# The Duration of Automation Bias in a Realistic Setting

Robert J. de Boer and Wijnand Heems

Amsterdam University of Applied Sciences

Karel Hurts

CogniTech

## Author's note

Robert J. de Boer, School of Technology, Amsterdam University of Applied Sciences, the Netherlands.

Wijnand Heems is currently at the Faculty of Mechanical Engineering, Eindhoven University of Technology, the Netherlands.

Karel Hurts, CogniTech, Leiden, the Netherlands

Correspondence should be addressed to Robert J. de Boer, Amsterdam University of Applied Sciences, School of Technology, PO Box 1025, 1000 BA Amsterdam, the Netherlands, rj.de.boer@hva.nl

# Abstract

Whereas in most studies conducted previously the effect of automation bias has been investigated in terms of an instantaneous decision, this study is aimed at quantifying its duration. Automation bias is modeled as a stochastic process using a unimodal log-log probability distribution. To validate the model, an experiment using an Airbus A320 fixed base flight simulator with a malfunction on the auto throttle was executed with 35 licensed pilots. The effect of pilot experience is investigated; results show that less experienced pilots are on average less sensitive to automation bias but have more variation in performance than more experienced pilots.

*Keywords*: Automation bias, probability distribution, detection time, aircraft

# The Duration of Automation Bias in a Realistic Setting

## INTRODUCTION

In the last decades the level of automation on the flight deck has increased significantly. This evolution has caused a shift in the dominant role of the flight crew, i.e., from manual operator to system supervisor (Baxter, Besnard, & Riley, 2007). One of the consequences of the increased automation is the risk of what has been called automation bias. This is the phenomenon that users of automation have a tendency to trust and follow the signals of the automated system to the extent that contradictory information available from other sources is ignored or is detected too late (Mosier & Skitka, 1996). This has been visible in aviation where numerous incidents have been attributed to automation bias (Ferris, Sarter, & Wickens, 2010; Mouloua, Hancock, Jones, Vincenzi, & Kantowitz, 2010).

In the case of the crash of Turkish Airlines flight TK 1951 near Amsterdam on February 25th 2009, it was clear that the crew depended on the auto-throttle to maintain velocity and height while they were busy finalizing the checklists and preparing for landing, unaware that a malfunctioning radio altimeter caused the auto-throttle to change to an inappropriate flight mode. Exactly one minute passed between the intercept of the glide scope and the initial activation of the stick shaker. Apparently, this was too little time to recognize the declining air speed, high pitch angle or the out-of-context flight mode annunciation (RETARD, indicating that the engines were in idle) on the primary flight display, despite the addition of a safety pilot to the standard crew. The remaining altitude after initial stick shaker activation was too limited to enable this crew to avoid a crash (Dekker, 2009; Dutch Safety Board, 2010).

Interestingly, while the equipment malfunction in this case tragically ended in a fatal accident, the investigation report mentions other, similar cases in which pilots were able to recover the situation in time to avert a crash. Previous to the Turkish Airlines tragedy, more than fifty instances of faulty auto throttle behavior were documented as a result of problems with the radio altimeter in the B737, in which more serious problems were averted (Dutch Safety Board, 2010). It is probable that in these cases the mode annunciation and the implications for the auto throttle may similarly not have been recognized immediately, but were detected in time (together with the declining air speed and high pitch angle) to prevent a major accident. The Board documented six cases in more detail in which the 'retard flare' mode for the auto-throttle was triggered inappropriately by a faulty radio altimeter. In these cases the situation was recognized and corrected by the crew without further consequences in a time span of between four and 94 seconds. The question arises how much more time would have been required for the crew of Turkish Airlines flight TK 1951 to save the situation. The objective of this study is therefore to investigate the episodic nature of automation bias: how long does the phenomenon of automation bias persist? And which factors influence its duration?

Automation bias has been defined "as a heuristic replacement for vigilant information seeking and processing" (Mosier & Skitka, 1996). Manzey, Reichenbach, and Onnasch (2012) use the term automation bias to describe three distinct but related automation induced phenomena: "(a) withdrawal of attention in terms of incomplete cross-checking of information, (b) active discounting of contradictory system information, and (c) inattentive processing of contradictory information analog to a "looking-but-not-seeing" effect." In contrast to suggestions from earlier research, the third category (c) was found by the authors to be quite significant: 11 out of 18 participants in their study exhibited this behavior, i.e., they had visually accessed all relevant system parameters but had nonetheless missed the contradictory signals. We follow Manzey et al. by adopting the term automation bias and by having it refer to all three phenomena.

In most studies conducted previously the effect of automation bias has been investigated in terms of an instantaneous decision influenced by work load and automation reliability (see for example the review by Wickens &

Dixon, 2007). The predominant focus on instantaneous decisions is justified by Parasuraman and Manzey (2010):

> *"Technically, the performance consequence [of automation bias] could also involve not an omission error but an extremely delayed reaction. However, in many contexts in which there is strong time pressure to respond quickly, as in an air traffic control conflict detection situation a delayed response would be equivalent to a miss."*

However, this stance can be challenged: in many instances automation bias manifests itself only to be resolved within the available time, as was the case in more than fifty reported instances before the Turkish Airlines crash (Dutch Safety Board, 2010). Undoubtedly in many more cases that went unreported, automation bias (i.e. trusting the system instead of vigilantly seeking contradictory information) occurred without severe ramifications. In any retrospective analysis, the time limit set for a manifestation of automation bias is determined by the severity of the consequences, and may therefore be considered to be subject to "hindsight bias" (Woods, Dekker, Johannesen, Cook, & Sarter, 2010) and even arbitrary.

In this study we propose an alternative approach, in which automation bias is modeled as a stochastic process and the probability of the detection of contradictory cues (i.e. the termination of automation bias) increases with time. Other human detection and recognition tasks have been found to be subject to a log-logistic probability function, like the visual detection of small flaws as a function of crack length (Georgio, 2006), for an operator's reaction time to an alarm (so-called time-reliability correlation, Dougherty & Fragola, 1988; Hollnagel, 2009), and to identify a mismatch between stimuli from a software tool and instructions given to them by the researcher (de Boer, 2012). The log-logistic probability function is generally unimodal (i.e. has a single peak). In the case of human detection and recognition tasks the mode (peak) of the distribution is thought to reflect the time required to perceive, recognize and identify the stimulus and includes the neurophysiological reaction time (Hollnagel, 2009). The tail of the distribution reflects a conscious or unconscious choice to ignore the stimulus.

The log-logistic probability function is described by equations (1) and (2):

$$f\ (x,\ \alpha,\ \beta) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left[1+\left(\frac{x}{\alpha}\right)^{\beta}\right]^{2}} \qquad (1)$$

$$F\ (x,\ \alpha,\ \beta) = \frac{x^{\beta}}{\alpha^{\beta}+x^{\beta}} \qquad (2)$$

In these equations: f (x, α, β) is the probability density function; F (x, α, β) is the cumulative distribution function; x is the duration of the contradictory stimuli ($x \in \mathbb{R}^{+}$); α is a scale parameter and is also the median of the distribution (α > 0); and β is a shape parameter. The probability density function is unimodal when the shape parameter β is larger than one. It is expected that automation bias - operationalized as the detection time for contradictory cues - will comply with a unimodal log-logistic probability function.

To investigate the episodic nature of automation bias, an experiment was designed in which the tendency to trust and follow the signals of the automated system and to ignore contradictory information available from other sources were observable, and the time scale for different participants was closely monitored. Pilots' reliance on the Electronic Centralized Aircraft Monitor (ECAM) was studied, a display that under normal circumstances presents every failure to the pilot. An experiment was conducted, in which a major automation failure was simulated that - contrary to pilot expectations - does *not* appear on the lower half of the ECAM. The time delay in detecting the failure was determined for licensed pilots with varying experience. Based on the FAA standard of 23 to 45 seconds for the detection of visible alarms (Pritchett, 2009; Veitengruber, 1978), it was initially expected that the fault would be identified within minutes, as several cues were distinctly visible and should have been part of the standard scanning cycle. A recent similar study, however, showed that the detection time for a discrepancy in Indicated Air Speed was not identified within 2 minutes (Dijk, Merwe, & Zon, 2011), suggesting that a longer period is required to identify contradictory information, particularly concerning systems that have proven to be reliable in the past (Wickens & Dixon, 2007). As automation bias is sensitive to system familiarity (Manzey et al., 2012) we compared novice and experienced pilots in our study. It is expected that experienced pilots will be less meticulous in their scanning behavior, given the highly reliable automated systems in aviation that increase automation bias (cf. Bahner, Elepfandt, & Manzey, 2008).

## METHOD

To investigate the episodic nature of automation bias, an automation malfunction was introduced during a 40 minute flight in a fixed base simulator. Cues about the malfunction were visible for a full 12 minutes. Every effort was made to make the flight as realistic as possible within the possibilities of the simulator, including take-off preparation and checklists, the actual take-off and climb to cruising altitude.

### Participants

Participants were recruited from alumni of the Amsterdam University of Applied Sciences and from our network of active pilots at different Dutch airlines. Thirty-five participants took part in the experimental study. The participants were given a small present as compensation for their participation and reimbursed for their travel costs.

### Experimental design

The experiment followed a randomized design with 1 discrete between-subjects factor consisting of 2 levels (experienced group vs. inexperienced group). All participants were exposed to the same sequence of events with respect to simulation setup, instructions and type and time course of the malfunction.

**Independent variables.** Pilot experience was the independent variable. Two groups were formed consisting of either inexperienced or experienced pilots. The inexperienced group consists of twenty pilots that recently graduated from flight school, but did not have a type rating or airline experience. Sixteen of these were licensed as an Airline Transport Pilot and had been awarded the Multi Crew Cooperation certificate, the others were licensed at the level of Commercial Pilot. The experienced group consists of fifteen airline pilots with between 1000 and 19000 (*M* = 9938; *SD* = 5105) hours flight time. The average age of the participating pilots was 34 (*SD* = 13) years. The average age of the inexperienced group was 24 (*SD* = 4) years, and of the experienced group it was 46 (*SD* = 10) years. The experienced group was mixed in current Boeing or Airbus type rating and/or experience.

**Dependent variable.** The dependent variable was the duration of automation bias, defined as the number of seconds from the onset of the malfunction until detection, as indicated by the pilot's behavior (exclamations and other behavioral changes). Two independent observers identified the moment in time, and video observations were available as a back-up in case of ambiguity. Preliminary experiments showed that the moment of detection is consistently visible, and post-test surveys verified this.

### Equipment

A fixed base Touch Screen Trainer (TST) flight simulator of the Airbus A320 was used for the experiment. The A320 TST is manufactured by Eca Faros (France) in 2006 and is equipped with official Airbus licensed software. The A320 TST consists of seven touch screens, which display the flight deck and is programmed in such a way that it can simulate one or multiple malfunctions simultaneously or consecutively. The simulator does not include an external vision system.

To enable the manipulation use was made of an inconsistency in the software provided to the simulator manufacturer in the original Airbus Data Package to subdue the flagging of an inoperative thrust lever ("ENG ONE TLA FAULT"; more recent versions of the simulator software have been updated and now exclude this possibility).

Additional equipment included a video camera (for post-experiment validation of the detection time) and a non-functioning heart rate monitor that was intended to mask the main intention of the study.

**Manipulation**

The manipulation consisted of a malfunction of the thrust lever for the left engine that was fixed at idle power *without* warning messages appearing on the lower part of the upper Electronic Centralized Aircraft Monitor (ECAM). This particular set of events contradicted with the expectation that all malfunctions in the A320 will appear on the lower part of the ECAM. Five cues were available to signal the presence of the malfunction of the thrust lever, despite the absence of a warning message on the ECAM: (1) a discrepancy between the N1 axis speed indications of the left and right engines, on the upper display of the ECAM; (2) deviations of the exhaust gas temperature (upper ECAM); (3) rotation speed of the secondary axis N2 (upper ECAM); (4) deviation in fuel flow between the two engines (upper ECAM); and (5) rudder deflection indication, which was presented on the middle console next to the rudder trim knob. These cues were in direct sight as long as the malfunction was present and are part of the standard scanning cycle for pilots. Another set of six cues was accessible only if the relevant page on the lower ECAM had been selected: deviations for the two engines in fuel consumed (two separate instances), oil pressure, vibrations, oil quantity, and remaining fuel. The manipulation does not lead to an undesired aircraft state, as the unbalance in thrust and the remaining thrust of both engines combined are within the limits of the flight envelope in this flight phase.

**Procedure**

Participants that were scheduled for the experiment received a short explanation beforehand by email, accompanied by: (1) flight plan, (2) A320 checklist, (3) Standard Instrument Departure chart and (4) Standard Arrival Route chart. Upon arrival at the research facility, a briefing document with experiment instructions and consent letter were handed to and signed by the participant. The stated purpose of the experiment was ”to gain a better understanding of human performance on the flight deck when a person is exposed to varying work load during various flight phases. Indicators of performance will be measured with the use of a heart rate monitor and a video camera”.

Each participant was first given a twenty-minute introduction to the A320 TST. The participant was designated pilot flying (PF), and was supported by a researcher in the role of pilot not flying (PNF). The participant was informed that the roles of PF and PNF in the study were similar to the roles in a grading flight. A second researcher acted as air traffic controller. A familiarization flight which was similar to the flight scenario used for the experiment was performed until the participant indicated that he was sufficiently acquainted with the system. This generally lasted longer for those without previous Airbus experience.

The experimental flight was then initiated, constituting a 40-minute flight from Amsterdam Schiphol Airport to London Heathrow. The flight departs from runway 24 after receiving clearance from Air Traffic Control. After this clearance, the plane takes off and heads due south-west for a VALKO 1S Standard Instrument Departure (SID). After reaching the transition altitude at Flight Level (FL) 60, the flight crew receives clearance to climb to an altitude of FL260 and accelerates to an Indicated Air Speed of 340 knots. At waypoint REFSO, Air Traffic Control orders the flight to turn in the direction of waypoint TRIPO and descend to FL70 earlier than anticipated with a rate of descent of 1.500 feet/minute. At the start of the descent, the malfunction is introduced (i.e. the thrust lever for the left engine remains stuck in idle). Although all the five cues that are described above are immediately accessible and visible, the resulting slip and unbalance in power is somewhat masked by a turn at the top of descent and the throttle reduction. The descent continues for twelve minutes until the aircraft automatically initiates a deceleration at the bottom of descent just before reaching FL100, where a speed limit of 250 knots must be maintained. This automatic deceleration requires both engines to run in idle, thereby effectively eliminating the cues that enables identification of the malfunction and signaling the end of the experimental phase.

After completion of the experimental study, the participants were debriefed in both a written and oral manner. Specific information was gathered about whether, at what time and how the malfunction of the throttle was detected, and corroborated with the registered observations. We checked for evidence that the reaction time was confounded by not recognizing the meaning of the indicators because of inexperience. A survey was filled out by the participants containing several demographic questions and questions about their work load and other experiences during the flight.

## RESULTS

### Manipulation check

At debriefing, all participants who had detected the failure conceded that they had been affected by the malfunction and none of them indicated that they had delayed or suppressed a verbal reaction. In response to the question "How did the realism effect your performance", respondents indicated there had been a limited effect (*M* = 3.2, *SD* = 0.8 on a 5 point Likert scale, 1 = not at all, 5 = totally, N=35) and confirmed this in verbal comments (e.g. "*I regularly checked for visual warnings signs at the ECAM, but there were none. I also heard no warning sound, so I thought everything was functioning ok.*"). It is concluded that the manipulation was successful and sufficiently realistic.

## Duration of automation bias

The time until failure detection was generally longer than expected. Four participants (11%) detected the failure within the norm time of 45 seconds, but 12 out of the 35 participants (34%) did not detect the failure even before the bottom of descent (12 minutes = 720 seconds = test end). The mean time to detect the failure was 240 seconds, and the median was 143 seconds (only those that detected before test end, N = 23).

The cumulative probability of detection as a function of time from onset of the malfunction follows a log-logistic distribution curve described by equations (1) and (2). The best-fitting solution (defined by the least sum for the squares of the difference for each of the data points below 720 seconds) for the cumulative distribution function is generated by $\alpha = 325$ and $\beta = 1.19$; $\chi 2$ (df = 37) = 14.95, 1-p < 0.001, see Figure 1.

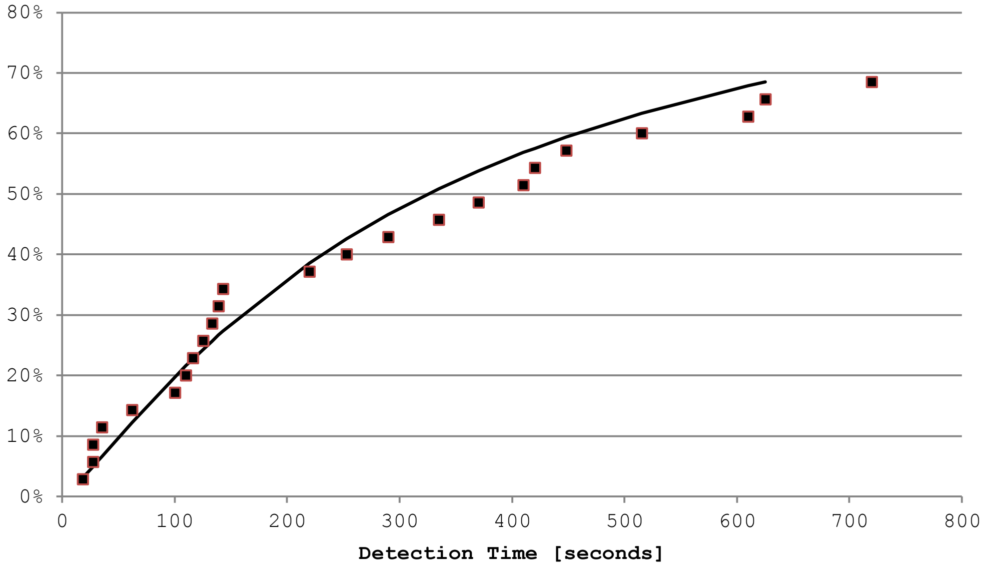

*Figure 1:* Graph showing cumulative probability of detection as a function of detection time. Shown are data points from the experiment (squares, N=23) and the best-fitting solution (line, $\alpha = 325$ and $\beta = 1.19$).

The best fitting solution takes into account a cumulative probability of 66% at 720 seconds due to the 12 participants that did not detect the failure before the bottom of descent. The other possible distributions that were investigated (the normal distribution and the exponential distribution) did not fit the data. The mode of the probability density function occurs at 42 seconds.

**Effect of experience**

The time until failure detection was generally longer for the group of pilots with no service experience. In this group two out of the 20 participants (10%) detected the failure within the norm time of 45 seconds, and ten (50%) did not detect the failure before the bottom of descent. We did not find evidence during the debrief that the reaction time was confounded by these participants not recognizing the meaning of the indicators because of inexperience. The mean time to detect the failure was 364 seconds, and the median was 390 seconds (only those that detected before test end, N = 10). The best fitting solution for the cumulative distribution function for this group is generated by $\alpha = 791$ and $\beta = 1.07$; $\chi2$ (df = 37) = 25.31, 1-p = 0.07. The mode of the probability density function occurs at 35 seconds (compared to 42 seconds for the whole group.

The time until failure detection was generally shorter for the group of more experienced pilots. In this group two participants out of the 15 participants (13%) detected the failure within the norm time of 45 seconds, and two (13%) did not detect the failure before the bottom of descent. The mean time to detect the failure was 146 seconds, and the median was 125 seconds (only those that detected before test end, N = 13). The best fitting solution for the cumulative distribution function for this group is generated by $\alpha = 137$ and $\beta = 2.07$; $\chi2$ (df = 37) = 13.89, 1-p < 0.001. Surprisingly, the mode of the probability density function occurs at 80 seconds, which is later than for the less experienced pilots. The difference between the two groups has been tested through a Wilcoxon rank sum test and was found to be significant: $U_A = 275.5$, $N_A=21$, $N_B=16$, $p < 0.001$ (one-tailed). The probability distributions for experienced and inexperienced participants are compared in Figure 2. Also shown is the probability distribution for all participants. As can be seen, the mode (peak) for the inexperienced pilots is earlier in comparison to the more experienced pilots. The probability distribution of the experienced pilots shows less variation than that of the less experienced pilots, as is evident by lower values in the tail of the distribution.

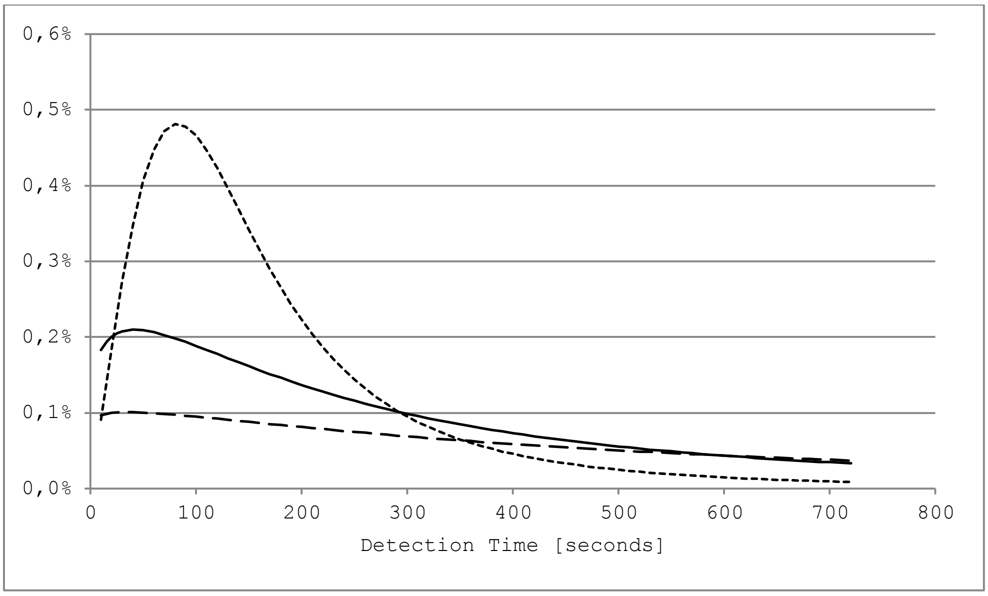

*Figure 2:* Comparison of the probability distributions for all participants (continuous line, mode = 42 seconds, N=23), experienced participants (dotted line, mode = 80 seconds, N=13), and inexperienced participants (dashed line, mode = 35 seconds, N=10).

We checked for an effect of current Boeing or Airbus type rating and/or experience on detection time. A Wilcoxon rank sum test showed that this was not the case: $U_A$ = 28.5, $N_A$=11, $N_B$=5, p = 0.45 (one-tailed).

**Correlation with flight hours**

A further analysis was conducted to investigate whether the effect of experience on the duration of automation bias was also apparent with an increasing number of flight hours. As can be seen in the scatter plot of detection time versus experience (Figure 3), two distinct groups are visible in the data. Therefore the group of experienced pilots was divided into two subgroups with more or less than 15000 flight hours. The time until failure detection was longer for the more experienced group: two out of four of these participants did not detect the failure before the bottom of descent (720 seconds), and the mean time to detect the failure for those that were successful was 350 seconds. In the subgroup with less than 15000 flight hours the mean time to detect the failure was 108 seconds, and all participants detected the malfunction before the bottom of descent. The difference between the two subgroups has been tested through a Wilcoxon rank sum test and was found to be significant: $U_A = 48$, $N_A=12$, $N_B=4$,$p < 0.001$ (one-tailed).

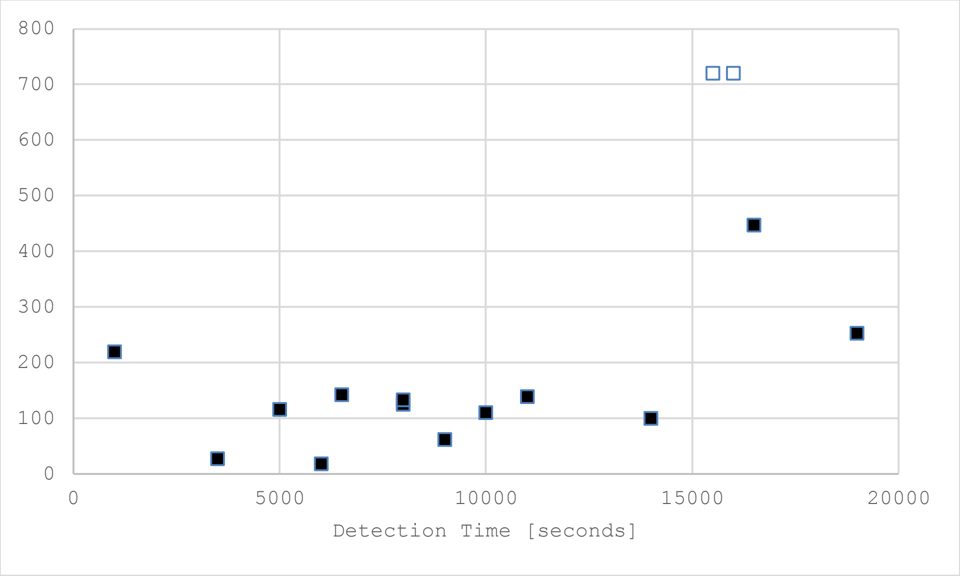

*Figure 3:* Detection time (in seconds) plotted as a function of experience (in Flight Hours, N=15). Two distinct groups are visible in the scatter plot, at below and above 15000 Flight Hours experience. The light squares represent the two participants that did not detect the failure before the bottom of descent (shown at 720 seconds detection time).

## DISCUSSION

In this work it was endeavored to investigate the duration of automation bias in a realistic setting, and to identify a possible relation with pilot experience. A major failure was introduced in the course of a simulator flight from Amsterdam Airport Schiphol to Heathrow *without* warning messages appearing on the lower part of the upper Electronic Centralized Aircraft Monitor (ECAM). This particular set of events contradicted with the expectation that all malfunctions in the A320 will appear on the ECAM, and therefore satisfies the requirements for automation bias in which the users display a tendency to trust and follow the signals of the automated system to the extent that contradictory information available from other sources is ignored or is detected too late (Mosier & Skitka, 1996). Five cues were available to signal the malfunction that were in direct sight and are part of the standard scanning cycle for pilots, but these were out-of-context and unexpected due to the absence of a warning message on the ECAM. 35 pilots with varying degrees of experience participated in the experiment.

The duration of failure detection for licensed pilots varied from 18 seconds to over 720 seconds. Only 11% of the participants met the standard for the detection of visible alarms of 23 to 45 seconds (Veitengruber, 1978) which is still currently being referenced in contemporary guidance material for engineers despite its age (Pritchett, 2009). These results match those of Dijk et al. (2011), who found that two minutes was insufficient to identify discrepancies in the Indicated Air Speed. These results also match incidents in practice, for instance Turkish Airlines flight TK 1951, where more than a minute was required to recognize and alleviate a critical situation after the onset of salient cues (Dutch Safety Board, 2010).

Our results show that a probabilistic approach rather than a rigid limit may be appropriate for the detection of contradictory cues. The authors acknowledge that this is an initial attempt at demonstrating this new approach to understand automation bias and further studies are required. Automation bias was found to follow a unimodal log-logistic probability density curve as a function of stimulus duration. This matches the results of earlier studies of human detection and recognition tasks (de Boer, 2012; Dougherty & Fragola, 1988; Georgio, 2006; Hollnagel, 2009). For the experimental manipulation of our study, the single peak (mode) for the whole group occurred at 42 seconds, which may be considered in line with the standard for the detection of visible alarms mentioned above (23 - 45 seconds, Pritchett, 2009; Veitengruber, 1978). However, the stimulus durations for a cumulative probability of 50% and 95% are 326 seconds and 3815 seconds respectively, implying that often much more time is required for the detection of contradictory cues. Obviously, the parameters of the time delay inherent in the probability distribution ( i.e. α and β in equations (1) and (2)) that were found in this study are probably dependent on the specific circumstances of the experiment such as task context and personal factors (cf. Parasuraman & Manzey, 2010).

We found a significant difference in the sensitivity to automation bias between junior and experienced pilots. We expected that young inexperienced pilots would show more meticulous scanning behavior than more experienced and veteran pilots, who may have become reliant on automation because they had never experienced such failures before (cf. Bahner et al., 2008). In line with this suggestion, the mode of the probability distribution for the less experienced pilots occurred at 35 seconds versus 80 seconds for the pilots with experience. However (based on the best-fitting distributions extrapolated beyond the 720 second cut-off) the stimulus durations for a cumulative probability of 95% are 12225 seconds for the novice pilots and 574 seconds for the experienced group, implying that the experienced group as a whole is less susceptible to automation bias and the novice group demonstrates a large variation in performance. A possible explanation is that less experienced pilots suffered from higher workload, resulting in a greater variation and on average a longer duration of automation bias (Manzey et al., 2012).

The unimodal log-logistic probability density function for automation bias may reflect satisficing behavior or a bounded rationality of the human operator (Simon, 1956). The mode (peak) of the distribution can be considered as the time required to perceive, recognize and identify the stimulus and includes the neurophysiological reaction time (Hollnagel, 2009). It is suggested that for the current study the greatest part of this period (35 – 80 seconds) reflects the duration of the scanning cycle over the instruments. (Note that inexperienced pilots seem to scan faster than their more experienced colleagues.) The tail of the probability distribution can be interpreted as a conscious or unconscious choice to discount information that contradicts the automation, i.e. the period in which the effort of acquiring and processing the additional information is considered to not weigh up against a better quality of the decision making process. Any *conscious* behavior of the pilots to withdraw attention or to actively discount contradictory information is supported by the findings related to the effects of previous experience with an automation aid on the likelihood of automation bias, as well as by those regarding the effects of workload (Manzey et al., 2012; Parasuraman & Manzey, 2010; Wickens & Dixon, 2007). The results of the current study suggest a similar type of bounded rationality in *unconscious* looking-but-not-seeing behavior. Apparently, the expected effort to process salient visual cues is subconsciously assessed in order to identify its relevance for the current task. Stimuli that are deemed task-irrelevant are likely to be missed under high work load, even if they are in the central visual field (the fovea) (Simons & Chabris, 1999). This behavior is functional in a dynamic reality in which a failure may automatically disappear without it having been noticed (e.g., if the operation of the thrust lever is somehow restored). Only after a period of persistent presentation are the cues properly recognized and interpreted.

Of the three types of automation bias described by Manzey, Reichenbach, and Onnasch (2012), (b - active discounting of contradictory system information) is described by the authors (p.59) as a conscious decision “to trust the provisions of the automated aid”; whereas (c - inattentive processing of contradictory information analog to a “looking-but-not-seeing” effect) is entirely subconscious. The authors are less clear as to whether they consider (a - withdrawal of attention in terms of incomplete cross-checking of information) a conscious choice or not. Under the circumstances of the current study none of the participants alluded to having consciously discounted the cues regarding the thrust lever failure. Furthermore, from the (video) observations it seemed that participants used conventional scanning techniques. This finding is in line with the high proportion of looking-but-not-seeing behavior (more than 60%) found in the experiment by Manzey et al. (2012).

The debriefing indicated that the current experiment was sufficiently realistic to give credibility to the results. The suppression of the ECAM message despite the malfunction was realistic (although extremely unlikely) because it was included in an initial version of the actual Airbus data package (Heems, Speet, & Stam, 2012). The experiment included take-off preparation and checklists, radio communication and flight plan changes. The masking of the cues by a turn at the top of descent, lack of undesired aircraft state and lack of support from the pilot not flying to detect the failure were realistic and intentional. In contrast to real life, the detection of the contradictory cues was made more difficult by the absence of motor sounds, the absence of an external view to detect slip, and the lack of motion.

## APPLICATION

This study represents an initial, proof-of-concept step to exploring automation bias and is hoped to be a good basis for further research to build upon. Its benefit to industry is threefold. Firstly, current standard for the detection of visible alarms of 23 to 45 seconds (Veitengruber, 1978) which is still currently being referenced in contemporary guidance material (Pritchett, 2009) seems to be too short in comparison to the duration of failure detection for licensed pilots. Avionics designers are wise to take a longer detection time into account. Secondly, in assessing detection failures (for instance in incident investigations), it is recommended to take into account a continuous function for the probability of detection, and avoiding a binary judgment for the detection of critical stimuli in hindsight (“too late”, “in time”). Finally, the current study suggests that in complex socio-technical systems like an aircraft error prevention strategies are doomed to fail – this study has shown that faultless behavior is nigh on impossible. Rather, error management, focusing on increasing the positive and decreasing the negative consequences of errors (Dimitrova, 2014) seems more fruitful within the context of cockpit crew training to prepare pilots to cope with surprise, ambiguity, and potentially conflicting information (Rankin, Woltjer, Field, & Woods, 2013).

## CONCLUSION

In the current study the duration of automation bias in a somewhat realistic setting has been investigated. Our results suggest that the duration to detect contrary or contradictory cues as is often necessary to break automation bias is much longer than currently specified in the literature. This can be understood as functional behavior under realistic, dynamic circumstances. Experience helps somewhat to reduce the variation in automation bias. The duration of automation bias can be modeled as a unimodal log-logistic probability density function.

## ACKNOWLEDGMENTS

We would like to thank Robert Stam, Alex Speet, Laurens Baardman and Jop Dingemans for their invaluable effort and support during the experimental phase of this study. We would also like to thank Moniek Gabel-Sombekke for her assistance in identifying potential participants, and those participating for their time and effort. We are grateful to two anonymous reviewers for their valuable comments.

## References

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008). Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1330–1334. doi:10.1177/154193120805201906

Baxter, G., Besnard, D., & Riley, D. (2007). Cognitive mismatches in the cockpit: Will they ever be a thing of the past? *Applied Ergonomics*, *38*(4), 417–423.

De Boer, R. J. (2012). *Seneca's Error: An Affective Model of Cognitive Resistance . Industrial Design Engineering*. Delft University of Technology, Delft.

Dekker, S. (2009). *Report of the flight crew human factors investigation conducted for the Dutch safety board into the accident of TK1951, Boeing 737-800 near Amsterdam Schiphol Airport, February 25, 2009* (p. 127). Lund: Lund University, School of Aviation.

Dijk, H. van, Merwe, K. van de, & Zon, R. (2011). A Coherent Impression of the Pilots ' Situation Awareness : Studying Relevant Human Factors Tools. *The International Journal of Aviation Psychology*, *21*(4), 343–356.

Dimitrova, N. G. (2014). *Rethinking errors: How error-handling Strategy Affects our Thoughts and Others' Thoughts about us*. Amsterdam: Vrije Universiteit.

Dougherty, E. M. J., & Fragola, J. R. (1988). Foundations for a time reliability correlation system to quantify human reliability. In *Conference Record for 1988 IEEE Fourth Conference on Human Factors and Power Plants* (pp. 268–278). IEEE.

Dutch Safety Board. (2010). *Crashed during approach, Boeing 737-800, near Amsterdam Schiphol airport, 25 February 2009*. the Hague, the Netherlands.

Ferris, T., Sarter, N., & Wickens, C. D. (2010). Cockpit automation: Still struggling to catch up. *Human Factors in Aviation*.

Georgio, G. A. (2006). *Probability of Detection (PoD) curves: Derivation, applications and limitations*. Health and Safety Executive, UK.

Heems, W., Speet, A., & Stam, R. (2012). *Automation Surprise, Mismatch between human-automation properties and capabilities*. Amsterdam, the Netherlands.

Hollnagel, E. (2009). *The ETTO principle: efficiency-thoroughness trade-off: why things that go right sometimes go wrong*. Ashgate Pub Co.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids : The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. doi:10.1177/1555343411433844.

Mosier, K., & Skitka, L. (1996). Human decision makers and automated decision aids: Made for each other. In Parasuraman R. & Mouloua M. (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Mahwah, NJ: Lawrence Erlbaum.

Mouloua, M., Hancock, P. A., Jones, L., Vincenzi, D., & Kantowitz, B. H. (2010). Automation in Aviation Systems: Issues and Considerations. In J. A. Wise, V. D. Hopkin, & D. J. Garland (Eds.), *Handbook of aviation human factors* (2nd ed.). Boca Raton: CRC.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. doi:10.1177/0018720810376055

Pritchett, A. R. (2009). Aviation Automation: General Perspectives and Specific Guidance for the Design of Modes and Alerts. *Reviews of Human Factors and Ergonomics*, *5*(1), 82–113. doi:10.1518/155723409X448026

Rankin, A., Woltjer, R., Field, J., & Woods, D. (2013). "Staying ahead of the aircraft" and Managing Surprise in Modern Airliners. In *Proceedings of the 5th Resilience Engineering Symposium*. Soesterberg, the Netherlands.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst : sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059–1074. doi:10.1068/p2952

Veitengruber, J. E. (1978). Design Criteria for Aircraft Warning, Caution, and Advisory Alerting Systems. *Journal of Aircraft*, *15*(9), 574–581. doi:10.2514/3.58409

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. doi:10.1080/14639220500370105

Woods, D. D., Dekker, S., Johannesen, L. J., Cook, R. I., & Sarter, N. (2010). *Behind human error* (2nd ed.). Ashgate.