the freely maneuvering aircraft did not vary, $F(3, 33) = 1.29$, $p > .05$. The omnibus analysis investigating ATSPs' self-confidence with freely maneuvering aircraft was significant, $F(3, 33) = 3.63$, $p = .023$. ATSPs rated that the freely maneuvering aircraft improved their performance more with automation support ($M = 76.53$, $SE = 8.50$) than without ($M = 66.67$, $SE = 11.32$), $F(1, 33) = 6.45$, $p = .02$. There were no other significant results.

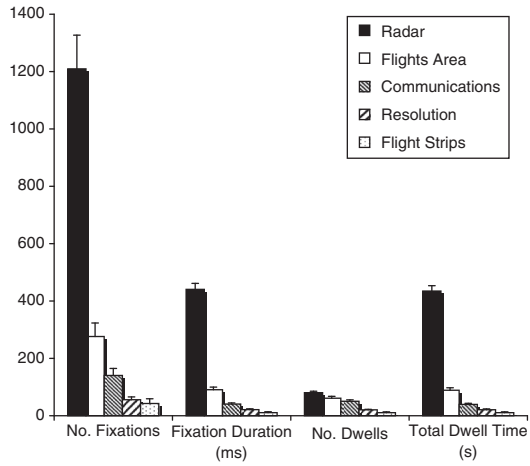

Figure 4. Eye movement data for each area of interest.

Eye Movements

The eye movement data were analyzed by first computing several metrics for each of the five AOIs previously defined (radar display, flights area, communications, resolution, and flight strips). These metrics included number of fixations, duration of fixations, number of dwells, and total dwell time. Figure 4 shows the values of these metrics for the five different AOIs. Next, an ANOVA was run to find whether eye movements varied by the five AOIs. Last, similar to results presented previously, the F value for the omnibus analysis (i.e., one-way repeated-measures ANOVA with four levels of the independent variable) is presented first, followed by the F values for the three planned orthogonal contrasts for each eye movement metric as appropriate.

Number of fixations, duration of fixations, number of dwells, and duration of dwells to AOIs. The number of fixations, $F(4, 44) = 150.02$, $p < .01$; duration of fixations, $F(4, 44) = 150.02$, $p < .01$; number of dwells, $F(4, 44) = 54.76$, $p < .01$; and duration of dwells, $F(4, 44) = 402.36$, $p < .01$, to the different AOIs was significantly different. As Figure 4 shows, most fixations were made to the radar display, followed by the flights area, communications, resolution, and flight strips. Duration of fixations, number of dwells, and duration of dwells reflected the same distribution to the AOIs.

Given that ATSPs mostly fixated the radar display, the remainder of the analyses focused on fixations to the radar display. Only for the number of dwells, the effect of the automation condition was marginally significant, $F(3, 33) = 2.50$, $p = .076$. There was an increased number of dwells on the radar display with imperfect automation ($M = 87.92$, $SE = 6.76$) as compared with reliable automation ($M = 66.92$, $SE = 7.9$), $F(1, 33) = 6.60$, $p = .015$. There were no other significant results.

Eye movements to aircraft projected to be in conflict. Using the sum of fixations to conflict pairs, we calculated a 4 (automation support: manual, reliable, miss, false alarm) × 3 (conflict: Conflict 1, Conflict 2, Conflict 3) ANOVA. A main effect of automation was found, $F(3, 33) = 4.82$, $p < .01$, with the most fixations made in the manual condition ($M = 72.19$, $SE = 12.35$) and the fewest in the false alarm condition ($M = 39.75$, $SE = 6.93$). A main effect of conflict was also found to be significant, $F(2, 22) = 29.22$, $p < .01$, for Conflict 1 ($M = 18.44$, $SE = 4.01$), Conflict 2 ($M = 72.60$, $SE = 12.45$), and Conflict 3 ($M = 69.75$, $SE = 11.56$). Last, an interaction between automation and conflict was found to be significant, $F(6, 66) = 3.30$, $p < .01$. Figure 5 shows that the effect of automation on number of fixations was greater with conflicts two and three as compared to conflict one.

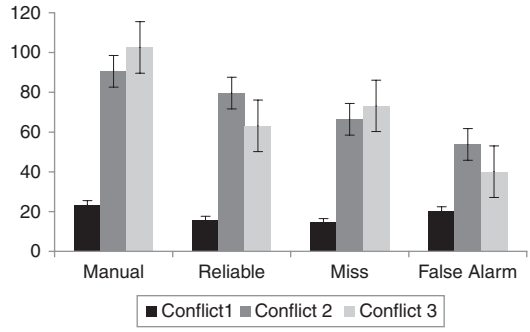

Figure 5. Fixations to conflict pairs.

DISCUSSION

This experiment examined two major issues concerning ATSP performance with imperfect conflict probe automation support in a simulated mixed-equipage environment of the type that will occur in the NextGen program. First, do the benefits of conflict probe automation occur when ATSPs are not responsible for conflict detection of freely maneuvering aircraft? Second, how do different types of automation imperfection (miss vs. false alarm) affect ATSP performance and attention allocation?

First, reliable automation led to ATSPs' detecting conflicts more accurately and sooner as compared with manual performance. As Table 2 depicts, conflict detection performance in the manual condition, as compared with when operators were supported with reliable automation, for the third conflict was particularly poor (25%). This poor performance could be a function of the conflict geometry. Although not a variable of interest in this study, previous work has found that sometimes conflict geometry does affect controller conflict detection performance (Wickens et al., 2009) and sometimes does not (Nunes & Scholl, 2004). Wickens et al. (2009) found increased anticipatory responses to conflicts when tracks were converging on the radar display as compared with when they were diverging or in parallel as in the instance of this third conflict.

Second, imperfect automation degraded conflict detection performance. As expected, this led to an increased number of visual dwells to the radar display following an imperfection. Furthermore, ATSPs felt more self-confident to perform without the automation when they were supported with imperfect automation as compared with reliable automation. With respect to types of automation imperfection, there was a trend for worse performance with miss automation as compared with false alarm automation.

The hypothesis that ATSPs would detect more conflicts and detect conflicts sooner when they were supported by reliable automation as opposed to performing the tasks manually was upheld. This finding, along with the previous results of Metzger et al. (2003) and Prevot et al. (2009), suggests that ATSP performance can be effectively supported in a mixed-equipage future ATM environment if reliable automation tools are available. Although there is some previous evidence that ATSP workload may be adversely affected by the requirement to work in a setting with both managed and unmanaged aircraft (Corker et al., 1999), the bulk of the evidence

indicates that this issue can be mitigated with effective automation support.

The problem arises, however, that conflict probe automation cannot be 100% reliable. When the automation was imperfect, operator conflict detection performance degraded catastrophically, also supporting the previous finding on automation imperfection by Metzger and Parasuraman (2005). This result is consistent with the view that although automation supports operators, it moves them farther away from the decision-making process; decisions are based on prior cognitive processing, including the acquisition and perception of information from multiple sources and the manipulation of information in working memory (Parasuraman, Sheridan, & Wickens, 2000). As a result, when the automation is imperfect, operators are caught overtrusting the automation, resulting in poor performance (Metzger & Parasuraman, 2001).

The increased number of dwells to the radar display with imperfect automation may explain why the typical eye movement related effects of overreliance were not found, that is, reduced visual attention with versus without automation support. A simple inspection of the data showed a decreased number of fixations with reliable automation support (M = 66.92, SE = 7.9) as compared with manual control (M = 78.38, SE = 8.86). It could be that following an automation imperfection, ATSPs made more dwells to the radar display in an effort to improve their mental picture or even to prevent future failed detections. This possibility is supported by improved conflict detection following an imperfect automation event. Similar eye movement and behavioral results were found by Dixon and Wickens (2006) to a concurrent task when operators were supported with imperfect automation.

Detection of the reliably cued conflict suffered when another conflict pair was incorrectly announced (false alarm event). Previous work (Meyer, 2001; Dixon & Wickens, 2006) has found evidence for the "cry-wolf" phenomenon, which may be defined as excessive and unnecessary alerts resulting in operator distrust in the automation. Although in this study, there was only one false alarm event, some evidence for the cry-wolf phenomenon was upheld in that 50% of the ATSPs failed to respond to the reliably cued conflict when an unnecessary alert for another conflict pair was present.

Recent work involving live conflict alert response data did not find a relationship between conflict alert rate and less safe separation performance (Wickens et al., 2009); this result is in contrast to our finding. Wickens et al. (2009) provides several reasons for not finding the cry-wolf effect, including controllers' perceiving some false alerts as "acceptable" because of conservative algorithms (Lees & Lee, 2007), cultural differences in acceptance among centers, and last, relative lack of evidence for the cry-wolf phenomenon when the primary task is automated, with the exception of the present study.

Alternatively, ATSPs may have failed to detect the reliably cued event because they experienced excessive workload and consequently could not respond because they were busy cross-verifying the false alarm event. However, neither an increase in eye movements nor ratings of subjective workload support this explanation.

The hypothesis that eye movements to the automated task would increase with miss automation compared with false alarm automation was partially upheld. Even though there were no differences in the number or duration of fixations and dwells to the radar display, ATSPs made fewer fixations to the aircraft projected to be in conflict in the false alarm condition. This is similar to previous work that has shown overreliance to be linked to suboptimal verification of the raw data input to automation (Bahner, Hüper, & Mantzey, 2008; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002; Metzger & Parasuraman, 2005).

Last, although conflict detection performance following a miss and false alarm was directly compared by evaluating the detection of the same conflicts, our hypothesis of improved conflict detection with miss automation and delayed or no response to false alarm automation was not upheld in the third conflict. The lack of performance differences in the third conflict could be due to feedback the ATSPs received. When ATSPs failed to detect a conflict, regardless of the type of automation imperfection, they were provided with feedback in the form of red bubbles appearing around the aircraft involved in the loss-of-separation

aircraft. It is very encouraging that following feedback, ATSPs' conflict detection performance for both miss and false alarm automation improved dramatically.

Limitations

The medium-fidelity en route ATC simulator used in this study may not be fully representative of real en route environments. Participants did not have to account for real-world stressors, including fatigue, environmental factors, and fatal implications. Second, in an effort to capture accurate response times and accuracies for all actions, participants did not have voice communication as a means to interface with the system; rather, they needed to complete all tasks using the mouse and keyboard. The task changes could have caused a disruption in their normal working patterns.

ATSPs are highly trained professionals, generally with years of experience working a particular sector. Another possible limitation of this research is the decision to not use live or recorded traffic. Rather, generic routes were selected to avoid having some air traffic controllers with more experience and familiarity than others. Consequently, the absolute levels of conflict detection accuracy (75.69%) cannot be compared with values in the real world, where these values would be totally disastrous and unacceptable.

The present study supported some but not all of the hypotheses that were put forward with respect to ATSP performance under a future ATM concept in a mixed-equipage environment. One limitation to the conclusions that can be drawn from the results is that only a single study was conducted. Ideally, multiple studies examining the effects of several factors should be carried out, but such studies can be expensive and time-consuming, particularly because they must require subject matter experts (air traffic controllers) who are difficult to recruit. Nevertheless, the results are consistent with the previous findings of Metzger et al. (2003) and add to the small but growing database of studies examining ATSP performance in the transition phase to future ATM.

Finally, the national airspace system is a very safe system, with only a few, rare anomalies. Hence collecting data in an environment that has few critical events is challenging, thus limiting the possibility of observing controller reactions to long-term incidence of automation failures. We did not obtain statistically significant evidence for differences in types of automation imperfection for conflict detection accuracy. The reason could be simply the low power associated with analyzing data from single trials (Wickens & Dixon, 2005).

Practical Implications

Traditionally, in highly critical situations, automation alerts tend to err in favor of false alarms versus misses. The current work found detrimental effects for both miss and false alarm automation when the primary conflict detection task was automated. This finding may suggest to designers that even highly reliable but imperfect automation that supports the primary task may be detrimental.

Following failed conflict detection, operators were provided with feedback; this procedure supported improved performance on the subsequent conflict detection event but did not alleviate poor ATSP performance when the automation failed again in subsequent scenarios. Perhaps feedback regarding not only operator performance but also system performance (e.g., when and why the automation may fail, probabilities of accuracy) may help to support better performance.

The current research found that if ATSPs are equipped with adequate pilot intent and performance feedback, but not held responsible for conflict detection assistance to equipped aircraft, ATSPs may not become overwhelmed by increased traffic density and imperfect automation as evidenced by mental workload, attention allocation, and performance. However, the rapid introduction of new technology and subsequent lack of proper system testing is problematic because the factors affecting human-automation interaction are not fully understood.

Last, although ATSPs in this study were not responsible for all aircraft in the mixed-equipage environment, marginal performance for managed aircraft in a time-critical situation with imperfect automation raises safety concerns.

ACKNOWLEDGMENTS

This work was supported by Grant GSRPORG 364230 from the National Aeronautics and Space Administration (NASA) Langley Research Center. The views expressed in this work are those of the authors and do not necessarily reflect official U.S. Army or NASA policy. Many thanks to the controllers who participated in the study and to colleagues who supported fruitful discussion: Ulla Metzger, Marla Zinni, Peter Squire, and Reshma Kumar. Last, thanks to Xiong Jiang for programming the air traffic management simulation.

KEY POINTS

- The effects of the type of automation imperfection (miss vs. false alarm) on air traffic service provider performance and attention allocation were investigated in a mixed-equipage environment.
- Imperfect conflict probe automation resulted in conflict detection performance degrading with both miss (25% conflicts detected) and false alarm automation (50% conflicts detected).
- When the primary task of conflict detection was automated, even highly reliable yet imperfect automation (miss or false alarm) resulted in serious negative effects on operator performance.
- Air traffic service providers in this study were not responsible for all aircraft in the mixed-equipage environment; marginal performance for managed aircraft in a time-critical situation with imperfect automation raises safety concerns.

REFERENCES

Applied Science Lab. (1999). Eyenal [Computer software]. Bedford, MA: Author.

Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*, 688–699.

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

Castaño, D., & Parasuraman, R. (1999). Manipulation of pilot intent under free flight: A prelude to not-so-free flight. In *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 170–176). Columbus, OH: The Ohio State University.

Corker, K., Fleming, K., & Lane, J. (1999). Measuring controller reactions to free flight in a complex transition sector. *Journal of ATC*, *4*, 9–16.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle flight control: Evaluating a reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*, 474–486.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (2nd ed.). Los Altos, CA: Peninsula.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 105–134). New York: Psychology Press.

Joint Planning and Development Office. (2007). *Concept of operations for the Next Generation of Air Transportation System: Version 2.0.* Available from http://www.jpdo.gov/

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*, 1243–1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human Computer Studies*, *40*, 153–184.

Lees, N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*, 1264–1286.

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault management in a simulated space flight microworld. *Aviation, Space, and Environmental Medicine*, *73*, 886–897.

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

Masalonis, A. J., Le, M. A., Klinge, J. C., Galster, S. M., Duley, J. A., Hancock, P. A., . . . Parasuraman, R. (1997). Air traffic control workstation mock-up for Free Flight experimentation: Lab development and capabilities. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting* (p. 1379). Santa Monica, CA: Human Factors and Ergonomics Society.

Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active vs. passive monitoring. *Human Factors*, *43*, 519–528.

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

Metzger, U., Rovira, E., & Parasuraman, R. (2003). Controller performance, workload and attention allocation in distributed air-ground traffic management: Effects of mixed equipage and decision support. In *Proceedings of the 12th International Symposium on Aviation Psychology* (pp. 803–809). Columbus, OH: The Ohio State University.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*, 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*, 196–204.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science*, *1*, 354–365.

National Aeronautics and Space Administration. (1999). *Concept definition for distributed air/ground traffic management (DAG-TM): Version 1.0.* Advanced Air Transportation Technologies (AATT) Project, Aviation System Capacity (ASC) Program. Available from http://www.asc.nasa.gov/aatt/dag.html

National Transportation Safety Board. (2006) National Transportation Safety Board (NTSB) Safety Recommendation A-06-44–A-06-47. Washington, DC: Author.

Nunes, A., & Scholl, B. J., (2004). Interactions between convergence angle, traffic load, and altitude distribution in air traffic control. In D. A. Vincenzi, M. Mouloua, & P. A. Hancock (Eds.), *Human performance, situation awareness and automation: Current research and trends* (pp. 180–185). Mahwah, NJ: Erlbaum.

Parasuraman, R., Hilburn, B., & Hoeskstra, J. (2001). Free flight. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (pp. 1005–1008). New York, NY: Taylor and Francis.

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, *30*, 286–297.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, *50,* 511–520.

Prevot, T., Homola, J., Mercer, J., Mainini, M., & Cabrall, C. (2009). Initial evaluation of NextGen air/ground operations with ground-based automated separation assurance. In *Proceedings of the 8th USA/Europe Air Traffic Management Research and Development Seminar.* Available from http://www.atmseminar.org/8th-seminar-united-states-june-2009/papers/paper_138/view

Radio Technical Commission for Aeronautics. (1995). *Report of the RTCA Board of Directors Select Committee on Free Flight.* Washington, DC: Author.

Wickens, C. D., & Dixon, S. R. (2005). *Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature* (Tech. Rep. No. AHFD-05-01/MAAD-05-1). Savoy: University of Illinois at Urbana-Champaign, Institute of Aviation, Human Factors Division.

Wickens, C. D., Mavor, A. S., & McGee, J. P. (1997). *Flight to the future: Human factors in air traffic control.* Washington, DC: National Academy Press.

Wickens, C. D., Mavor, A. S., Parasuraman, R., & McGee, J. P. (1998). *The future of air traffic control: Human operators and automation.* Washington, DC: National Academy Press.

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors*, *51*, 446–462.

Ericka Rovira is an assistant professor at the Engineering Psychology Lab in West Point, New York. She received her PhD in applied experimental psychology from the Catholic University of America in Washington, D.C., in 2006.

Raja Parasuraman is a professor of psychology at the Arch Lab at George Mason University in Fairfax, Virginia. He received his PhD in psychology from Aston University in Birmingham, United Kingdom, in 1976.

Date received: August 19, 2009
Date accepted: April 21, 2010