Applied Ergonomics 118 (2024) 104288

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Operator selection for human-automation teaming: The role of manual task skill in predicting automation failure intervention[☆]

Natalie Griffiths[a], Vanessa Bowden[a,*], Serena Wee[a], Luke Strickland[b], Shayne Loft[a]

[a] The University of Western Australia, Australia
[b] Curtin University, Australia

ARTICLE INFO

Handling Editor: Dr. P.R. Pennathur

Keywords:
Human-automation teaming
Automation failure intervention
Air traffic control

ABSTRACT

Humans working in modern work systems are increasingly required to supervise task automation. We examined whether manual aircraft conflict detection skill predicted participants' ability to respond to conflict detection automation failures in simulated air traffic control. In a conflict discrimination task (to assess manual skill), participants determined whether pairs of aircraft were in conflict or not by judging their relative-arrival time at common intersection points. Then in a simulated air traffic control task, participants supervised automation which either partially or fully detected and resolved conflicts on their behalf. Automation supervision required participants to detect when automation may have failed and effectively intervene. When automation failed, participants who had better manual conflict detection skill were faster and more accurate to intervene. However, a substantial proportion of variance in failure intervention was not explained by manual conflict detection skill, potentially reflecting that future research should consider other cognitive skills underlying automation supervision.

1. Introduction

Automation is changing how humans manage work by assisting task completion (Bhaskara et al., 2021; Rieth and Hagemann, 2022; Vagia et al., 2016). Many modern work systems provide automated assistance – from surgeons receiving assistance during medical procedures, to pilots largely flying aircraft through inputs to automated flight deck systems. High-risk operations such as those in mining and oil and gas are also increasingly automated (Loughney and Wang, 2018). Automation can make work safer and more efficient when implemented appropriately because it can typically perform pre-determined steps more efficiently and with less variability than humans (Endsley, 2017). Automated systems function effectively in routine situations but can potentially be erroneous in situations outside those anticipated by developers. Under these circumstances, automation can become 'brittle' and risks failure (Woods and Cook, 2012).

Humans are therefore required to supervise automated systems in case they need to intervene to resolve failures of automation. Automation supervision refers to operators noticing that automation may have failed, deciding that it has, and then intervening effectively. This is assumed to primarily involve the cognitive skills that would be required to manually perform the automated task. However, the sub-process of monitoring for automation failure differs from manual performance because operators are largely removed from active situational control, while still being required to remain vigilant. Unfortunately, humans are "magnificently disqualified for this particular form of sustained attentive response" (Hancock, 2013, p.98). Supervising automation is challenging because it becomes increasingly difficult to remain vigilant over time (*vigilance decrement*; Mackworth, 1948; Warm et al., 2008). Although certain work procedures can offset the impact of this (e.g., Crew Resource Management in piloting; Flin et al., 2002), with increasing automation, there is an increased potential for operators to notice fewer automation failures as a function of time on task, compared to performing the task manually (Molloy and Parasuraman, 1996). This issue is compounded when automation failures are rare (Bowden et al., 2023; Taleb, 2005). Examining human-automation teaming is critical given that automation supervision is an increasing job requirement in modern work environments, such as in military surveillance, airport baggage inspection, medical screening, remote (uninhabited) vehicle management, and air traffic control (ATC).

[☆] This research was supported by an Australian Research Council (ARC) Discovery Grant (DP160100575) awarded to Loft, an ARC Future Fellowship (FT190100812) awarded to Loft, and an ARC Discovery Early Career Researcher Award (DE230100171) awarded to Strickland.

[*] Corresponding author. School of Psychological Science, The University of Western Australia, Crawley, Western Australia, 6009, Australia.
E-mail address: vanessa.bowden@uwa.edu.au (V. Bowden).

https://doi.org/10.1016/j.apergo.2024.104288
Received 5 November 2023; Received in revised form 25 February 2024; Accepted 9 April 2024

To adapt to the increasing requirements for human-automation teaming in modern work, organisations may need to re-evaluate their personnel assessment, selection, and training processes. An assumption likely underlying current organisational processes is that manual task skill (ability to perform a task manually/without automation) and automation supervision performance when automation completes the same task should be relatively highly related. For instance, we might assume that our ability to drive a car relates to our ability to detect and correct the erroneous performance of an autonomous vehicle. However, to our knowledge, this assumption has not yet been empirically examined.

Practically, it is important to know the extent to which manual task skill predicts the quality of automation supervision, and thus the extent to which organisations need to consider selecting operators based on other cognitive traits and skills related to superior automation supervision. Continuing our driving example, given the finite resources likely to be available to train operators of future autonomous vehicles, there will be a trade-off between training operators to effectively monitor automation for failure and training them to drive manually. Organisations will likely be tempted to spend more resources on what they believe (either correctly or incorrectly) predicts automation supervision as this will be the operator's task the majority of the time, at the expense of training for manual skills, which may be perceived as only being required in the rare instances when automation fails. However, such an approach could backfire if manual task skill itself is predictive of automation supervision performance.

Manual task skill should certainly be related to performance *following* detection of an automation failure, when returning to manual task performance (return to manual; RTM). In many settings, human intervention to an automation failure is the last line of defence before disaster. For instance, the pilots of Air France flight 447 were required to RTM after a blocked sensor caused the autopilot to turn off mid-flight. A major contributing factor in the accident (crash) was that the two pilots who were sharing first officer responsibilities when the autopilot switched off were inexperienced in manually handling the aircraft (BEA, 2012). Thus, although pilots spend most of their time supervising highly reliable autopilot systems, their manual flying skills are likely critical after automation failures are detected. However, the initial detection of automation failure, a necessary condition for successful automation failure intervention, may depend on cognitive processes additional to manual task skill (e.g., vigilance performance, prospective memory, etc.).

The current study examined the extent to which manual task skill predicted automation supervision in a medium-fidelity ATC task. ATC is a complex, safety-critical, work environment where automation is being increasingly used to handle traffic load and reduce operator workload (e.g., Airservices Australia, 2018; FAA, 2020), making it a prototypical task to address our research question. Controllers complete multiple tasks, including monitoring and communicating with aircraft entering/exiting their sector, scheduling take-offs/landings, and diverting aircraft around weather. However, their key responsibility is to ensure minimum separation between aircraft. While automation can assist, it is ultimately the operators' responsibility to ensure that automated conflict avoidance systems perform appropriately.

In the current study, aircraft were in conflict if they would simultaneously violate 5 nautical miles (nm) lateral and 1000 ft vertical separation standards. When aircraft are cruising (i.e., not ascending/descending), manual conflict detection requires individuals to assess the relative position and speed of converging aircraft pairs that share a common altitude, in order to judge their relative-arrival time at common intersection (Loft et al., 2009; Neal and Kwantes, 2009; Rantanen and Nunes, 2005; Vukovic et al., 2013; 2014). We asssesed manual conflict detection skill using a conflict discrimination task (CDT). Participants were presented successive pairs of converging aircraft (cruising at the same altitude) with varying future minimum lateral separation based on their relative position and speed. Participants were required to make a lateral-relative-arrival-time judgement to determine if each presented aircraft pair would conflict or not at their closest point of approach.

Participants' performance on the CDT was then used to predict their automation supervision performance in a medium-fidelity ATC task. In the ATC task, participants were responsible for handling a sector with multiple aircraft that required acceptance and handoff from the sector, as well as maintaining minimum separation standards between multiple aircraft with the assistance of either high or low degree automation. Degree of automation (DOA) is a metric that ranks the level of automated assistance, where higher DOA indicates more automated support (Parasuraman et al., 2000; Sheridan and Verplank, 1978). In both the high and low DOA conditions, automation reliably detected 24 conflicts (80%), and participants were required to intervene to six automation failures. Automation supervision performance was operationalized as the accuracy and speed of intervening to prevent the aircraft conflicts that automation missed (failure intervention).

In the low DOA condition, automation highlighted potential conflicts (converging aircraft at the same altitude) and participants were required to manually determine if highlighted aircraft pairs would conflict. In the high DOA condition, the automation detected conflicts and intervened without participant input. To effectively supervise low or high degree conflict detection automation, participants needed to make relative-arrival time judgements on converging aircraft pairs flying at the same altitude that they deemed automation may have not highlighted (low DOA) or failed to detect/resolve (high DOA).

In summary, the current study examined the extent to which manual CDT task performance (i.e., the ability to discriminate conflicts from non-conflicts) predicted automation supervision (i.e., more accurate and/or faster failure intervention) when either low or high degree conflict detection automation was provided. In a recent meta-analysis by Onnasch et al. (2014), automation supervision (i.e., failure intervention) was shown to be better when using low compared to high DOA across various task domains. Due to this difference in automation supervision across DOAs, it was also of interest whether CDT skill and DOA would interact such that CDT skill would be a stronger predictor of automation supervision under low compared to high DOA.

2. Methods

2.1. Participants

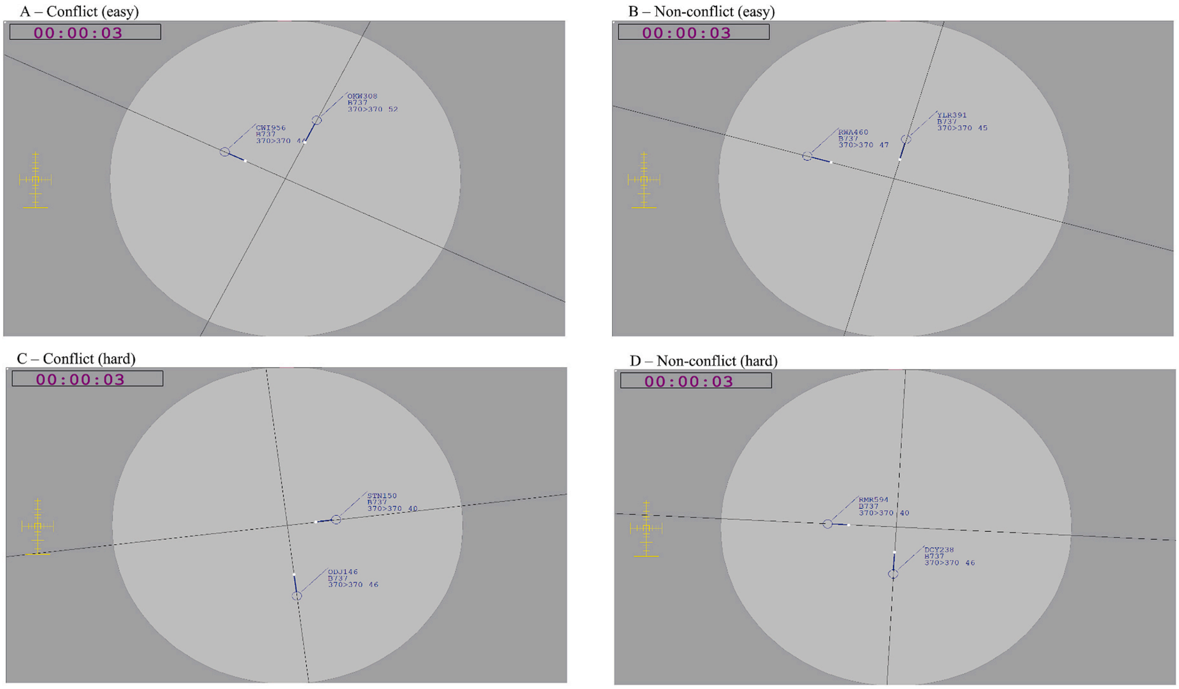
Undergraduate students from The University of Western Australia ($N = 204$) participated in exchange for course credit or AUD$40 (60% female, $M_{\text{age}} = 21.4$ yrs; $SD_{\text{age}} = 7.1$, range = 17–63) and an additional performance-based incentive of AU$5-$20. This research complied with the American Psychological Association Code of Ethics and was approved by The University of Western Australia Human Research Ethics Office.

2.2. Equipment and tasks

2.2.1. Manual conflict discrimination task

CDT trials were presented as a dark grey screen with a centred light grey circle representing an air traffic sector. Two black lines ran across the display from edge to edge, representing flight paths. These flight paths randomly rotated around the screen for each trial, but always bisected at a 90-degree angle. When a trial started, participants were shown two aircraft, one on each flight path. Aircraft were represented by a circle icon with an attached projection line indicating where that aircraft would be in 1 min. Each aircraft icon had an attached data block containing the aircraft callsign (e.g., PZV599), type (e.g., B737), current/cleared altitude (e.g., 370 > 370 indicates flying at 37,000 ft, cleared to fly at 37,000 ft), and speed (e.g., 43 indicates 430kn). Examples of CDT trials are presented in Fig. 1.

Participants were presented with aircraft pairs with systematically varied minimum separation distances, and were asked to determine if


A – Conflict (easy)

B – Non-conflict (easy)

C – Conflict (hard)

D – Non-conflict (hard)

Fig. 1. CDT trials. Blue circles with attached data blocks represent aircraft approaching common intersections along the black flight paths. White dots show projected aircraft location in 1 min. A countdown timer was in the top left corner (i.e., 3s left). The yellow scale on the left-hand side of the trial represented 5NM and the y-axis represented 10 nM. Panel A presents an easy conflict trial (1.5 nm DOMS). Panel B presents an easy non-conflict trial (9.2 nm DOMS). Panel C presents a more difficult conflict trial (4.65 nm DOMS). Panel D presents a more difficult non-conflict trial (5.7 nm DOMS). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

each aircraft pair were in conflict or not. All aircraft pairs were cruising (at the same altitude) so only future lateral separation required assessment. On each trial, participants watched a pair of aircraft move along their respective flight paths. After 3.5s, the screen proceeded from the aircraft display to a subsequent message asking participants to: "Please indicate whether these aircraft are in conflict. Press ***f*** for conflict or ***j*** for non-conflict" (response keys counterbalanced). To determine if the pair would conflict, participants were required to judge if aircraft would violate lateral separation standards of 5 nm in the future based on their current relative position and speed. If participants responded incorrectly, a feedback screen advising them of the error was displayed. Participants then saw a grey screen with an instruction to press the spacebar when they were ready to proceed to the next trial.

We used the Method of Constant Stimuli (MOCS), a procedure commonly used in psychophysics for determining sensory thresholds (Gescheider, 1985). Trial difficulty varied through the degree of lateral minimum separation (DOMS) between aircraft, that is, the lateral distance between aircraft at their closest point of passing. DOMS ranged from 0.45 to 9.55 nm in increments of 0.35 nm (26 DOMS in total). Aircraft with a DOMS less than 4.9 nm would conflict, while those with a DOMS more than 5.1 nm would not. At more extreme DOMS values (i.e., further from 5 nm), it was easier to discriminate conflicts from non-conflicts. For example, at the smallest DOMS it was relatively easy to discern that aircraft would conflict, as aircraft in these trials would significantly overlap at their closest point of approach (Fig. 1A). At the largest DOMS, it was relatively obvious that aircraft would not conflict, as aircraft in these trials would have the largest distances between them at their closest point of approach (Fig. 1B). More difficult trials were those with DOMS closest to 5 nm, at which distance it was more ambiguous whether the aircraft would conflict (Fig. 1C and D). Participants completed 20 trials for each DOMS (520 trials in total, with 260 conflict and 260 non-conflict trials).

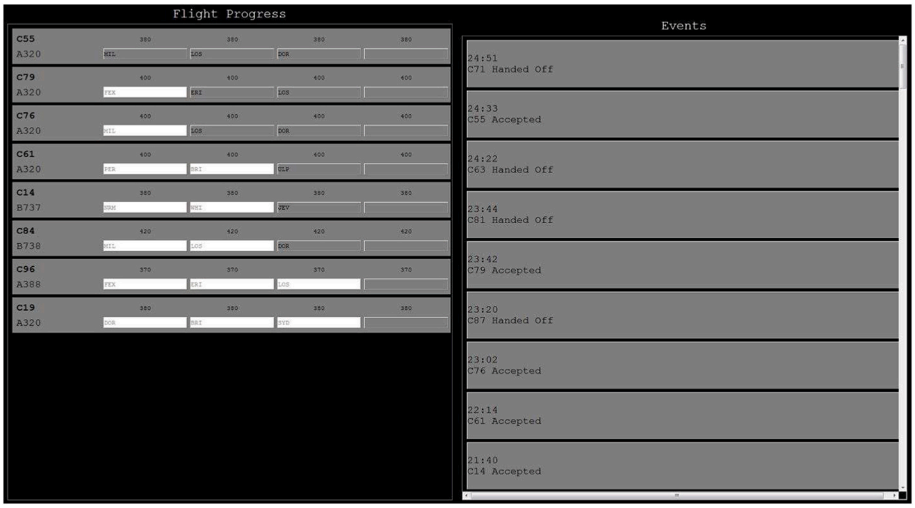
2.2.2. *Air traffic control task*

The medium fidelity ATC task (Fothergill et al., 2009) was presented on two 22-inch monitors and participants used a computer keyboard and mouse. While the ATC display and aircraft interaction features contained realistic elements, participant tasks represented a simplification of the real work of an air traffic controller. For example, participants made no radio communications and had limited aircraft controls available to them, such as not being able to re-route aircraft or change their speed. There were also a relatively low number of aircraft under participant control and minimal variability in air traffic volume.

The right-hand monitor (Fig. 2) contained flight progress strips and an event log, designed based on Masalonis et al. (1997). Flight strips contained the callsign, altitude, and route information for each aircraft. The event log displayed actions performed by the participant or the automation. It was updated as participants accepted and handed off aircraft, or when either participants or the high DOA resolved conflicts. Each event in the log described the involved aircraft, action taken, and time.

The left-hand monitor contained a sector map with an inner controlled sector (light grey) surrounded by an uncontrolled airspace (Fig. 3). Aircraft entered the map and travelled unidirectionally along flight paths before exiting. Aircraft were represented by circle icons, with a data block and an attached projection line, indicating where aircraft would be in 20s. Aircraft remained at the same speed and altitude, unless instructed to ascend to avoid a conflict by either the participant or the high DOA. Participants controlled a median of eight aircraft at any one time.

Participants accepted aircraft into and handed-off aircraft out of the controlled sector. Aircraft flashed blue when 20s from entering the controlled sector to indicate the need for acceptance. Participants accepted aircraft by pressing the 'A' key and clicking on the aircraft. Accepted aircraft turned green to indicate they were under participant control. Aircraft flashed orange for hand-off when 20s away from exiting the sector. Participants handed-off aircraft by pressing the 'H' key and


Flight Progress

C55 A320	380 MIL	380 LOS	380 DOR	380
C79 A320	400 FEX	400 ERI	400 LOS	400
C76 A320	400 MIL	400 LOS	400 DOR	400
C61 A320	400 PER	400 BRI	400 ULP	400
C14 B737	380 NRM	380 WMI	380 JEV	380
C84 B738	420 MIL	420 LOS	420 DOR	420
C96 A388	370 FEX	370 ERI	370 LOS	370
C19 A320	380 DOR	380 BRI	380 SYD	380

Events

24:51
C71 Handed Off

24:33
C55 Accepted

24:22
C63 Handed Off

23:44
C81 Handed Off

23:42
C79 Accepted

23:20
C87 Handed Off

23:02
C76 Accepted

22:14
C61 Accepted

21:40
C14 Accepted

Fig. 2. Flight strips (left) and event log (right) presented on the right-hand monitor. The flight strips displayed aircraft callsigns, altitude, and route information. Waypoints along the flightpath were displayed on the strips and changed from grey to white as aircraft flew past them. The event log displayed actions as they were performed by the participant or the automation, as well as the time and involved aircraft.

clicking on the aircraft. Handed-off aircraft turned black to indicate that they were no longer under participant control. Participants were notified of missed acceptances and hand-offs by an auditory alert.

Minimum separation standards remained the same as in the CDT. To detect conflicts, participants projected the future lateral separation of aircraft cruising at the same altitude and on intersecting flightpaths to judge whether lateral separation would be violated. To intervene, participants clicked on both relevant aircraft and confirmed their selection via a pop-up dialogue box. If the pair were in conflict, one aircraft would begin to ascend 1000 ft to avoid the conflict and a notification would be added to the event log. If a conflict was missed and minimum separation was lost, then the aircraft involved turned yellow until separation was re-established (and they returned to green). If participants attempted to intervene to aircraft that were not in conflict, the aircraft did not change altitude. Auditory alerts informed participants of conflict misses and false alarms. Participants were presented 30 conflicts in total (10 per scenario) and 18 'near-misses', where pairs flew at the same altitude but were ~10s away from being classified as a conflict. Intervening to near-misses was classified as a near-miss false-alarm. Participants detected potential aircraft conflicts with assistance from high or low DOA (a between-subjects manipulation), as described below.

2.2.3. *Low degree of automation*

Low DOA (Fig. 3, upper) highlighted (either purple or red) any pair of aircraft flying at the same flight level and on converging flight paths, which resulted in highlighting both conflicts and near-misses. Participants were thus instructed that highlighting did not guarantee a conflict and that they were to decide the conflict status of the highlighted pair. Participants were also required to intervene to prevent conflicts missed (i.e., not highlighted) by automation. The automation failed to detect (highlight) six conflicts, two per scenario. Participants were thus required to manually intervene to resolve the conflicts that the automation missed (i.e., failed to highlight).

2.2.4. *High degree of automation*

High DOA (Fig. 3, lower) detected and resolved conflicts upon entry of the second aircraft in the pair to the controlled sector. Participants were notified of automated conflict resolution actions with a time-stamped event in the event log detailing the aircraft involved in the conflict and the action taken (Fig. 2). The high DOA failed to resolve the same six conflicts as the low DOA failed to highlight. Participants were thus required to intervene to resolve the conflicts that the automation missed (i.e., failed to change altitude).

2.3. *Operator state measures*

Participants completed in-task questions regarding their perceived workload, fatigue, trust in automation, and task engagement every 3-min. Additionally, participants completed multi-item measures of the constructs after completing each scenario. These measures were designed to answer auxiliary research questions and are not further examined here.

2.4. *Procedure*

Participants were briefed on the experiment and provided consent, before being instructed on the CDT. They then completed 40 CDT practice trials (approximately 5-min). They then completed 520 CDT experimental trials (approximately 55-min). Participants then had the opportunity for a 5 to 10-min break.

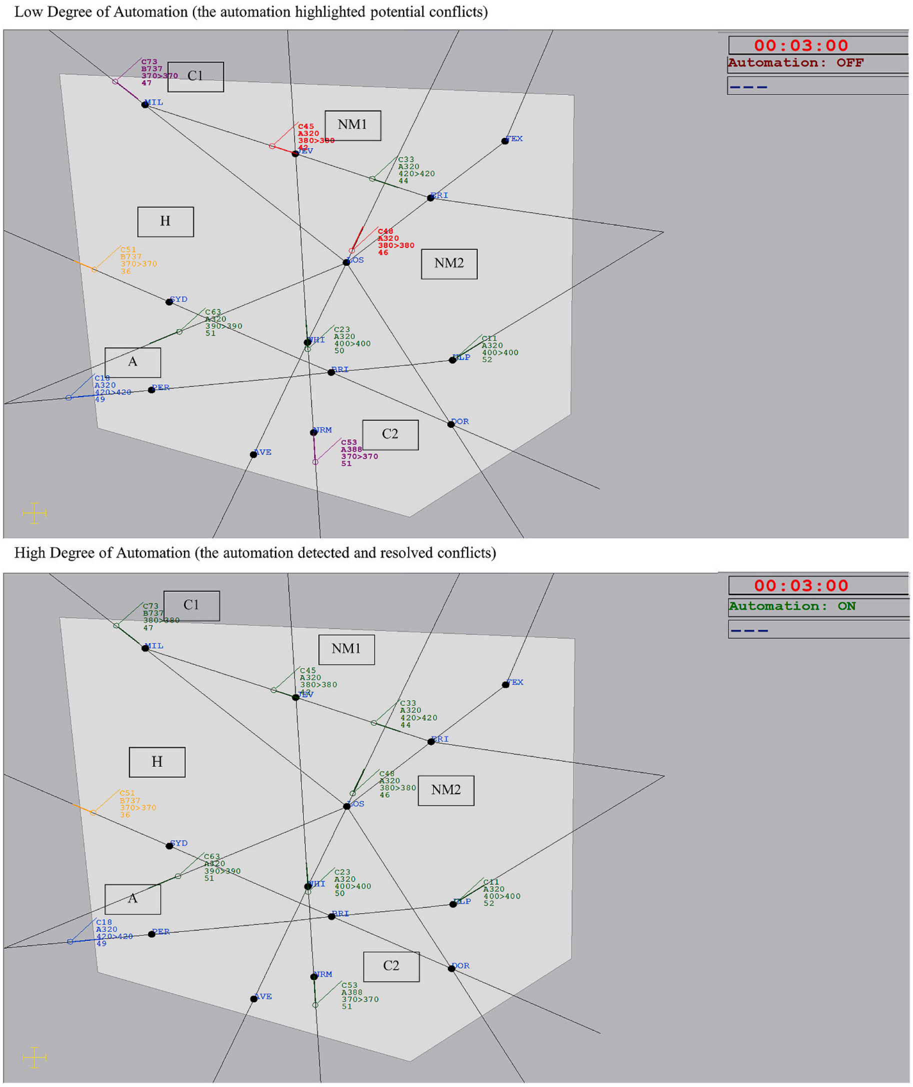
For the ATC simulation task, participants first completed a 25-min audio-visual training on the ATC tasks. They then completed a 30-min manual practice scenario. While participants were provided with conflict detection automation in all the experimental scenarios, the practice ATC scenario required manual conflict detection to ensure participants in the low DOA condition could make conflict status decisions on highlighted aircraft in the experimental scenarios, and to ensure participants in both DOA conditions could intervene to automation failures in the experimental scenarios. Following this practice scenario, participants completed a 5-min audio-visual training introducing either high or low DOA (condition specific), before they completed three 30-min experimental ATC scenarios with either high or low DOA automation. Participants answered questionnaires at the end of each scenario and were given the chance to have a short break (2-5-min) between scenarios.

2.5. *Data analyses*

2.5.1. *CDT performance metrics*

Psychometric functions relate physical stimuli to perception/sensory thresholds by presenting stimuli of different magnitudes and asking participants to rate the stimuli in a two-alternative forced-choice format (e.g., conflict/non-conflict).

We examined participant response probabilities, specifically, the probability of responding "conflict" at each DOMS, by fitting a Cumulative Gaussian distribution to each participant's data. Fig. 4 shows an example of this function fit to the average of all participant data.


Low Degree of Automation (the automation highlighted potential conflicts)

High Degree of Automation (the automation detected and resolved conflicts)

Fig. 3. Sector map of the airspace (labels added to figure for clarity, not presented in-task). Aircraft requiring acceptance (A) and hand-off (H). The box at the top right displays the scenario run-time (red when paused, black when running), and indicates whether the conflict detection automation was active. The top panel shows the low DOA highlighting aircraft pairs at the same altitude. The bottom panel shows the same example with high DOA, where aircraft were not highlighted and conflicts were automatically resolved. The low DOA (upper panel) highlighted all aircraft flying on intersecting flight paths at the same flight level, which includes both conflicts (C1 & C2), and near-misses (NM1 & NM2). The high DOA (lower panel) automatically resolved conflicts (C1 & C2) by allocating one aircraft in the conflict pair to a new altitude (C1 has ascended to 38,000 feet). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Fitting the function to each participant's data generated a function for each participant given by the equation:

$$p(conflict\ response) = 0.5 \times \left(1 + \text{erf}\left(\frac{\mu - \text{DOMS}}{\sqrt{\sigma}}\right)\right) \tag{1}$$

In this equation, *erf* refers to the error function. Two parameters were extracted from fitting this function. Sigma (σ) reflects the slope of the function, with larger values indicating a shallower slope (Fig. 5A). Lower sigma scores reflect participants who were better at discriminating smaller DOMS differences when determining aircraft conflict status. Sigma was used as our main CDT performance metric and predictor of automation failure intervention in the simulated ATC task. For multiple regression analyses, CDT sigma was mean-centred around the sample average.

For completeness, we also analysed CDT mu (μ). Mu reflects the DOMS at which a participant was equally likely to respond 'conflict' or 'non-conflict' (*p(conflict)* = 0.5). Participants with mu > 5 were more likely to respond 'conflict' to presented aircraft pairs, while those with mu < 5 were more likely to respond 'non-conflict' (Fig. 5B). Changes in mu reflect a shift of the function along the x-axis without a change in slope. We first determined whether direction of mu mattered (e.g., conflict versus non-conflict bias), and found no effect of bias direction on failure intervention accuracy or response time outcomes. We therefore converted mu to *response bias*, by taking the absolute value centred at 5. Thus, smaller response bias values represent participants who were well calibrated to the 50:50 conflict/non-conflict CDT task base presentation rate. In contrast, larger response bias values indicate either a tendency to classify aircraft pairs as in conflict, or a tendency to classify aircraft as not in conflict. For multiple regression analyses, CDT bias was also mean-centred.

2.5.2. *Automation supervision: failure intervention*

Failure intervention accuracy and response time were recorded. We defined failure intervention hit rate as the proportion of conflicts correctly detected by the participant when the automation failed. Failure intervention false alarm rate was the proportion of the near-miss

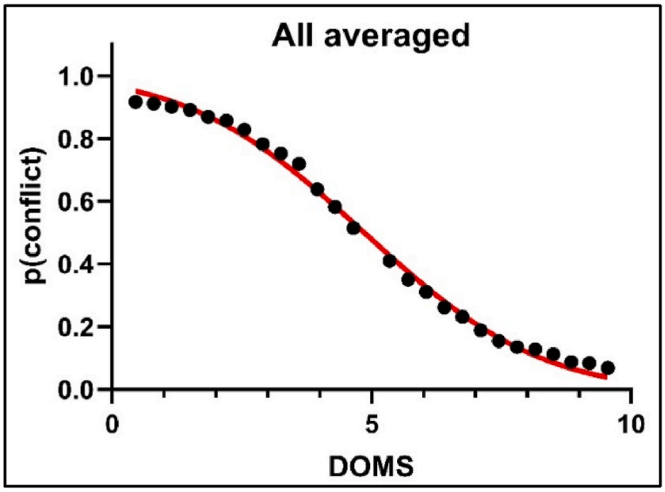


Fig. 4. A Cumulative Gaussian function (red line) fit to the average of all participant data. Black circles indicate the average probability of all participants at each DOMS. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

aircraft pairs that participants incorrectly identified as conflicting. Overall failure intervention accuracy was calculated by subtracting the failure intervention false alarm rate from the failure intervention hit rate (hit – false alarm rate). This quantified the degree to which participants could discriminate between instances of automation successfully detecting a conflict from instances of automation missing a conflict. For completeness, the hit and false alarm data were also analysed separately. Failure intervention response time was the time taken to correctly resolve conflicts missed by automation.

3. Results

Nine participants' data was excluded due to extremely poor CDT task performance resulting in a poor fit of the Gaussian function to their data ($R^2 < 0.80$). Two additional participants were excluded for not intervening to any automation failure (11 participants excluded in total, 93% of the sample retained). The Gaussian function fit the remaining data adequately (low DOA: $\overline{x} = 0.93$, $sd = 0.05$; high DOA: $\overline{x} = 0.94$, $sd = 0.03$). Descriptive statistics are provided in Table 1. Welch's t-tests were run to account for unequal sample sizes between conditions following data cleaning, and effect sizes are given using Cohen's d, with weak, moderate, and strong effect sizes 0.10, 0.30, and 0.50, respectively (Cohen, 1992).

The hit-false alarm rate data indicated that participants were significantly more accurate in their decision to intervene when automation failed in the high compared to the low DOA condition, $\overline{X}_{diff} = 0.14$, $t(184.74) = 3.44$, $p < 0.001$; $d = 0.49$. Failure intervention hit rate did not differ significantly between high and low DOA conditions, $\overline{X}_{diff} = 0.04$, $t(187.92) = 1.32$, $p = 0.19$. Rather, participants made more failure intervention false alarms in the low compared to high DOA condition, $\overline{X}_{diff} = 0.10$, $t(187.20) = 3.81$, $p < 0.001$; $d = 0.55$. However, participants were faster to correctly intervene when provided low compared to high DOA, $\overline{X}_{diff} = 10.6\text{s}$, $t(190) = 3.81$, $p < 0.01$; $d = 0.40$.

To determine whether CDT skill was related to automation failure intervention, we ran multiple linear regression analyses with CDT sigma and CDT response bias as predictors regressed on the dependent variables of failure intervention hit-false alarm rate, failure intervention hit rate, failure intervention false alarm rate, and failure intervention response time. The interaction between CDT skill and DOA condition was also assessed to determine if manual CDT skill was more closely related to automation failure intervention in the low compared to high DOA condition. Where the effect of CDT was significant, we conducted follow-up bivariate regressions to determine the strength of the relationship between CDT skill and the dependent variable of interest. Correlations are presented in Table 2.

3.1. CDT performance: sigma

Lower sigma scores reflected participants better at discriminating aircraft conflict status during the CDT task. High DOA was coded as 0 and low DOA was coded as 1, with negative beta weights representing the change in the dependent variable when moving from high DOA to low. The main automation failure intervention variable of interest was the hit-false alarm rate. CDT sigma, DOA, and their interaction accounted for 23.7% of the variance in the failure intervention hit-false alarm rate, $F(3,189) = 19.59$, $p < 0.001$. Participants who could better discriminate aircraft conflict status in the CDT task were more accurate

Table 1
Descriptive statistics for manual (conflict discrimination task; CDT) and failure intervention by condition (low and high DOA).

Variables	Low DOA			High DOA		
	Mean	SD	Range	Mean	SD	Range
Hit-False Alarm Rate	0.51	0.31	0.33–1.00	0.65	0.25	0.06–1.00
Hit Rate	0.78	0.21	0.17–1.00	0.82	0.18	0.17–1.00
False Alarm Rate	0.28	0.20	0.00–0.94	0.17	0.16	0.00–0.72
RT (s)	90.4	27.5	16.7–168.9	101.0	25.3	29.3–151.9
CDT Sigma	2.56	0.92	1.48–7.93	2.51	0.78	1.18–4.97
CDT Response Bias	0.42	0.31	0.002–1.30	0.48	0.48	0.02–2.58

Note. RT = Response Time. CDT = Conflict Discrimination Task.

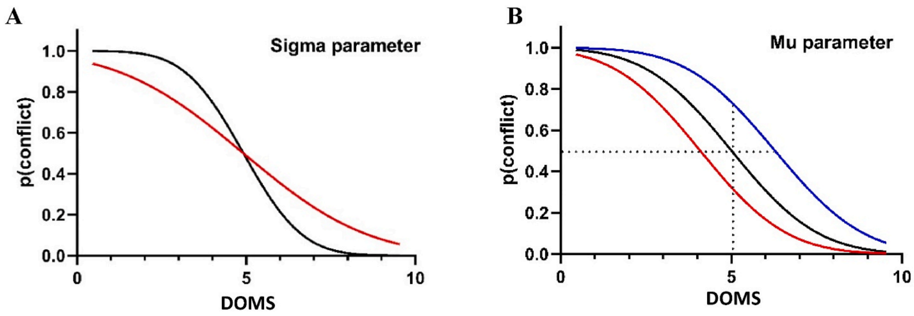


Fig. 5. A) The sigma parameter. The black line represents a participant who was better able to discriminate conflicts from non-conflicts (smaller sigma = steeper slope), compared to the red line who was poorer at discrimination (larger sigma = shallower slope). B) The mu parameter. The black line represents an unbiased participant (i.e. neither conflict nor non-conflict biased when responding). The blue line represents a conflict-biased participant (i.e. more like to consider aircraft pairs to be conflicts), while the red line represents a non-conflict-biased participant (i.e. considers aircraft pairs to be more likely non-conflicts). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Correlation matrix for manual (manual conflict discrimination task; CDT) and automated task performance measures by condition (low and high DOA).

	Low DOA						HIGH DOA					
	1.	2.	3.	4.	5.	6.	1.	2.	3.	4.	5.	6.
1. Hit-False Alarm Rate	–						–					
2. Hit Rate	**0.77**	–					**0.74**	–				
3. False Alarm Rate	**−0.73**	−0.13	–				**−0.68**	−0.06	–			
4. RT	**−0.40**	**−0.27**	−0.08	–			−0.10	**−0.27**	0.15	–		
5. CDT Sigma	**−0.47**	**−0.37**	**0.34**	0.11	–		**−0.37**	**−0.23**	**0.31**	0.10	–	
6. CDT Response Bias	0.05	−0.01	>-0.001	0.04	0.0007	–	−0.21	**0.34**	−0.23	0.10	**0.54**	-

Note. RT = Response Time. CDT = Conflict Discrimination Task. Significant values are bolded (two-tailed, α = 0.05).

in their decision to intervene when automation failed ($B = -0.12$, $SE = 0.03$, $p < 0.001$). DOA also predicted failure intervention hit-false alarm rate such that participants in the high DOA condition were more accurate in their decision to intervene when automation failed ($B = -0.13$, $SE = 0.04$, $p < 0.001$). There was no interaction between CDT sigma and DOA, ($B = -0.04$, $SE = 0.04$, $p = 0.38$). The bivariate regression indicated that CDT sigma accounted for 18.2% of the variance in the failure intervention hit-false alarm rate $F(1,191) = 42.51$, $p < 0.001$.

CDT sigma, DOA, and their interaction accounted for 10.8% of the variance in the failure intervention hit rate, $F(3,189) = 7.65$, $p < 0.001$. Participants who were better at discriminating conflicts from non-conflicts in the CDT task had a higher failure intervention hit rate ($B = -0.05$, $SE = 0.02$, $p < 0.05$). DOA did not predict failure intervention hit rate ($B = -0.03$, $SE = 0.03$, $p = 0.21$). There was no interaction between CDT sigma and DOA, ($B = -0.03$, $SE = 0.03$, $p = 0.33$). The bivariate regression indicated that CDT sigma accounted for 9.65% of the variance in the failure intervention hit rate $F(1,191) = 20.39$, $p < 0.001$.

CDT sigma, DOA, and their interaction accounted for 17.0% of the variance in the failure intervention false alarm rate, $F(3,189) = 12.80$, $p < 0.001$. Participants who were better at discriminating conflicts from non-conflicts in the CDT made fewer failure intervention false alarms, ($B = 0.07$, $SE = 0.02$, $p < 0.01$). Participants in the low DOA condition made more failure intervention false alarms than those in the high DOA condition ($B = 0.10$, $SE = 0.03$, $p < 0.001$). There was no interaction between CDT sigma and DOA, ($B = 0.007$, $SE = 0.03$, $p = 0.83$). The bivariate regression indicated that CDT sigma accounted for 10.3% of the variance in the failure intervention false alarm rate, $F(1,191) = 21.95$, $p < 0.001$.

CDT sigma, DOA, and their interaction accounted for 4.94% of the variance in failure intervention response time, $F(3,189) = 3.27$, $p < 0.05$. CDT sigma did not predict failure intervention response time ($B = 3.24$, $SE = 3.46$, $p = 0.35$). Participants in the high DOA condition were slower to respond to automation failures ($B = -10.80$, $SE = 3.81$, $p < 0.01$. There was no interaction between CDT sigma and DOA, ($B = -0.08$, $SE = 4.54$, $p = 0.99$).

3.2. *CDT performance: response bias*

Lower response bias indicated a participant better calibrated to the 50:50 conflict/non-conflict CDT task base presentation rate. CDT response bias, DOA, and their interaction accounted for 8.00% of the variance in the failure intervention hit-false alarm rate, $F(3,189) = 5.31$, $p < 0.01$. CDT response bias significantly predicted failure intervention hit-false alarm rate, such that more biased participants had lower failure intervention accuracy ($B = -0.12$, $SE = 0.06$, $p < 0.05$). DOA significantly predicted the failure intervention hit-false alarm rate, such that low DOA was associated with lower failure intervention accuracy ($B = -0.14$, $SE = 0.04$ $p < 0.05$). There was no interaction between CDT response bias and DOA, ($B = 0.12$, $SE = 0.11$, $p = 0.28$). The bivariate regression indicated that CDT response bias only accounted for 0.97% of the variance in the failure intervention hit-false alarm rate, $F(1,191) = 1.87$, $p = 0.17$.

CDT response bias, DOA, and their interaction did not significantly predict failure intervention hit rate, $R^2 = 0.02$, $F(3,189) = 1.33$, $p = 0.26$.

CDT response bias, DOA, and their interaction accounted for 8.30% of the variance in the failure intervention false alarm rate, $F(3,189) = 5.70$, $p < 0.001$. DOA significantly predicted the failure intervention false alarm rate, such that low DOA was associated with more failure intervention false alarms ($B = 0.10$, $SE = 0.03$, $p < 0.001$). Neither CDT response bias ($B = 0.06$, $SE = 0.04$, $p = 0.13$), nor the interaction between CDT response bias and DOA ($B = -0.03$, $SE = 0.07$, $p = 0.71$) significantly predicted the failure intervention false alarm rate.

CDT response bias, DOA, and their interaction accounted for 6.82% of the variance in failure intervention response time, $F(3,189) = 4.61$, $p < 0.01$. More biased CDT participants were slower to correctly intervene, ($B = 13.55$, $SE = 5.62$, $p < 0.05$). Participants in the high DOA condition were slower to intervene ($B = -10.27$, $SE = 3.79$, $p < 0.01$). There was no interaction between CDT response bias and DOA, ($B = -16.00$, $SE = 10.42$, $p = 0.13$). The bivariate regression indicated that CDT response bias accounted for 2.18% of the variance in failure intervention response time, $F(1,191) = 4.26$, $p < 0.05$.

4. Discussion

In modern work environments such as ATC, automation is essential for efficiency and safety. However, the human supervision of automation provides a critical last line of defence between automation failure and complete system failure. Our aim was to examine the extent to which manual task skill predicted automation supervision. Participants who were better able to discriminate conflicts from non-conflicts on the CDT were more accurate in their decision to intervene when automation failed in the ATC task, indicating superior automation supervision. That is, those participants with better manual conflict detection skill had a higher automation failure intervention hit-false alarm rate. This observed relationship likely reflects the fact that the cognitive process involved in making aircraft relative-arrival judgments (Loft et al., 2009; Neal and Kwantes, 2009; Rantanen and Nunes, 2005) were required for both the CDT task and for the effective supervision of conflict detection automation. It is likely that individuals estimated the relative arrival-times of certain selectively attended aircraft pairs in order to detect whether automation may have failed to highlight an aircraft pair (low DOA) or had failed to intervene (high DOA). This provides empirical support for the assumption that automation supervision can be reliant on the cognitive skills required to manually perform tasks currently being automated.

We additionally measured CDT response bias, which reflects whether participants were biased toward classifying aircraft pairs as conflict or non-conflicts, with less bias indicating better calibration to CDT conflict/non-conflict base presentation rate (i.e., 50:50). More biased participants were found to be less accurate in their decision to intervene when automation failed in the ATC task, but this effect was not significant when examining hit rate or false alarm rates separately. A positive association between CDT response bias and failure intervention response time indicated that participants who were more biased were slower to intervene to automation failures in the ATC task.

In line with previous research and the Onnasch et al. (2014) meta-analysis, participants were faster to correctly intervene to automation failures when assisted by low compared to high DOA. However, in contrast to Onnasch et al. participants were more accurate in their decision to intervene when automation failed in the high compared to the low DOA condition. That is, DOA significantly predicted the failure intervention hit-false alarm rate, and this was driven by an increased false alarm rate when using low DOA. However, we found no effect of DOA on the association between CDT skill and automation supervision performance. This finding suggests that manual conflict detection skill similarly predicted automation supervision across DOA. Future research should examine whether manual skill can predict automation supervision across other DOAs and/or other whether there are other potential moderators such as automation design features or environmental conditions (Karpinsky et al., 2018).

Although manual conflict detection skill was associated with more accurate decisions to intervene when automation failed in the ATC task, substantial variability in failure intervention remained unaccounted for. Future research should also examine other potential underlying cognitive mechanisms that may predict automation supervision. In particular, supervising conflict detection automation also requires individuals to visually search for aircraft pairs that share a common altitude and are heading towards a common intersection (Galster et al., 2001; Gronlund et al., 1998; Remington et al., 2000), whereas the CDT task, for the purpose of experimental control, presented aircraft pairs in isolation and thus had no such visual search requirement. Additionally, given competing demands on attention in simulated ATC (e.g., aircraft acceptance and hand-off requirements), effective automation supervision may have required individuals to defer task actions (e.g., defer the checking of the relative-arrival time of a particular aircraft pair) and thus requiring them to remember to come back and complete that action (i.e., prospective memory; Loft, 2014; Loft et al., 2019). Lastly, performance on vigilance tasks (requiring sustained attention to detect and respond to a stimuli) might predict automation supervision in simulated ATC (Helton and Wen, 2023).

As noted earlier, the current study is limited in that it used a medium-fidelity ATC task (e.g., no radio communications, limited aircraft controls, relatively low aircraft volume). In addition, as we tested novice participants, future research should examine the applicability of the current outcomes to expert controllers who are undoubtedly more experienced, competent, and motivated to detect automation failures (Jamieson and Skraaning, 2020). This extension would also require the use of a higher fidelity ATC task. Related to this, we operationalized automation supervision performance as the degree to which participants could discriminate between instances of automation successfully detecting a conflict from instances of automation missing a conflict, whereas in real ATC missed conflicts are far more concerning than making false alarms (see Loft et al., 2009). These points notwithstanding, the current study presents an initial examination of the relationship between manual task performance and automation failure detection using a medium fidelity simulation relatively representative of current ATC automation designs (International Civil Aviation Organization: Asia and Pacific Office, 2022; Skybrary, 2024).

In conclusion, the current findings indicate a significant degree of overlap in the cognitive processes (i.e., relative-arrival time judgment) underlying manual task skill and automation supervision within the context of this medium-fidelity ATC simulation. Should these findings be replicated in higher fidelity studies using experts, this would suggest that practically, manual task skill is potentially relevant for organisations when selecting future operators to work in automated work environments. However, as a reasonably large amount of variance in automation failure intervention was left unexplained, future research should also examine other cognitive traits and skills potentially related to automation supervision such as visual search performance, attentional management, and prospective memory.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Airservices Australia, 2018. OneSKY Australia. Airservices Australia. https://www.airservicesaustralia.com/wp-content/uploads/OneSKY-Brochure-Detailed-Overviewp-web.pdf.

BEA, 2012. Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro—Paris (Accident Investigation F-GZCP). Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf.

Bhaskara, A., Duong, L., Brooks, J., Li, R., McInerney, R., Skinner, M., et al., 2021. Effect of automation transparency in the management of multiple unmanned vehicles. Appl. Ergon. 90, 103243.

Bowden, V.K., Griffiths, N., Strickland, L., Loft, S., 2023. Detecting a single automation failure: the impact of expected (but not experienced) automation reliability. Hum. Factors 65 (4), 533–545. https://doi.org/10.1177/00187208211037188.

Cohen, J., 1992. A power primer. Psychol. Bull. 112, 155–159.

Endsley, M.R., 2017. From here to autonomy. Hum. Factors 59 (1), 5–27. https://doi.org/10.1177/0018720816681350.

FAA, 2020. NextGen Annual Report (Pp. 1–155). U.S Department of Transportation; Federal Aviation Administration. https://www.faa.gov/sites/faa.gov/files/2022-06/NextGenAnnualReport-FiscalYear2020.pdf.

Flin, R., O'Connor, P., Mearns, K., 2002. Crew resource management: improving team work in high reliability industries. Team Perform. Manag.: Int. J. 8 (3/4), 68–78. https://doi.org/10.1108/13527590210433366.

Fothergill, S., Loft, S., Neal, A., 2009. ATC-lab Advanced: an air traffic control simulator with realism and control. Behav. Res. Methods 41 (1), 118–127. https://doi.org/10.3758/BRM.41.1.118.

Galster, S.M., Duley, J.A., Masalonis, A.J., Parasuraman, R., 2001. Air traffic controller performance and workload under mature free flight: conflict detection and resolution of aircraft self-separation. Int. J. Aviat. Psychol. 11 (1), 71–93. https://doi.org/10.1207/S15327108IJAP1101_5.

Gescheider, G.A., 1985. Psychophysics: Method, Theory, and Application. Psychology Press.

Gronlund, S.D., Ohrt, D.D., Dougherty, M.R.P., Perry, J.L., Manning, C.A., 1998. Role of memory in air traffic control. J. Exp. Psychol. Appl. 4, 263–280.

Hancock, P.A., 2013. In search of vigilance: the problem of iatrogenically created psychological phenomena. Am. Psy. Ass. 68 (2), 97–109. https://doi.org/10.1037/a0030214.

Helton, W.S., Wen, J., 2023. Will the real resource theory please stand up! Vigilance is a renewable resource and should be modeled as such. Exp. Brain Res. 241 (5), 1263–1270. https://doi.org/10.1007/s00221-023-06604-x.

International Civil Aviation Organization: Asia and Pacific Office, 2022. Air traffic management automation system implementation and operations guidance document. https://www.icao.int/APAC/Documents/edocs/cns/ATMASImplementationAndOperationsGuidance-v1.0.pdf.

Jamieson, G.A., Skraaning, G., 2020. The absence of degree of automation trade-offs in complex work settings. Hum. Factors 62, 516–529. https://doi.org/10.1177/0018720819842709.

Karpinsky, N.D., Chancey, E.T., Palmer, D.B., Yamani, Y., 2018. Automation trust and attention allocation in multitasking workspace automation trust and attention allocation in multitasking workspace. Appl. Ergon. 70, 194–201. https://doi.org/10.1016/j.apergo.2018.03.008 (March).

Loft, S., 2014. Applying psychological science to examine prospective memory in simulated air traffic control. Curr. Dir. Psychol. Sci. 23 (5), 326–331. https://doi.org/10.1177/0963721414545214.

Loft, S., Bolland, S., Humphreys, M.S., Neal, A., 2009. A theory and model of conflict detection in air traffic control: incorporating environmental constraints. J. Exp. Psychol. Appl. 15 (2), 106–124. https://doi.org/10.1037/a0016118.

Loft, S., Dismukes, K., Grundgeiger, T., 2019. Prospective memory in safety-critical work contexts. In: Rummel, J., McDaniel, M. (Eds.), Current Issues in Memory, pp. 170–185.

Loughney, S., Wang, J., 2018. Bayesian network modelling of an offshore electrical generation system for applications within an asset integrity case for normally unattended offshore installations. Proc. IME M J. Eng. Marit. Environ. 232 (4), 402–420. https://doi.org/10.1177/1475090217704787.

Mackworth, N.H., 1948. The breakdown of vigilance during prolonged visual search. Q. J. Exp. Psychol. 1 (1), 6–21. https://doi.org/10.1080/17470214808416738.

Masalonis, A.J., Le, M.A., Klinge, J.C., Galster, S.M., Duley, J.A., Hancock, P.A., Hilburn, B.G., Parasuraman, R., 1997. Air traffic control workstation mock- up for free flight experimentation: lab development and capabilities. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 41 https://doi.org/10.1177/1071181397041002166, 1379–1379.

Molloy, R., Parasuraman, R., 1996. Monitoring an automated system for a single failure: vigilance and task complexity effects. Hum. Factors 38 (2), 311–322. https://doi.org/10.1518/001872096779048093.

Neal, A., Kwantes, P.J., 2009. An evidence accumulation model for conflict detection performance in a simulated air traffic control task. Hum. Factors 51 (2), 164–180. https://doi.org/10.1177/0018720809335071.

Onnasch, L., Wickens, C.D., Li, H., Manzey, D., 2014. Human performance consequences of stages and levels of automation: an integrated meta-analysis. Hum. Factors 56 (3), 476–488. https://doi.org/10.1177/0018720813501549.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Syst. Hum. 30 (3), 286–297. https://doi.org/10.1109/3468.844354.

Rantanen, E.M., Nunes, A., 2005. Hierarchical conflict detection in air traffic control. Int. J. Aviat. Psychol. 15 (4), 339–362. https://doi.org/10.1207/s15327108ijap1504_3.

Rieth, M., Hagemann, V., 2022. Automation as an equal team player for humans?-A view into the field and implications for research and practice. Appl. Ergon. 98, 103552.

Remington, R.W., Johnston, J.C., Ruthruff, E., Gold, M., Romera, M., 2000. Visual search in complex displays: factors affecting conflict detection by air traffic controllers. Hum. Factors 42 (3), 349–366. https://doi.org/10.1518/001872000779698105.

Sheridan, T.B., Verplank, W.L., 1978. Human and Computer Control of Undersea Teleoperators. Massachusetts Institute of Technology.

SKYbrary, 2024. Medium term conflict detection (MTCD). SKYbrary Aviation Safety. https://skybrary.aero/articles/medium-term-conflict-detection-mtcd.

Taleb, N.N., 2005. The Black Swan: Why Don't We Learn that We Don't Learn?, vol. 1145. Random House.

Vagia, M., Transeth, A.A., Fjerdingen, S.A., 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Appl. Ergon. 53, 190–202.

Vuckovic, A., Kwantes, P.J., Humphreys, M., Neal, A., 2014. A sequential sampling account of response bias and speed-accuracy tradeoffs in a conflict detection task. J. Exp. Psychol. Appl. 20, 55–68.

Vuckovic, A., Kwantes, P.J., Neal, A., 2013. Adaptive decision making in a dynamic environment: a test of a sequential sampling model of relative judgment. J. Exp. Psychol. Appl. 19, 266–284.

Warm, J.S., Parasuraman, R., Matthews, G., 2008. Vigilance requires hard mental work and is stressful. Hum. Factors 50 (3), 433–441. https://doi.org/10.1518/001872008x312152.

Woods, D., Cook, R.I., 2012. Incidents – Markers of Resilience or Brittleness?, pp. 69–76.