Accepted 5/November/2022, *Journal of Experimental Psychology: Applied*

Automated decision aids: When are they advisors and when do they take control of human decision making?

Luke Strickland[1], Russell J. Boag[2], Andrew Heathcote[3,4], Vanessa Bowden[2] & Shayne Loft[2]

[1] The Future of Work Institute,

Curtin University, Australia

[2] The School of Psychological Science,

The University of Western Australia, Australia

[3]The School of Psychology, University of Newcastle, Australia

[4]Department of Psychology, University of Amsterdam

Address for Correspondence

Luke Strickland,

Future of Work Institute,

Curtin University,

78 Murray Street,

6000 Perth, Australia

Email: luke.strickland@curtin.edu.au

Author Note

Task simulation materials are available on request. The data and analysis code associated with the manuscript are available at: https://osf.io/q3b2r/

Acknowledgements

This research was supported by an ARC discovery grant, DP200101842, awarded to Loft and Strickland.

Declarations of interest: none

Abstract

We applied a computational model to examine the extent to which participants used an automated decision aid as an advisor, as compared to a more autonomous trigger of responding, at varying levels of decision aid reliability. In an air traffic control conflict detection task we found higher accuracy when the decision aid was correct, and more errors when the decision aid was incorrect, as compared to a manual condition (no decision aid). Responses that were correct despite incorrect automated advice were slower than matched manual responses. Decision aids set at lower reliability (75%) had smaller effects on choices and response times, and were subjectively trusted less, than decision aids set at higher reliability (95%). We fitted an evidence accumulation model to choices and response times to measure how information processing was affected by decision aid inputs. Participants primarily treated low-reliability decision aids as an advisor rather than directly accumulating evidence based on its advice. Participants directly accumulated evidence based upon the advice of high-reliability decision aids, consistent with granting decision aids more autonomous influence over decisions. Individual differences in the level of direct accumulation correlated with subjective trust, suggesting a cognitive mechanism by which trust impacts human decisions.

Keywords: human-automation teaming; automation reliability; trust in automation; cognitive control; evidence accumulation model

Public Significance Statement

In the modern workplace, humans can receive imperfect decision advice from automated systems. A computational cognitive model fit to behavioral data from a simulated air traffic control conflict detection task indicated that participants used a low-reliability decision aid as an advisor, but allowed a high-reliability aid more autonomous influence over their decisions.

In safety-critical industries such as defence, aviation, and healthcare, people increasingly work with automated decision aids. For example, decision aids can advise operators how to manage the deployment of multiple uninhabited vehicles for activities such as military operations, delivering cargo and rescuing civilians (Calhoun et al., 2018; Loft et al., 2021), and in air traffic control (ATC[1]), decision aids can advise controllers of potential future conflicts (i.e., loss of minimum separation standards) between aircraft (Musialek et al., 2010; Noskievič & Kraus, 2017). Providing operators with decision aids can improve performance and reduce workload (see meta-analysis by Onnasch et al., 2014). However, decision aids can be imperfect for a variety of reasons, including hardware issues, software bugs, and automation "brittleness" (Billings, 1997) wherein the decision aid performs incorrectly in an unforeseen or untested context. Thus, operators often need to scrutinize advice from decision aids to decide whether to accept or reject it. This introduces risks of automation misuse (reliance on incorrect automated advice) and disuse (rejection of correct automated advice) (Lee & See, 2004). For future systems designers to anticipate and prevent such errors, research into the cognitive processes underlying how humans incorporate advice from decision aids into their decisions is essential.

Automation reliability (accuracy) is a crucial factor in determining human reliance on automation and subsequent human-automation teaming[2] (HAT) performance outcomes, both when human operators passively supervise automation that is performing tasks for them (e.g.,

[1] Abbreviations used in this article: HAT (human-automation teaming), ATC (air traffic control), RT (response time).

[2] In this article, we adopt a broad definition of human-automation teaming, referring to any situation where a human performs a task with the assistance of automation. However, it should be note that some literature adopts a narrower definition with more specific requirements, such as bidirectional communication between the human and the automation (e.g., Lyons et al., 2021).

Ferraro et al., 2018), and when operators perform tasks themselves whilst provided with decision aids that advise actions (e.g., Hussein et al., 2020a; Rovira et al., 2007; Shah & Bliss, 2017; Wiegmann et al., 2001). On the one hand, low-reliability automation is often undesirable because it can cause human operators to drop below manual (no automation) levels of accuracy (Wickens & Dixon, 2007). On the other, humans tend to over-trust (Lee & See, 2004) and misuse high-reliability automation (Barg-Walkow & Rogers, 2016; Parasuraman & Riley, 1997). Consequently, if high-reliability automation does fail to function as expected, humans are relatively less likely to detect it (Bailey & Scerbo, 2007), which can contribute to serious accidents (e.g., NTSB, 2014).

When a human decision maker works alongside an automated decision aid, some system designers assume the human will independently verify decision aid advice. However, the degree to which people verify decision aid advice will almost certainly be affected by the reliability of automation (i.e., the extent to which it advises correct decisions; Lee & See, 2004; Parasuraman & Manzey, 2010) and resulting human trust in the automation (Lee & See, 2004; Merritt & Ilgen, 2008; Muir, 1994; Muir & Moray, 1996). Lee and See (2004) proposed a simple definition of trust as: “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 54), which is the most widely used definition in the HAT literature. When a decision aid is not reliable it is important that it not be the sole source of evidence that triggers a decision, as this would encourage automation misuse. Instead, it is desirable that choices can only be triggered when the human has assessed the available evidence, even though the decision aid might still provide input into which choice is made. This enables humans to better detect cases where the decision aid is incorrect, while still making proper use of it in other cases, minimizing misuse and disuse. In contrast, when decision aids are highly reliable people may allow them to trigger choices in a relatively autonomous manner, freeing up attention for non-automated

concurrent tasks in the wider context of the work system (Bagheri & Jamieson, 2004; Moray, 2003; Moray & Inagaki, 2000). Although this could be viewed in many operational contexts as an adaptive strategy (Kaber, 2018), it also increases the risk of automation misuse.

Recently, a quantitative cognitive model has been proposed which has the potential to identify whether humans use decision aids as automated advisors or whether humans allow them to more autonomously trigger their decisions (Strickland et al., 2021). To date, the model has only been applied to circumstances where levels of human manual performance and automation reliability were approximately equal, and in that case indicated that decision aid input influenced decision making in a way that would not autonomously trigger decisions. In the current study, we expand on this to manipulate decision aid reliability by including a relatively high-reliability condition (better than average participant manual performance) and a low-reliability condition (poorer than average participant manual performance). Our aim was to examine whether decision aid reliability affected the balance of human strategies employed, and particularly the extent to which humans allowed high-reliability decision aids to autonomously contribute to their decision making.

An Evidence Accumulation Model of Automation Use

Cognitive models are important tools for understanding HAT because they quantify complex relationships between observed behaviour and latent cognitive processes. Notably, signal detection theory has been applied to assess how automated advice influences decision making accuracy (e.g., Bartlett & McCarley, 2017; 2021). This approach provides a benchmark for optimality and can be used to distinguish between predictions of a broad range of potential human strategies for incorporating automated advice into decisions.

Although signal detection theory is useful for understanding HAT, it does not incorporate RT and hence cannot identify cognitive processes related to RT, such as speed-accuracy tradeoffs (e.g., Balakrishnan et al., 2002). Evidence accumulation models are a class

of cognitive models that predict human choices and response times (RTs) by assuming that decision making involves accumulating evidence until reaching a threshold. They have the potential to provide novel insights about HAT because they account for accuracy and RT distributions in a unified way, and hence can identify cognitive processes underlying performance. They have been shown to accurately describe choices and distributions of RTs in a range of simple tasks (Ratcliff et al., 2016). Recently, they were shown to be applicable to complex tasks such as ATC conflict detection (Boag, Strickland, Heathcote, et al., 2019) and maritime surveillance (Palada et al., 2016). They can also describe how humans adapt to time pressure and secondary demands in such complex tasks, and how they allocate attention according to task priorities (Boag, Strickland, Loft, et al., 2019).

Strickland et al. (2021) applied an evidence accumulation model to investigate how humans use automated decision aids in an ATC conflict detection task, which requires deciding whether pairs of aircraft will violate minimum separation standards in the future. Figure 1 depicts their model. At the start of a conflict detection trial, some initial amount of evidence is drawn from a distribution of start points. Evidence then accumulates towards each possible decision at a speed determined by the accumulation rate. For a two-choice task, there are two types of accumulation rate: match and mismatch. The match accumulation rate refers to the accumulation rate matching the correct decision (e.g., ‘conflict’ accumulation when aircraft are in conflict), and mismatch the rate for incorrect decisions (e.g., ‘nonconflict’ accumulation when aircraft are in conflict). Accumulation continues until an accumulator reaches threshold. The first accumulator to reach threshold determines which choice the model predicts will be made.

Strickland et al. (2021) proposed an evidence accumulation framework for understanding how task (stimuli) information and decision aid advice are integrated to make decisions (Figure 1). Evidence accumulation rates are driven by two sources of input – task

inputs and automation (decision aid) inputs. Task inputs are derived from characteristics of the presented stimuli. In the case of conflict detection for aircraft at the same cruising altitude and approaching a common intersection, this input would reflect the relative position and speed of aircraft. With this information, participants could mentally project the future lateral positions of each aircraft over time (relative arrival time judgment; Law et al., 1993; Loft et al., 2007) to assess future conflict status. This could be accomplished with a range of cognitive/perceptual strategies (Xu & Rantanen, 2003), which in Strickland et al.'s model would contribute to evidence accumulation. The other source of inputs is the decision aid – for example, if the decision aid advises "conflict" then attending to the decision aid largely provides inputs corresponding to the conflict accumulator.

Specifying an evidence accumulation model requires assumptions about the temporal organization of decision making. For example, an accumulation model must specify whether the decision aid is processed in parallel with the task stimulus, or in series (e.g., process stimuli and then consider automated advice). In many psychological experiments, serial and parallel architectures can mimic each other's predictions, implying data could be fitted by some models from either architecture (Townsend, 1990). However, for a specific model, the descriptive adequacy of its combination of assumptions can be tested by examining fit to observed choices and RTs. In Strickland et al. (2021)'s model, the processing of task inputs and automation inputs occurs in parallel. Their model fitted observed patterns of automation use accurately and has the advantage of being highly tractable. Further, its parameters provide a plausible psychological account of how automated advice can be incorporated into decisions by affecting evidence accumulation, discussed below.

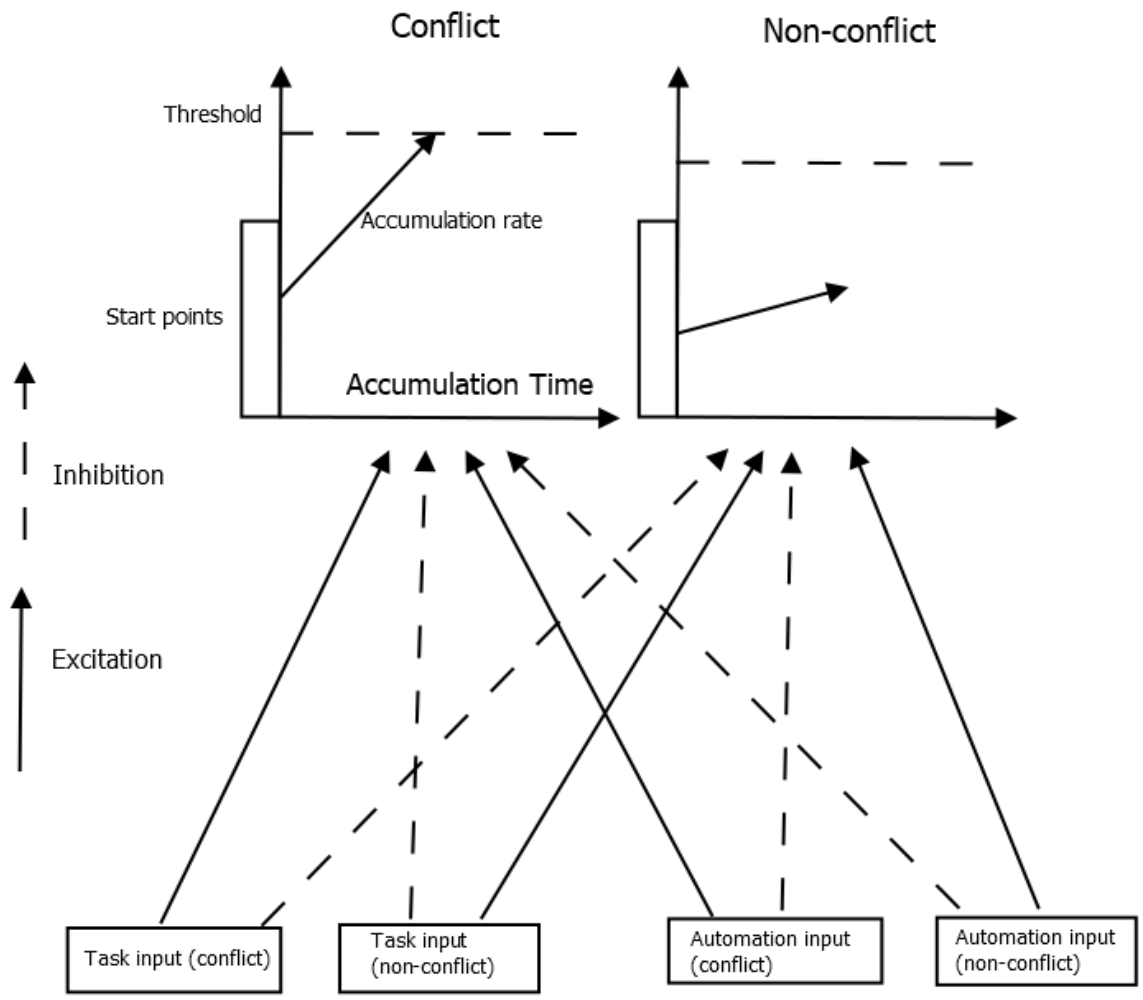

Figure 1. An evidence accumulation model of conflict detection performed with a decision aid (Strickland et al., 2021). Evidence towards each decision begins at some start point which varies from trial to trial, and then accumulates over time towards threshold. The first accumulator to reach its threshold determines the predicted decision. Accumulation rates are determined by inputs from the task stimuli (e.g., aircraft characteristics), and the automation decision aid. Inputs can excite (increase) accumulation towards corresponding accumulators and inhibit (decrease) accumulation towards opposing accumulators.

Strickland et al. proposed two mechanisms by which processing task and/or automation inputs could affect accumulation rates (see also Boag, Strickland, Heathcote, et al., 2019; Strickland et al., 2018). The first mechanism, *excitation*, refers to increased accumulation rates towards the decision matching the inputs (e.g., conflict inputs increase accumulation

towards the conflict decision). If inputs from a decision aid produce excitation (e.g., a decision aid recommending conflict stimulates evidence accumulation towards the "conflict" decision), the human could potentially make a decision without processing task inputs (e.g., they could respond "conflict" without assessing the aircraft), thus allowing the decision aid autonomy in triggering decisions. The second mechanism, *inhibition*, refers to when inputs matching a decision reduce accumulation rates towards the opposing decision (e.g., non-conflict inputs inhibit accumulation towards the conflict decision). Inhibition provides a mechanism for decision aids to influence the decision process without autonomously triggering decisions. Total evidence accumulation rates are determined by combined (summed) excitation and the inhibition derived from each independent input source (task and automation). Decisions require evidence to be accumulated (excitation) and thus cannot be triggered by inhibition alone. Therefore, if a decision aid provokes only inhibition (and no excitation) then the operator must still process task inputs to trigger decisions.

The Strickland et al. (2021) model was tested in a simulated ATC conflict detection task in which participants were provided a 90% reliable decision aid recommending a classification (conflict or non-conflict). The decision aid improved conflict detection accuracy on trials where it was correct and decreased accuracy on trials where it was incorrect. Further, the decision aid slowed mean correct RT on trials where it was incorrect, consistent with previous HAT studies (Bailey & Scerbo, 2007; Bowden et al., 2021).

Strickland et al. (2021) found that the effects of the decision aid were primarily on evidence accumulation rates, and not thresholds. Evidence accumulation rates were slowed towards decisions that disagreed with the automated advice relative to manual (no automation) trials, consistent with inhibition. There was little evidence that accumulation rates were faster for decisions agreeing with automated advice than on matched manual trials, suggesting that inputs from the decision aid did not excite (independently trigger) evidence

accumulation. The finding that participants used inhibition, rather than excitation, to integrate decision aids into their decisions fits with a cognitive strategy of allowing decision aids to influence decisions without autonomously triggering them, because no decision can be made without accumulating evidence based on task inputs.

Although Strickland et al. (2021) found evidence for a single cognitive process supporting automation use (inhibition), they only examined an approximately equal-ability human/automation team (90% reliable automation on a task where participants manual performance was approximately 90% accurate). However, in many cases decision aid reliability differs from human accuracy. Many HAT studies indicate that the extent of automation reliance increases with increased automation reliability (Bailey & Scerbo, 2007; Hussein et al., 2020a; Rovira et al., 2007; Shah & Bliss, 2017; Wiegmann et al., 2001). Some authors have suggested this is mediated by an increased human trust in automation (Dixon & Wickens, 2006; Dzindolet et al., 2001; Meyer & Lee, 2013), and this has been supported empirically (Hussein et al., 2020b). In particular, automation use is affected by the perceived reliability of the automation relative to the humans perceived/actual manual ability (Bailey & Scerbo, 2007; Hutchinson et al., 2022; Liang et al., 2022; Wiegmann et al., 2001). Thus, operators may apply different strategies when using decision aids with differing levels of reliability relative to their own. Deferring information processing to highly reliable decision aids could be adaptive if it frees up cognitive resources for other tasks or speeds up decisions with little risk of error (Kaber, 2018; Moray 2003, Moray & Inagaki, 2000). Avril et al. (2021) found evidence for such a transition, with participants spending less time visually sampling their primary task when decision aids were most reliable. However, the extreme version of this strategy, wherein humans rely entirely upon automation rather than processing information on their own, is a form of "automation complacency" (Parasuraman & Manzey, 2010), and would lead to misuse if decision aids fail. In contrast, responding to automation

inputs with only inhibition avoids such complete automation complacency, because decisions are not made until the human has processed primary task inputs.

Given that deferring autonomous information processing to decision aids is a strategy that people could apply to speed performance or free cognitive capacity, but also one that could lead to significant misuse, it is important to develop models that can identify this strategy and formally describe it. Such a strategy is likely present in situations where people highly trust automation. Trust is a key component underlying automation use (Dixon & Wickens, 2006; Hoff & Bashir, 2015; Meyer & Lee, 2013). However, much of the focus in the literature is on the subjective (attitudinal) aspects of trust, with less focus on establishing the cognitive mechanisms by which trust affects decisions. High levels of trust would be expected to encourage operators to accept automation without independently checking the automation's work, consistent with the proposal in our computational model regarding how humans transfer autonomy to the decision aid (i.e., via excitation).

Although the Strickland et al. (2021) model has the potential to identify the cognitive mechanisms by which increased trust in automation allows a decision aid to autonomously trigger decisions (i.e., excitation), to date the model has only implicated inhibition. This suggests that the decision aid in Strickland et al. played the role of advisor, without potential to autonomously trigger decisions. Thus, to date the Strickland et al. model has not been applied to situations where humans allow automated advice a degree of autonomy (independence) in triggering their decisions. As reviewed, it is possible that humans grant higher reliability decision aids more autonomy over their decisions (excitation in the Strickland et al. model), and this could explain increases reliance on automation associated with higher reliability. However, the effects of reliability could also be explained merely by humans treating higher-reliability aids as a more influential advisor without risking completely autonomous automation decisions (increased inhibition but no increased

excitation). To extend the Strickland et al. model, and to understand the cognitive mechanisms underlying the effects of automation reliability on decisions, the current study applies the Strickland et al. model to scenarios in which decision aids varied in reliability.

The Current Study

We used a similar ATC conflict detection paradigm to Strickland et al. (2021) and employed large trial numbers (4,680 per participant) to facilitate reliable modelling (Smith & Little, 2018). Our task was calibrated to achieve approximately 85% participant accuracy, so that we could include a condition with lower reliability than participants, and another with higher reliability. Our experiment included a manual (no decision aid) baseline condition, a "low-reliability" condition where participants were provided a 75% reliable decision aid, and a "high-reliability" condition where participants were provided a 95% reliable decision aid.

We expected to replicate Strickland et al. (2021)'s findings that correct decision aids increase accuracy, and that incorrect decision aids decrease accuracy and increase correct RTs. Further, we expected these effects would be stronger with higher automation reliability. Specifically, we expected that correct decision aids would lead to larger increases in accuracy, and incorrect decision aids would lead to larger decreases in accuracy and larger increases in RT, in the high-reliability condition than the low-reliability condition.

Although previous research has identified the behavioral manifest effects of automation reliability on accuracy and RT, it has not formally identified the latent cognitive mechanisms underlying these effects. Strickland et al. found that decision aids inhibited, but did not excite, evidence accumulation with an approximately equal-ability (90% accuracy) human-automation team and argued that this was an adaptive way to benefit from automation whilst mitigating misuse. Given our low-reliability condition's decision aid was less reliable (75%) than that of Strickland et al., and lower relative to expected participant accuracy, an excitation strategy would pose an even greater risk to automation misuse and thus we

expected that participants would rely primarily on inhibition. In the high-reliability condition, we also expected to find inhibition (to the extent that participants took advice from the decision aid without giving it full autonomy to trigger decisions), however, it could be adaptive for participants to allow the high reliability decision aid more autonomy in triggering responses. Should this be the case, we would expect to find excitation, that is, higher accumulation rates towards decisions that agree with automation relative to matched manual (unaided) decisions.

Given that Strickland et al. (2021) did not previously find effects of automation on thresholds, we expected to replicate that. However, participants can control their thresholds to affect the relative speed versus accuracy of their responding (e.g., Boag, Strickland, Loft et al., 2019), and tend to optimize how quickly they can complete an experiment whilst maintaining an acceptable level of accuracy (Hawkins et al., 2012). Thus, to the extent participants become satisfied with their accuracy with high-reliability automation, they might lower their thresholds to complete the task faster.

We measured subjective trust in automation after each task block that provided a decision aid. To the extent that excitation measures the objective effect of trust on decisions, we would expect the two to co-vary. Thus, in addition to expecting excitation in the high-reliability condition, we also expect automation trust ratings to be greater. We also investigate whether individual differences in our model parameters (e.g., excitation) predict individual differences in subjective trust, thereby revealing a cognitive mechanism by which increased trust in decision aids influences human decisions.

Method

Transparency and openness

We report the rationale for our sample size, all data exclusions, all manipulations, and all measures recorded in the study. Data and the analysis code associated with the manuscript

are available on the open science forum (https://osf.io/q3b2r). Task simulation materials are available on request. This study's design and analysis were not preregistered.

Participants

The University of Western Australia's Human Research Ethics Office approved the study. We aimed to test 24 participants. This is adequate to apply our cognitive model, which requires large trial numbers for individual-subjects modelling, rather than large participant numbers. Our design achieves a full counterbalance at 24 participants, but due to a miscommunication between experimenters 28 participants were tested. Our analysis focuses on 24 participants to maintain a balanced design and minimize confounds from practice and fatigue. For the counterbalanced analysis, one of the first 24 participants' data were replaced with one of the additional participants (who fortuitously matched for counterbalancing purposes), due to low accuracy in manual blocks (<60%) for sessions one and two. Of these 24 participants, 22 were undergraduate students participating for course credit and two were reimbursed $100 for their travel and time spent. Rewards were also paid out to participants based on overall percentage accuracy (reward = proportion correct x $40 AUD). The average age of participants was 21.3 years (*SD* = 5.80), excluding one participant who did not report their age. Seventeen participants identified as female and seven identified as male.

Design

Participants completed three sessions. Session one lasted approximately 2 hours (due to including training), and sessions two and three were approximately 1.5 hours each. In each session, participants performed three blocks of 520 trials: one high-reliability automation block, one low-reliability automation block, and one manual block. Block order was counterbalanced such that participants got each condition in a different position on each day (e.g., if they got the manual block first in session one, it would not be first in session two or three). This resulted in twelve possible block orders, which were balanced across condition.

To submit responses, participants were either required to press ‘f’ for conflicts and ‘j’ for non-conflicts or vice versa. Response key assignment was counterbalanced across participants, leading to the full counterbalance requiring 24 participants.

Materials

ATC Conflict Detection Task. The ATC task, depicted in Figure 2, was developed by Fothergill et al. (2009) and was used by Strickland et al. (2021). Participants viewed a 180 nautical miles (nmi) by 112.5nmi sector of airspace, with a 10nmi by 20nmi scale fixed on the left of the display. On each trial, two aircraft appeared within the light grey circle and flew straight paths towards the point of intersection. Although aircraft were moving towards the point of intersection, they did so very slowly, and so the available information about conflict status effectively did not change throughout the trial. Each aircraft was accompanied by a data block including callsign (e.g., WVH619), type (e.g., B737), flight level (e.g., 370 indicates 37,000 feet), and speed in knots (nmi per hour) divided by 10. Aircraft were also accompanied by a probe vector indicating heading and predicted position in one minute. Participants decided whether aircraft pairs would violate minimum separation, corresponding to a distance of 5nmi (laterally) and 1,000ft (vertically), at any time during their flight.

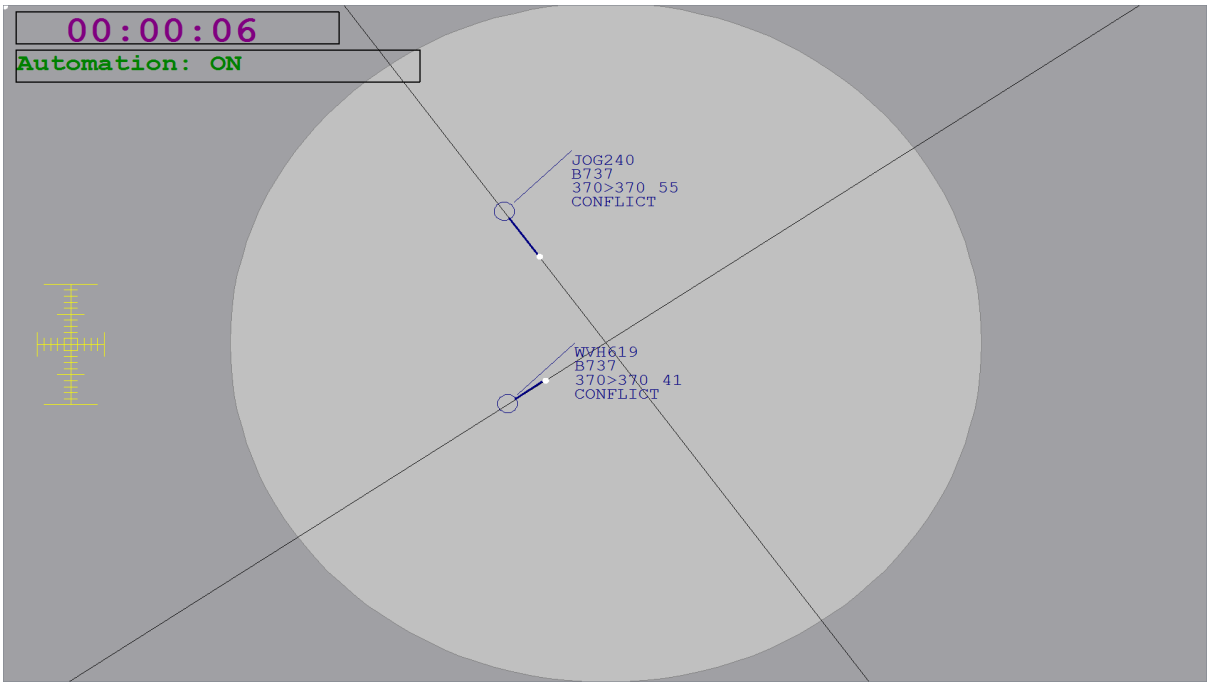

Figure 2. Display of the ATC conflict detection simulation. Next to each aircraft callsign (e.g., WVH619), aircraft type (e.g., B737), current and cleared altitude (e.g., 370>370), and airspeed (e.g., 41 standing for 410 knots) are displayed. A trial countdown timer is displayed on the top left. The automated decision aid advice is placed under the data block of each aircraft (e.g. in this screenshot, it is recommending responding 'conflict'). In manual conditions, this space was filled with a string with no special meaning, '########'.

Conflict detection stimuli. The altitude of all aircraft was fixed at 37,000 feet, and thus participants only needed to evaluate lateral separation to determine conflicts status. We randomly sampled the angle of approach of one of the aircraft, but fixed the relative angle of approach between aircraft at 90 degrees. Evaluating the conflict status of aircraft pairs requires extrapolating the future lateral separation of the aircraft using current locations and speeds. Conflict status was defined by distance of minimum lateral separation (d_{min}), with a d_{min} of 5 being the boundary between conflict and non-conflict. The spatial variables relevant to lateral separation of the aircraft pairs were similar to Strickland et al. (2021). In Strickland et al., 90% manual accuracy was achieved with d_{min} for conflicts drawn from the uniform distribution U [0, 1.5] nmi and non-conflicts drawn from the uniform distribution U [8.5, 10]

nmi. We aimed for a slightly lower manual accuracy (i.e., 85%) than Strickland et al., because it was the mid-point between high (95%) and low (75%) automation reliability. Thus, we shifted each of d_{min} distribution 0.5 nmi closer to the 5 nmi minimum lateral separation standard (0.5-2 and 8-9.5). Across trials, aircraft speeds varied randomly between 400 and 700 knots, and time to minimum separation between 120s and 210s, and both speeds and time to minimum separation were held constant within each trial.

Automated decision aid. As displayed in Figure 2, the decision aid was presented under the data block of each aircraft in both high- and low-reliability automation conditions (in manual conditions, there was a string '########' that participants were informed had no special meaning). The decision aid advised participants whether to classify aircraft as in conflict ("CONFLICT") or not in conflict ("NON-CONF"). On most trials, the automation advised the correct classification ('automation-correct trials'), but on some trials, the automation advised the incorrect classification ('automation-incorrect trials'). In the high-reliability condition, the automation was 95% reliable, and so there were 494 automation-correct trials and 26 automation-incorrect trials. In the low-reliability condition, the automation was 75% reliability, and there were 390 automation-correct and 130 automation-incorrect trials. Automation was unbiased such that in each automation block, half the automation-incorrect trials were conflicts, and half were non-conflicts.

For each session, stimuli in high- and low-reliability blocks were matched to manual blocks in terms of aircraft speeds and distances from intersection, to control conflict detection difficulty. However, new callsigns and angle of approach were generated for each block (whilst maintaining a relative angle of 90 degrees between aircraft) to avoid instance-based learning (Bowden & Loft, 2016), and the order of stimulus presentation was randomized each block. The stimuli that were selected to be automation-incorrect trials were chosen randomly for the low-reliability block, and a random subset of those stimuli were assigned to

automation-incorrect trials in the high-reliability blocks. In subsequent data analysis, the performance on manual stimuli matching to the low-reliability automation-incorrect stimuli is used as a baseline to compare to performance on automation trials.

Automation Trust Questionnaire. After completing each high- and low-reliability automation block, participants rated their trust in the decision aid (adapted from Merritt, 2011; supplementary materials). Participants rated six trust questions, such as "I believe the Automated Decision Aid is a competent performer", on a five-point scale from 'strongly disagree' to 'strongly agree'.

Procedure

Trial Procedure. Before each trial, participants were presented a screen that read "press space to continue". Participants were then presented with a pair of aircraft heading towards a common intersection, and had 7s to respond. This was 1s faster than Strickland et al. (2021) because they found non-responses were extremely rare, and to potentially help to reduce accuracy towards the target manual level of 85%. If participants responded accurately, they proceeded immediately to the next trial. Otherwise, they received feedback (e.g., "*Incorrect! This pair was in conflict*") and then pressed an 'ok' button to proceed.

Experimental Procedure. After providing informed consent, participants viewed training instructions and a demonstration of aircraft pairs with different $d_{\min}$. They then completed 40 training trials and proceeded to their first experimental block (manual, high-reliability automation, or low-reliability automation).

Before each automation condition, participants were instructed regarding their decision aid, and shown examples of how it provides advice. Participants were not told the exact percentage reliability of decision aids. However, in the high-reliability condition participants were instructed that "*Although the automation is highly reliable, it is not perfect, and automation advice errors are unlikely but still possible*". In the low-reliability condition,

participants were instructed that "*Although the automation is reasonably reliable, it is not perfect, and automation advice errors may be relatively common*". In manual conditions, participants were informed that there would be a string ('########') presented under each aircraft data block with no special meaning.

Participants were informed that they would receive a financial reward for task performance, ranging from $0 to $40 AUD. In both automation conditions, participants were instructed that "*In the event that the automation makes an incorrect recommendation, it is essential that you perform the correct action. Rejecting the automated recommendation when it is actually correct will reduce your performance score and subsequent bonus*". They were also told that "*Accepting the automated recommendation when it is wrong will result in a substantially greater reduction in your performance score and subsequent bonus"*, to emulate the common real-world situation in which automation misuse can have significant consequences. In manual conditions, they were instructed that "*Incorrect responses will reduce your performance score and subsequent bonus*".

Participants were presented with feedback on their accuracy at the end of each block and prompted to have a self-paced break. In the high- and low-reliability automation blocks, their feedback informed them of the percentage of trials on which they incorrectly accepted incorrect automated advice and the percentage of trials on which they incorrectly rejected correct automated advice. The procedures for sessions two and three were the same as session one, except they included only "refresher" instructions rather than training trials.

Results

All data analysis was conducted using the R programming language (R Core Team, 2021). Our analyses focus on the fully counterbalanced set of 24 participants. Follow up analyses including the extra 3 participants with interpretable data produced similar results, except for some minor differences in cognitive modelling (discussed in the supplementary

materials) which arose due to an influential participant who used the decision aid in the low-reliability condition in a manner more characteristic of the high-reliability condition. We excluded trials where participants did not respond (0.19% of trials) or responded in less than 0.2s (0.07% of trials) from analysis.

We first applied linear mixed-effects models to identify the effects of our manipulation on accuracy and RT, using the lme4 package (Bates et al., 2015). These mixed models included a random intercept for participants, and no random slopes. To analyze accuracy, we fit a generalized linear mixed model with a probit link function for each trial. To analyze RT, we fit a linear mixed effect models to mean correct RT for every participant. We also analyzed mean trust in automation with a linear mixed effects model. We tested for significance using Type II Wald Chi-Square tests, which are reported in supplementary materials. We followed up on significant effects using the *emmeans* package (Lenth et al., 2019), reported in supplementary materials. Thus, for mixed-model analyses all relevant *p*-values are tabulated in supplementary materials. In the individual differences section below, *p*-values and correlation coefficients are reported in text.

The factors in our mixed effects models included stimulus type (conflict/non-conflict), experimental session (one/two/three), blocked automation condition (manual/low-reliability automation/ high-reliability automation), and automation accuracy (automation-correct/automation-incorrect). For the high- and low-reliability conditions, automation-correct trials refer to trials where automation advised the correct decision and automation-incorrect to trials where the automation advised the incorrect decision. For the manual condition, this factor indexes trials that were matched to automation-incorrect trials and automation-correct trials from the low-reliability condition in terms of stimulus properties (e.g., distance of minimum separation, speed). As stimulus properties were randomly assigned across automation correct/incorrect trials, this factor is expected to have no strong

effect in the manual condition (and indeed it did not). Because we were interested in identifying strong effects, we set our significance criterion at $p < .005$ (Benjamin et al., 2018) for the manifest analyses reported below. The within-subjects SEs reported in text and figures follow the Morey (2008) bias-corrected method.

Accuracy

Participant accuracies are displayed in Figure 3. Accuracy was higher on conflict trials ($M = 0.82$, $SE = 0.05$) than on non-conflict trials ($M = 0.80$, $SE = 0.05$). Accuracy increased from session one ($M = 0.78$, $SE = 0.05$) to session two ($M = 0.82$, $SE = 0.04$), and was higher still on session three ($M = 0.84$, $SE = 0.04$). There were main effects of condition and of automation accuracy on participant accuracy, and these effects interacted. On trials where the automation was correct, accuracy was higher in the low-reliability condition ($M = 0.89$, $SE = 0.02$) than on matched manual trials ($M = 0.86$, $SE = 0.01$), and higher still in the high-reliability condition ($M = 0.95$, $SE = 0.02$). On trials where the automation was incorrect, accuracy was lower in the low-reliability ($M = 0.79$, $SE = 0.03$) condition than on matched manual trials ($M = 0.86$, $SE = 0.01$), and lower still in the high-reliability condition ($M = 0.52$, $SE = 0.05$). Because the high-reliability condition had a high proportion of automation-correct trials, overall participant accuracy (i.e., marginal accuracy over automation-correct and automation-incorrect trials weighted by proportion) ($M = 0.93$, $SE = 0.01$) was higher than that of the low-reliability condition ($M = 0.87$, $SE = 0.01$), which was in turn slightly higher than that of the manual condition ($M = 0.86$, $SE = 0.01$).

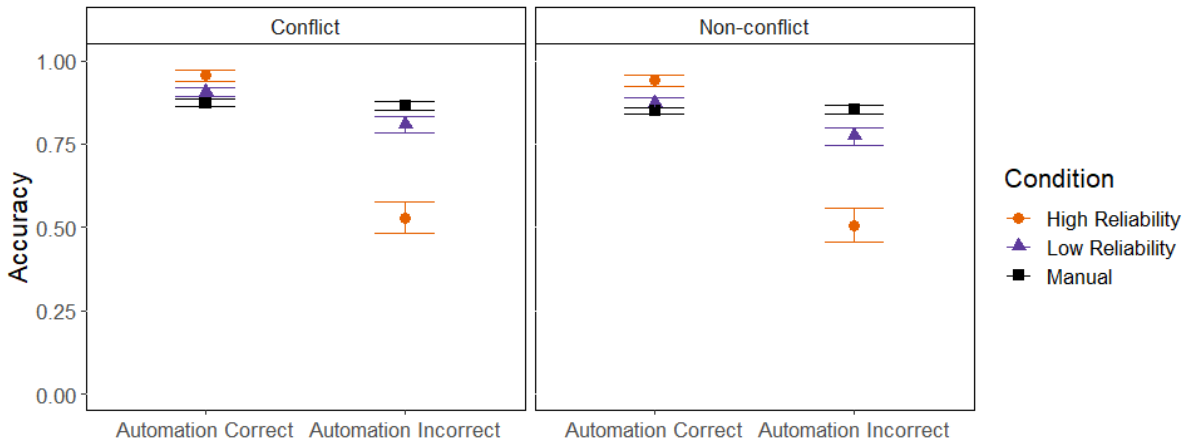

Figure 3. Conflict detection accuracy. Each panel corresponds to one stimulus type, on one experimental session. We depict within-subjects error bars that were calculated using the Morey (2008) bias-corrected method.

Response Times

Mean correct RTs are displayed in Figure 4. Correct RTs were longer on conflict trials (M = 2.06s, SE = 0.17s) than on non-conflict trials (M = 1.87s, SE = 0.16s). Correct RT decreased from session one (M = 2.41s, SE = 0.05s) to session two (M = 1.83s, SE = 0.04s), and decreased further on session three (M = 1.64s, SE = 0.04s). There were main effects of condition and automation accuracy on correct RTs, and these effects interacted. On trials where automation was correct, RTs were significantly faster in the high-reliability condition (M = 1.71s, SE = 0.11s) than in either of the other conditions. However, there were not significant differences in correct RTs between the low-reliability condition (M = 1.87s, SE = 0.09s) and matched manual trials (M = 1.88s, SE = 0.1s). For trials where the automation was incorrect, correct RTs were slower in the low-reliability condition (M = 2.06s, SE = 0.12s) than in the manual condition (M = 1.88s, SE = 0.1s), and slowest of all in the high-reliability condition (M = 2.41s, SE = 0.17s).

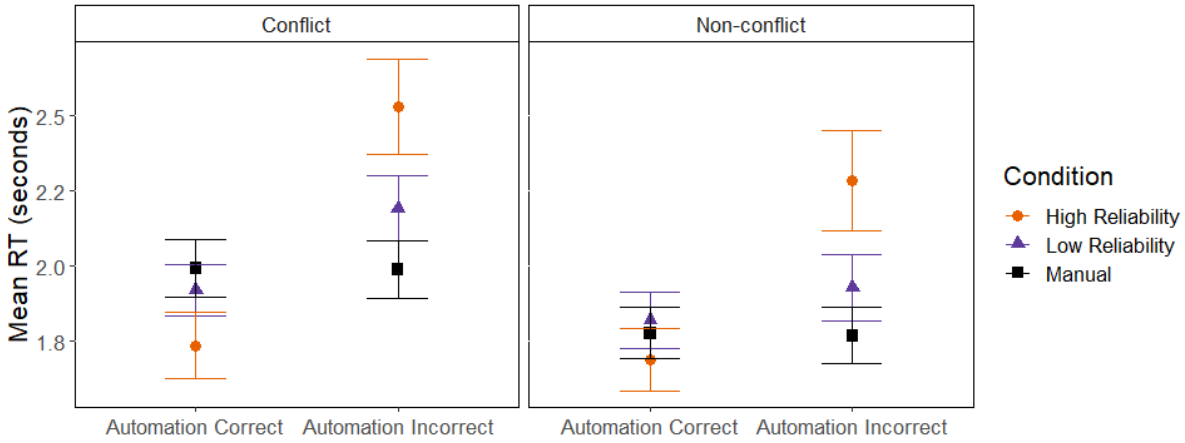

Figure 4. Correct conflict detection response times (RT) in seconds. Each panel corresponds to responses to one type of stimulus on one experimental session. We depict within-subjects error bars that were calculated using the Morey (2008) bias-corrected method.

Trust in Automation

We report the internal consistency and test-retest reliability of the trust in automation scale in the supplementary materials. Automation trust scores were created by averaging participants' answers to the six trust questions they completed after each automation block. One participant had two missing responses, and one participant had one missing response, so for those participants their average excluded those questions. We analysed the effects of automation condition and session on trust in automation. Trust was higher in the high- (M = 3.58, SE = 0.13) than the low-reliability condition (M = 2.32, SE = 0.13). The effect of session, and its interaction with condition, did not reach significance.

Individual Differences

Our highly reliable measurement of each participant (due to high trial numbers) meant that some clear individual-difference trends emerged even in our counterbalanced sample of 24 participants (see Rouder et al., 2019 for the importance of high trial numbers for studying individual differences). We examined the relationships between the effects of automation including the benefits to accuracy from correct automated advice, the costs to accuracy of

incorrect advice, the benefit to RT of correct advice and the cost to RT of incorrect advice. Accuracy benefits were calculated by subtracting accuracy on manual trials from accuracy on automation-correct trials, and costs by subtracting accuracy on automation-incorrect trials from accuracy on manual trials. Similarly, RT benefits were derived by subtracting RT on automation-correct trials from RT on matched manual trials, and RT costs by subtracting RT on matched manual trials from RT on automation-incorrect trials.

The accuracy benefits of correct automation were positively correlated with the accuracy costs of incorrect automation for both the high-, r (22) = .88, $p < .001$, and the low-reliability condition, r (22) = .59, p = .002. Correlation between RT benefits and RT costs did not reach significance for the high-reliability condition, r (22) = -.38, p = .065, but did for the low-reliability condition, r (22) = -.69, $p < .001$, in which there was a negative correlation between RT benefits of correct automation and RT costs of incorrect automation.

We also explored correlations between the effects of automation and trust in automation. In the high-reliability condition, trust in automation was correlated both with the accuracy benefits of correct automation, r (22) = .65, $p < .001$, and the accuracy costs of incorrect automation, r (22) = .60, p = .002. In contrast, in the low-reliability condition trust was correlated with neither the accuracy benefits of correct automation, r (22) = .28, p = .179, nor the accuracy cost of incorrect automation, r (22) = .11, p = .595. The benefits of correct automation to RT were not significantly ($p < .005$) correlated with trust in either the high-, r (22) = .41, p = .049, or low-reliability conditions, r (22) = -.02, p = .935, although the high-reliability condition reached the conventional p threshold of <.05. The costs of incorrect automation to RT were not significantly correlated with trust in the high-, r (22) = -.30, p = .158, or low-reliability condition, r (22) = -.06, p = .788.

We also examined individual differences using difference scores calculated on an alternative scale for both accuracy and RT. Specifically, we converted accuracy to the probit

scale, and converted mean RTs to mean [log (RT)] before taking differences (results tabulated in the supplementary materials). Trends in the correlations were similar, but some patterns of significance differed. Specifically, the positive correlations in the high-reliability condition between accuracy costs of automation and trust, and between accuracy benefits of automation and trust, no longer reached our strict criterion of $p < .005$ (in both cases, $p \sim .01$). In addition, the positive correlation in the low-reliability condition between automation accuracy costs and automation accuracy benefits no longer reached significance ($p \sim .05$).

Model Results

Model specification. We applied a two-accumulator linear ballistic accumulator (Brown & Heathcote, 2008) as depicted in Figure 1, with one accumulator corresponding to responding 'conflict' and one to responding 'non-conflict'. On each trial, evidence in each accumulator begins at a start point drawn from a uniform distribution U [0, A]. Thus, start point noise is estimated by the parameter A. Evidence then accumulates linearly at an accumulation rate drawn from the truncated normal distribution N (v, sv) with a lower bound of 0. The first accumulator to reach its threshold, b, determines the decision made, and total response time is equal to accumulation time plus non-decision time, a constant which is estimated. We report thresholds in terms of B, which is equal to b - A.

We fixed the variability in accumulation rates (*sv*) for mismatching accumulators at 1 as a scaling parameter. For each participant we estimated only one sv for matching accumulators, one A parameter and one non-decision time, to encourage sound parameter estimation. Thresholds and mean accumulation rates were varied across experimental factors to address theoretical questions of interest. Accumulation rates varied by stimulus type (conflict/non-conflict), condition (manual/low reliability/ high reliability), match to stimulus type (match/mismatch), and automation accuracy (automation-correct/automation-incorrect). Thresholds varied by latent accumulator (conflict/non-conflict), by condition (manual/low

reliability/ high reliability), and by experimental session (one/two/three), with the latter to account for practice effects. To avoid circularity, thresholds were not allowed to vary by stimulus type – if the threshold could be set based on stimulus type, this would imply the participant already knew the correct answer, and so would not need to accumulate evidence. Similarly, thresholds did not vary based upon whether automation was correct, as they are assumed to be set before participants accumulate evidence.

Model Fit

We obtained Bayesian estimates of model parameters using the Dynamic Models of Choice R Suite (Heathcote et al., 2019). These estimates are posterior samples, proportional to probability distributions of the model parameters given the data and prior information about the parameter values. The details of estimation are in the supplementary materials. The priors we used were the same as Strickland et al. (2021), except that the prior on the *A* parameter was tightened as this was necessary to achieve convergence for two participants (results were similar excluding these participants and using the Strickland et al. prior).

Figure 5 displays fit of the posterior predictions of the model to the data. Generally, the model fitted the data well. There was some minor miss-fit to median correct RT in the high-reliability condition on automation-incorrect trials, with the model somewhat underestimating the slowing induced by the incorrect advice. However, the model did fit the direction of this effect, and that this represents a small number of responses (there are relatively few trials where highly reliable automation was incorrect). Given the relatively limited data for this cell, this miss-fit to rare RTs may result from influence from the prior.

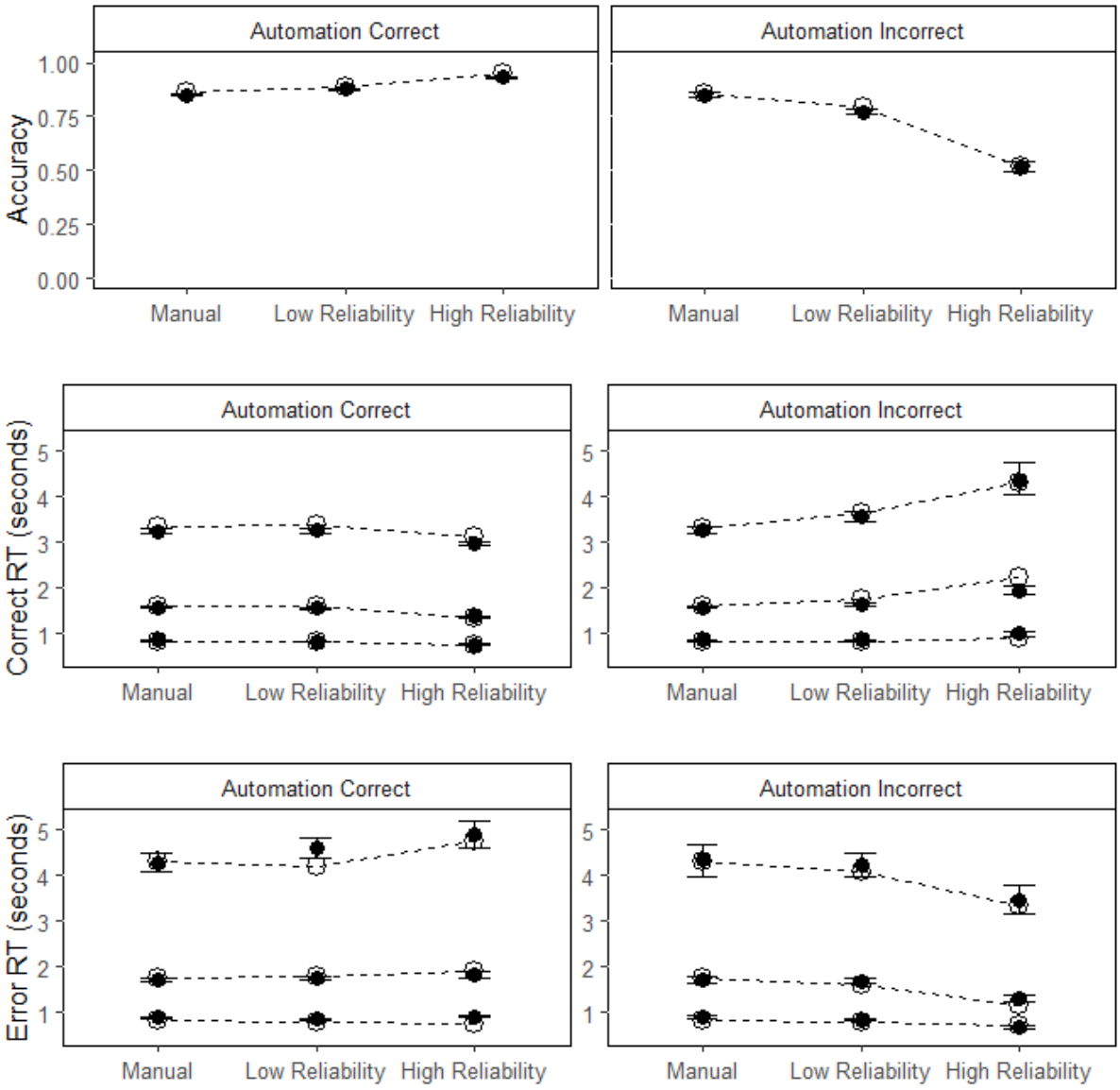

Figure 5. Posterior model predictions of accuracy and response time (RT). The model predictions correspond to the white circles, the posterior means correspond to the black shaded dots. The error bars display the 95% posterior credible intervals of the predictions. For both model and data, we created a grouped data frame concatenating each participants' data, and the plot contains accuracy and quantiles of response time calculated using the grouped data. Three quantiles of RT are depicted, with the 0.1 quantile of RT grouped on the bottom, the median RT at the middle, and the 0.9 quantile of RT at the top.

Parameter Inference

We created a set of group-averaged posterior samples, which we use for the purpose of inference, by averaging parameter values across participants for every posterior sample. The model parameters resulting from these group-averaged posteriors samples are reported in the Supplementary Materials. In text, we focus on comparisons of accumulation rates and thresholds that allow us to answer the key theoretical questions. We test effects with a Bayesian one-tailed posterior *p* value, indexing the proportion of posterior samples for which that effect was smaller than 0 (i.e., small *p*-values suggest there is a difference). To estimate effect size, we report the posterior mean of each effect divided by the posterior standard deviation, referred to as *Z*.

Excitation and Inhibition

Figure 6 depicts estimates of evidence accumulation rates. Measuring excitation and inhibition effects requires comparing accumulation rates on (high- or low-reliability) automation-block trials with corresponding accumulation rates in manual blocks. Excitation increases the accumulation rate towards the accumulator agreeing with the decision aid. Thus, for trials where the decision aid advises classifying aircraft as in conflict, excitation is measured by the *increase* in accumulation rates towards deciding aircraft *are* in conflict, as compared with corresponding accumulation rates on matched manual trials. Similarly, for trials where the decision aid advises classifying aircraft as not in conflict, excitation is measured by the increase in accumulation rates towards deciding aircraft are not in conflict, as compared with corresponding accumulation rates on matched manual trials. Inhibition reduces accumulation towards the accumulator that disagrees with the decision aid. Thus, for trials where the decision aid advises that aircraft are in conflict, inhibition is measured by the *decrease* in accumulation rates towards deciding aircraft are *not* in conflict, as compared with corresponding accumulation rates on matched manual trials. Similarly, for trials where the

decision aid advises classifying aircraft as not in conflict, inhibition is measured by the decrease in accumulation rates towards deciding aircraft are in conflict, as compared with corresponding accumulation rates on matched manual trials.

Table 1 contains statistical tests of excitation and inhibition effects. We found strong evidence of inhibition in both conditions, with accumulation rates lower to decisions disagreeing with automation than on matched manual trials, replicating Strickland et al. (2021). Further, inhibition was stronger in the high compared to low-reliability condition. Importantly, and in contrast to Strickland et al.'s results, in the high-reliability condition, we also found strong evidence of excitation on all trial types: accumulation rates were higher to decisions agreeing with automation than on matched manual trials.

In the low-reliability condition, we did not find substantial evidence of excitation on automation-correct trials. In fact, we found an effect in the opposite direction of excitation for low-reliability automation: non-conflict accumulation rates were lower on automation-correct trials than on matched manual trials. We did find significant excitation effects on automation-incorrect trials. However, these effects were weaker than in the high-reliability condition, and automation-incorrect trials are relatively less well estimated than automation-correct trials because there are less of them. Thus, evidence for excitation in the low-reliability condition is less consistent than the high-reliability condition, and the detected effects are weaker.

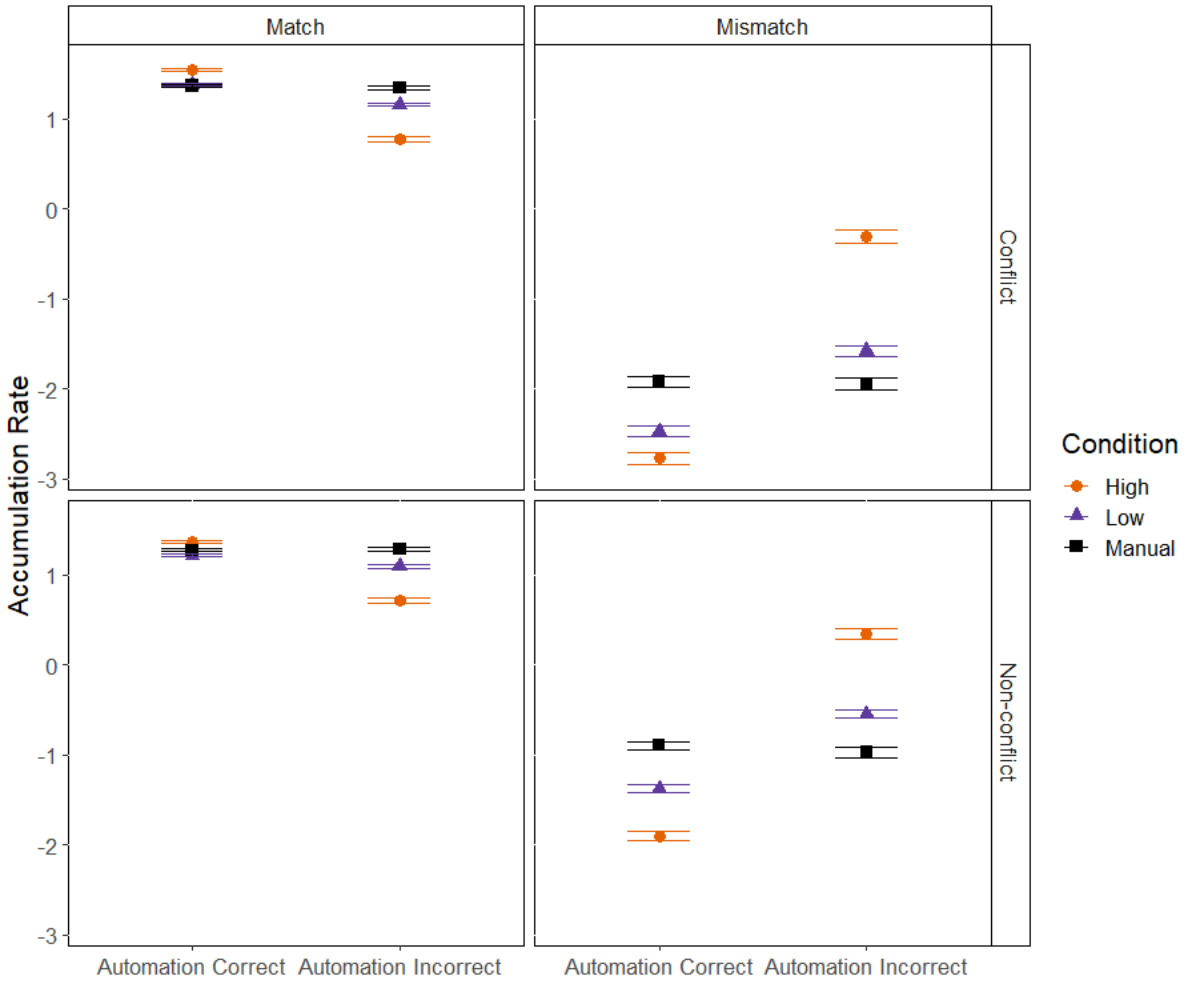

Figure 6. Estimates of group-averaged accumulation rates. Excitation is indicated by increased 'match' accumulation rates, as compared with manual conditions, when decision aids were correct (leftmost points on the figure, substantial for the high-reliability condition) and increased 'mismatch' accumulation when decision aids were incorrect (rightmost points on the figure, substantial for both automation conditions). Inhibition is indicated by decreased 'match' accumulation rates when decision aids were incorrect (middle-left points on the figure, substantial for both automation conditions) and decreased 'mismatch' accumulation rates with automation when decision aids were correct (middle-right points on the figure, substantial for both automation conditions). Shapes indicate the posterior means and error bars indicate the mean plus or minus the posterior standard deviation. In cases where the shapes are overlapping, the condition means are very close together (e.g., match, automation-correct, conflict, low reliability vs manual).

Table 1

Statistical tests of automation-induced excitation and inhibition effects. We depict Z(p), where Z is the posterior mean of the effect divided by the standard deviation, and p is the one-tailed posterior probability against their being an effect. L = Low Reliability Condition, H = High Reliability condition.

Trial Type	Excitation (L)	Inhibition (L)	Excitation (H)	Inhibition (H)
Conflict Trials				
Automation Correct	0.92 (.178)	9.1 (<.001)	13.05 (<.001)	13 (<.001)
Automation Incorrect	4.62 (<.001)	11.02 (<.001)	17.71 (<.001)	18.45 (<.001)
Non-conflict Trials				
Automation Correct	-4.63 (<.001)	10.61 (<.001)	7.24 (<.001)	20.06 (<.001)
Automation Incorrect	7.02 (<.001)	11.19 (<.001)	17.89 (<.001)	16.67 (<.001)

Threshold effects

Figure 7 depicts estimates of thresholds, and Table 2 contains statistical tests of decreases in thresholds in two automation conditions compared with manual conditions. There were some small drops in threshold levels in the low-reliability condition compared with the manual condition, suggesting participants may have become less cautious when provided automation. There were relatively stronger decreases in thresholds observed in the high-reliability condition when compared with the manual condition.

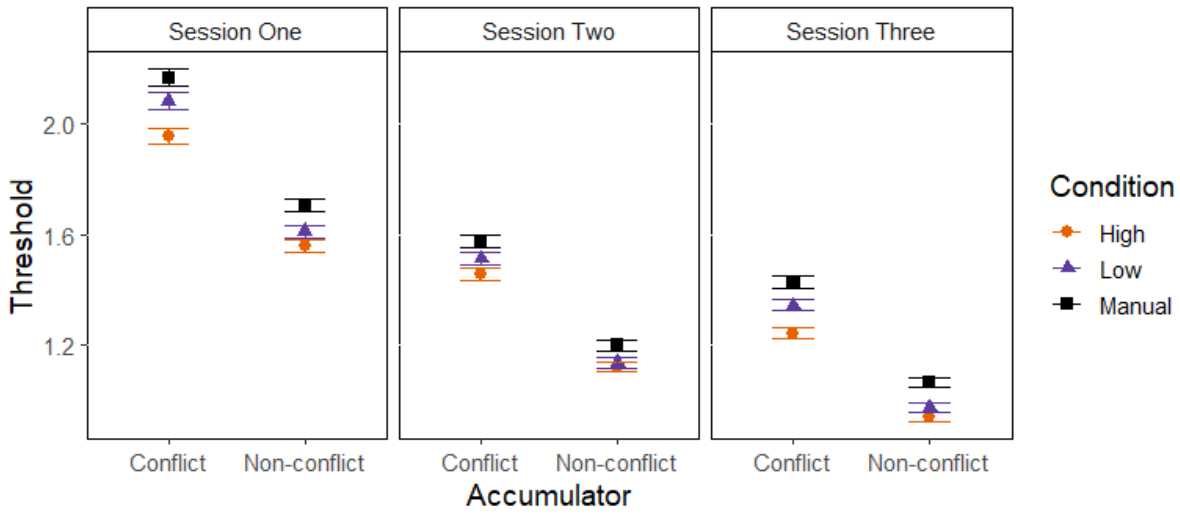

Figure 7. Estimates of group-averaged thresholds. Shapes indicate the posterior means and error bars indicate the mean plus or minus the posterior standard deviation.

Table 2

Statistical tests of differences in thresholds across automated and manual conditions. We report Z(p), where Z is the posterior mean of the parameter difference divided by the standard deviation of the parameter difference, and p is the one-tailed posterior probability against their being an effect. L = Low Reliability Condition, H = High Reliability condition.

Accumulator (Condition)	Session One	Session Two	Session Three
Conflict Accumulator (L)	3.47 (<.001)	3.39 (<.001)	4.83 (<.001)
Non-conflict Accumulator (L)	4.53 (<.001)	4.17 (<.001)	6.57 (<.001)
Conflict Accumulator (H)	9.01 (<.001)	6.6 (<.001)	11.05 (<.001)
Non-conflict Accumulator (H)	7.36 (<.001)	5.24 (<.001)	9.28 (<.001)

Posterior Exploration

To summarize, we found substantial evidence that both high- and low- reliability decision aids are incorporated into decisions with an inhibition mechanism, strong evidence

that participants relied on an additional excitation mechanism when using high-reliability automation, and weaker evidence of excitation for low-reliability automation. We also found some evidence that thresholds were reduced when using highly reliable automation. However, in evidence accumulation models the relation between model predictions and model parameters is complex and nonlinear. Thus, we used simulations to shed light upon how the mechanisms estimated in our model contributed to its predictions of the costs and benefits of automation on accuracy and RTs, by removing mechanisms from the model by changing parameter values. To the extent removing mechanisms causes miss-fit to observed effects that the model previously fitted, those mechanisms were responsible for the estimated model's fit of the effects. To remove excitation from the model, accumulation rates to accumulators agreeing with the decision aid advice were set to the values estimated from matched manual trials. To remove inhibition, accumulation rates to decisions disagreeing with the decision aid were set equal to their value in matched manual trials. To remove threshold effects, thresholds were set to the levels observed in manual conditions.

Figures 8 and 9 depict the results of the posterior exploration. We found major miss-fits to automation's effects on accuracy when inhibition was removed for both high and low reliability conditions, and also major miss-fits to accuracy effects for the high-reliability condition when excitation was removed. This suggests that inhibition was important for explaining automation's effects on accuracy in both conditions, and excitation important in the high-reliability condition. We also found that removing inhibition from the model caused major miss-fits to the effects of automation on RT for both high and low reliability conditions, particularly costs of incorrect automation to RT, demonstrating the importance of the inhibition mechanism in driving this RT cost. In contrast to these findings, setting thresholds to manual levels had little effect on predictions regarding automation's effect, suggesting they were substantially less important to model predictions.

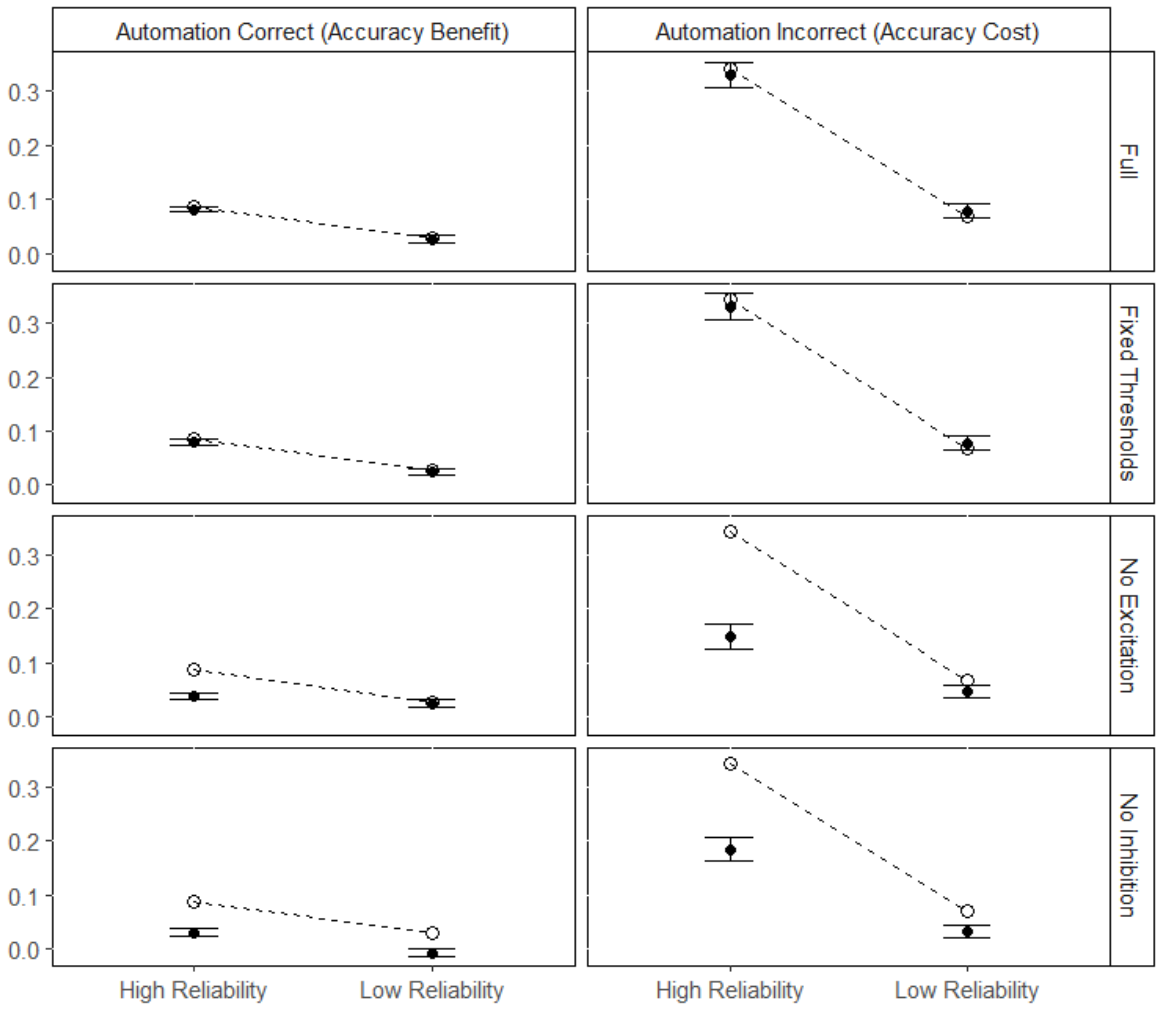

Figure 8. Exploration of the importance of model mechanisms in explaining automation's effects on accuracy. Accuracy benefit refers to improvement of accuracy on automation-correct trials, and accuracy cost to decrements in accuracy on automation-incorrect trials. Model mechanisms were removed by setting parameter values equal to matched manual conditions and resulting miss-fit indicates the degree to which that mechanism was responsible for the full model's ability to predict effects. Model predictions correspond to the white circles, the posterior means correspond to the black shaded dots. The error bars display the 95% posterior credible intervals of the predictions.

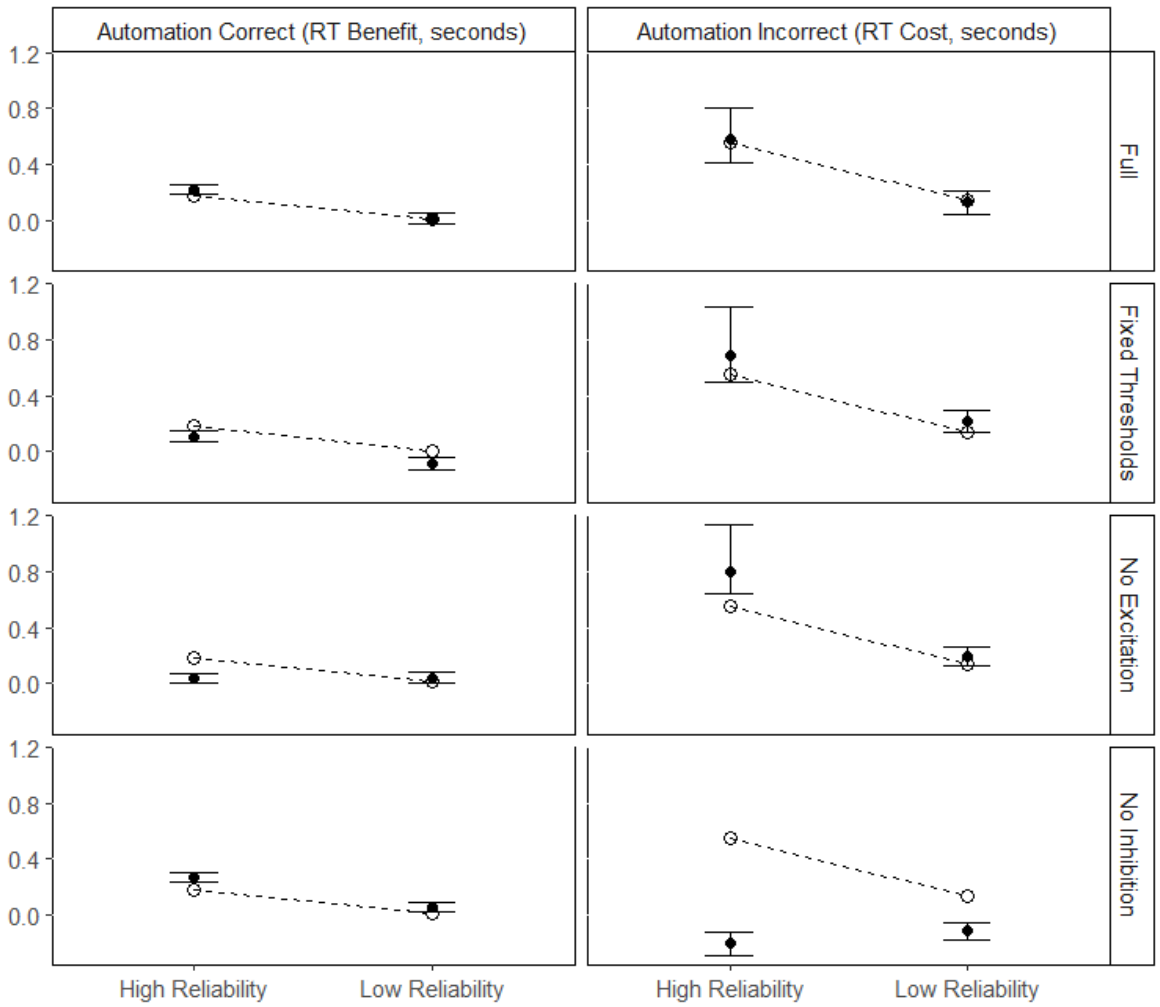

Figure 9. Exploration of the importance of model mechanisms in explaining automation's effects on RT. RT benefit refers to the speeding of correct RT on trials where automation was correct, and RT cost refers to the slowing of RT on trials where the automation was incorrect. Model mechanisms were removed by setting parameter values equal to matched manual conditions and resulting miss-fit indicates the degree to which that mechanism was responsible for the full model's ability to predict effects. Model predictions correspond to the white circles, the posterior means correspond to the black shaded dots. The error bars display the 95% posterior credible intervals of the predictions.

Individual differences (Model Mechanisms)

We examined how individual differences in model parameters related to the costs and benefits of automation to accuracy. To do so, we constructed distributions of "plausible

value" correlations (Ly et al., 2017), in which a correlation between model parameters and the data of interest is calculated for every posterior sample, creating a distribution of correlations. We then adjusted this plausible value distribution to correct for population-level inference (i.e., inferences about new participants, Ly et al., 2018). This approach of correcting for sample size is quite conservative, and so we are confident in individual-difference trends identified by this method. In Table 3 we report posterior means of plausible value correlations and accompanying 95% credible intervals. Although our framework provides evidence for correlations that can be assessed in a graded, rather than binary, manner, for brevity in text we focus on "plausible" correlations, referring to those with 95% credible intervals that do not overlap 0. We calculated correlations using a single estimate of excitation and inhibition for each participant by averaging over conflict and non-conflict trials, and over automation-correct and automation-incorrect trials. We also examined correlations of parameters with automation costs and benefits calculated on transformed scales: the probit scale for accuracy, and log[mean(RT)] for RTs. These correlations, which are similar to those reported in text, are included in the supplementary materials.

In the high-reliability condition, we found a plausible positive correlation between the benefits of correct automation to accuracy and excitation, but we did not find a plausible correlation between automation accuracy benefits and inhibition. In contrast, in the low-reliability condition, there was a plausible positive correlation between the benefits of correct automation and inhibition, but no plausible correlation between accuracy benefits and excitation. In the high-reliability condition there was a plausible positive correlation between the costs of incorrect automation to accuracy and excitation but not inhibition. In the low-reliability condition, there was a plausible positive correlation between the costs to accuracy of incorrect automation and inhibition, but also a plausible positive correlation between the costs to accuracy of automation and excitation.

We found that benefits to RT on automation-correct trials were (plausibly) positively correlated with excitation for both reliability conditions, whereas they were more weakly (not plausibly) negatively correlated with inhibition in both reliability conditions. In contrast, costs to RT on automation-incorrect trials were (plausibly) positively correlated with inhibition for both reliability conditions, and weakly (not plausibly) negatively correlated with excitation for both conditions.

We also examined how model parameters correlated with trust in automation. In the high-reliability condition, there was a plausible positive correlation between excitation and trust and a near-plausible (credible interval barely overlapping 0) negative correlation between inhibition and trust. In contrast, in the low-reliability condition there was nothing approaching a plausible correlation between trust and either excitation or inhibition.

Table 3

Correlations between model mechanisms, benefits of automation, costs of automation, and trust in automation. We constructed distributions of "plausible values" (Ly et al., 2017) of correlations by calculating the Pearson correlation across participants between model parameters and dependent variables for every posterior sample. This was then corrected for population-level inference (Ly et al., 2018). We report posterior mean (95% credible interval) of these correlation distributions. In text we refer to "plausible" correlations as those with 95% credible intervals that do not overlap 0. We obtained a single estimate of excitation and inhibition for each participant by averaging over conflict and non-conflict trials, and over automation-correct and automation-incorrect trials.

Condition/Dependent Variable	Excitation	Inhibition
High Reliability	Posterior mean *r* (CI)	Posterior mean *r* (CI)
Accuracy Benefit	0.57 (0.24 - 0.79)	-0.12 (-0.49 - 0.27)
Accuracy Cost	0.63 (0.33 - 0.83)	-0.10 (-0.48 - 0.29)
RT Benefit	0.75 (0.51 - 0.89)	-0.29 (-0.63 - 0.09)
RT Cost	-0.25 (-0.60 - 0.14)	0.54 (0.19 - 0.77)
Trust	0.53 (0.19 - 0.77)	-0.33 (-0.65 - 0.05)
Low Reliability	Posterior mean *r* (CI)	Posterior mean *r* (CI)
Accuracy Benefit	0.19 (-0.22 - 0.55)	0.59 (0.26 - 0.8)
Accuracy Cost	0.40 (0.01 - 0.69)	0.62 (0.31 - 0.82)
RT Benefit	0.50 (0.14 - 0.76)	-0.28 (-0.62 - 0.10)
RT Cost	-0.11 (-0.50 - 0.28)	0.62 (0.31 - 0.82)
Trust	0.01 (-0.39 - 0.41)	0.03 (-0.36 - 0.41)

Discussion

This study examined the use of automated decision aids in a simulated ATC conflict detection task including blocks of trials with low-reliability and high-reliability decision aids. We found higher accuracy on trials where decision aids were correct compared with matched

trials in a manual (no automation) block, and lower accuracy on trials where decision aids were incorrect, demonstrating that participants used the automation to inform their decisions. As predicted, automation reliability moderated the effects of the decision aid on accuracy, with participants more accurate when high-reliability automation was correct than low-reliability automation (when both compared to manual performance), and less accurate when high-reliability automation was incorrect than low-reliability automation. This replicates the benchmark effect of automation reliability demonstrated in the HAT literature (e.g., Bailey & Scerbo, 2007; Hussein et al., 2020a; Rovira et al., 2007). Because of the speed-accuracy tradeoff, it is important to also interpret these effects in light of changes in RT (Heitz, 2014; Pachella & Pew, 1968). We found that correct RTs were also slower when automation was incorrect as compared with manual trials, replicating Strickland et al. (2021; see also Bailey & Scerbo, 2007; Bowden et al., 2021). As expected, this effect was stronger in the high- compared to low-reliability condition. One novel effect found in the high-reliability condition, but not the low-reliability condition, was a decrease in correct RT (an improvement) when automation was correct as compared with manual trials. Participants also reported higher trust in automation after high compared to low-reliability blocks.

We fit an evidence accumulation model to identify the cognitive mechanisms underlying how the decision aid affected performance. We replicated Strickland et al. (2021)'s finding that evidence accumulation rates were lower to decisions that automation advised against as compared with manual trials. This effect was stronger in the high than the low-reliability condition. This indicates that to some extent, participants relied upon an inhibition mechanism which allows automation to act as an advisor, affecting decisions, without necessarily autonomously triggering them, and that automated advice was weighted more heavily in the high-reliability condition (hence the need for greater inhibition of decisions that automation advised against). It corresponds to a type of speed-accuracy

tradeoff, in which the benefits of automation to accuracy trade against slowing of RT to responses disagreeing with automation. It differs from traditional speed-accuracy tradeoffs in that it plays out in accumulation rates rather than thresholds, although in some cases even traditional speed-accuracy tradeoffs can also involve accumulation rates (Rae et al., 2014).

In the high-reliability condition, we found substantial evidence that automation increased the evidence accumulation rates of decisions agreeing with automation advice (excitation). This demonstrates a novel mechanism underlying automation use that was not identified in Strickland et al. (2021)'s examination of an equal-reliability human/automation team, consistent with participants allowing higher-reliability automation more autonomy in triggering their decisions. There was also some evidence of excitation in the low-reliability condition, but this evidence was inconsistent and relatively weaker than the high-reliability condition. Further, model simulations indicated that excitation was a relatively weak contributor to performance outcomes for the low-reliability condition, but a strong contributor for the high-reliability condition.

To summarise, our findings regarding accumulation rates represent a significant theoretical advance in explaining benchmark effects in the HAT literature regarding increased automation reliance with increased reliability (e.g., Hussein et al., 2020a; Wiegmann et al., 2001), and more specifically, increased automation reliance when decision aids are more reliable relative to human manual ability (e.g., Avril et al., 2021; Bailey & Scerbo, 2007; Hutchinson et al., 2022). Our computational model suggested that two processes drove the effects of automation reliability. First, inhibition of decisions opposing automated advice was stronger with high-reliability decision aids than with low-reliability decision aids, although substantial for both. Second, with high-reliability decision aids, but not low-reliability decision aids, we found substantial evidence that participants granted automation a degree of autonomy over decisions (via an excitation process).

Contrasting with Strickland et al. (2021), who did not find an effect of decision aids on thresholds, we found a small decrease in thresholds in the low-reliability condition, and a more substantial decrease in the high-reliability condition, relative to the condition with no decision aid. Thresholds may have been reduced in the high-reliability condition because participants were satisfied with the accuracy that could be achieved and were motivated to decrease RT to complete the experiment (Hawkins et al., 2012). However, our model simulations indicated that the observed threshold effects had relatively minor implications for the observed effects of automation, and thus they should be interpreted with caution.

It is interesting to consider our results alongside the findings of Bartlett and McCarley (2021), who applied signal detection theory to examine the effects of automation reliability. They too found that participants tended to use high-reliability decision aids more than low-reliability decision aids, and that higher-reliability decision aids benefitted accuracy overall. Their modelling approach spoke to the optimality of decision aid use under signal detection theory. They found that decision aid use was suboptimal for high-reliability aids, and closer to optimal for low-reliability aids, with the latter consistent with participants reducing reliance on the low-reliability advice. Our model provides new insights into how reliability affects decision aid use. Specifically, our model indicated different patterns of inhibition and excitation of evidence accumulation, depending on reliability. These findings complement Bartlett and McCarley, speaking to how incorporating RT distributions, in addition to accuracy, can provide novel insights into psychological processes. However, to account for RT distributions our model committed to specific assumptions about how decisions unfold over time. An advantage of the Bartlett and McCarley approach is the ability to assess a variety of strategies to using automation, such as a “coin flip” (in case of disagreement, accepting automated advice with 50% probability), probability matching (if disagreement, accepting advice at a rate proportional to its reliability), and contingent criterion shifts (bias

to respond in favour of the response agreeing with automation). Although our model provided an adequate and coherent account of the behaviour observed in this study, future data may require alternative architectures to capture a broader set of cognitive strategies.

Our design focused on large trial numbers to enable individual-subject modelling. However, a benefit of this design was that the large trial numbers reduced measurement noise to the extent that we could identify individual-difference correlations between patterns of automation use, trust in automation, and model parameters. The benefits of correct automated advice to accuracy correlated strongly with the costs of incorrect advice. This is consistent with the "lumberjack analogy" invoked in the HAT literature (Sebok & Wickens, 2017), in which higher degrees of automation provide increasing benefits to performance but leaves humans at greater risk of failing to identify automation failures (Onnasch et al. 2014). Another interesting finding was that trust was correlated with both automation costs and benefits in the high-reliability condition, but not the low-reliability condition. This might owe to individual differences in the degree to which participants afforded decision aids more autonomy in triggering their decision making in the high-reliability condition. As little autonomy was allocated to the low-reliability decision aid, there would be less scope for this individual difference process to co-vary with perceived trust in automation.

We found that excitation, that is the extent to which decision aids autonomously triggered evidence accumulation, correlated with the benefits of correct automation to accuracy and the costs of incorrect automation to accuracy for the high-reliability condition, but only correlated with automation costs to accuracy for the low-reliability condition. In contrast, inhibition correlated with the benefits of correct automation and the costs of incorrect automation to accuracy for the low-reliability condition and not the high-reliability condition. Thus, our model suggests that qualitatively different cognitive mechanisms underlie individual differences in automation's effects on accuracy in the high-reliability

condition as compared to the low-reliability condition. However, we found that automation benefits to RTs were correlated with excitation for both conditions, but not inhibition, and automation costs to RTs were correlated with inhibition for both conditions, but not excitation. This suggests that individual differences in RTs were driven by the same mechanisms in both conditions.

We also examined correlations between model mechanisms and trust. We found a positive correlation between the excitation parameter and trust in automation in the high-reliability condition, consistent with assumptions in the HAT literature that trust mediates the effect of reliability on automation use (e.g., Lee & See, 2004; Muir & Moray, 1996). This is also consistent with the theoretical concept of excitation developed in this paper, in which humans turn over control to the information input provided by the decision aid so that it can more autonomously trigger their responding, and extends that concept by establishing a cognitive mechanism by which trust can affect decisions in HAT settings.

Practical Implications

We found evidence that higher automation reliability encouraged participants to give automation more autonomy over their decisions. In some operational settings, this could be an adaptive way to free up cognitive capacity for concurrent tasks, at the expense of tasks that automation can, in the vast majority of cases, reliably handle (Kaber, 2018; Moray 2003). Indeed, in HAT contexts the probability that an operator will attend to information (task inputs as conceptualized in our model) will likely depend on the relative expected value and cost of accessing that information (Senders, 1983; Wickens et al., 2015). Further, to the extent that humans integrate decision aids into their decisions via excitation, correct decision aids are likely to speed up RTs.

However, in safety-critical work settings it is desirable that choices only be triggered when the human has independently validated the decision aid advice before responding.

Unfortunately, the outcomes of the current study indicate that this is less likely (a form of automation complacency; Parasuraman & Manzey, 2010) with high-reliability automation because it promotes excitation from decision aid inputs, increasing risk of misuse. In such safety-critical work settings, training, alerts, or warning messages could be implemented to encourage operators to independently verify decision aid advice to minimize misuse.

In operational settings human reliance on automation is often employed as a proxy measure of trust because it is not always possible to interrupt the operator to measure trust subjectively, or experts may not be able or willing to accurately report trust. However, a limitation of using manifest behavioural measures of trust is that it is difficult to separate the effects of trust from changes in strategy resulting from extraneous factors such as workload and time pressure. A key advantage of our model is that it can be used to measure the relative contribution of these factors to human performance (Boag, Strickland, Heathcote, et al., 2019; Boag, Strickland, Loft, et al., 2019), and with the contribution of the current paper it is now possible to measure them alongside the objective effects of trust on decision-making.

The proposed model could potentially function as a "human performance model" (Byrne & Pew, 2009), helping to predict and understand workplace performance in situations where decision makers are increasingly reliant on decision aid advice. It both underscores the importance of interpreting patterns of accuracy and RT in combination, which is necessary to avoid confounding from speed-accuracy tradeoffs, and provides psychological process interpretations of variations in speed and accuracy. For example, our model indicates that degree of benefit to accuracy and RT when automated advice is correct strongly corresponds to the extent that decision aid autonomously triggers decisions (excitation), whereas costs to RT when automated advice is incorrect demonstrates the extent to which automation acts as an advisor that inhibits decisions it disagrees with. Such signatures of the different patterns of automation use could emerge for any task context (e.g., uninhabited vehicle control,

healthcare) in which decision aids are used, although this remains to be empirically tested.

Limitations and Future Directions

An obvious limitation is that we examined the performance of undergraduate participants, rather than experts. Participants completed approximately 4.5 hours of testing, allowing them more potential to develop an efficient approach to their task than in many other prior HAT studies, but it is possible there were still gains to be made with further practice, and even qualitative differences in cognitive strategies with extensive practice. Because experts tend to learn the contexts in which decision aids are likely to be inaccurate and adjust their behaviour (Cohen, 2000; Jamieson & Guerlain, 2000), it is likely that with experience experts learn to adaptively allocate control to decision aids as a function of the appropriateness of task context. Workplace automation, such as aircraft conflict detection decision aids, would be more likely to reliably work under particular conditions (as opposed to randomly failing as in the current study), leading to more accurate operator understanding of automation capabilities and limitations and greater predictability of reliability under various conditions (Balfe et al., 2018; Khastgir et al., 2018). In forming trust in and reliance on automation, experts undoubtedly consider the environment in which automation operates, for example distinguishing between incorrect advice resulting from an unstable environment from inherent problems in decision aid functionality (Muir, 1994). Future work should seek to characterize how automation is incorporated into the performance of experts.

Another limitation is our focus on decisions in a single task environment, rather than a more realistic ATC environment in which controllers contend with multiple stimuli and concurrent goals. This was necessary to develop a detailed evidence accumulation model of the effects of reliability. Notwithstanding the relevance of our earlier discussion that affording decision aids more autonomy would be adaptive to free up cognitive capacity for other concurrent tasks, this was obviously not the motivation for our participants, whom more

likely were interested in using excitation to maintain good accuracy while finishing the experiment reasonably quickly (Hawkins et al., 2012). It is also likely that in a more realistic ATC environment automation would be biased to classify aircraft as conflicts, as missing a potential conflict is more dangerous than miss-classifying a non-conflict (Loft et al., 2009). Future work could apply our model to manipulations of automation bias, to determine how the extent of automation bias affects underlying cognitive mechanisms.

In our task, information for stimulus inputs and decision aids inputs was simultaneously available, whereas in some operational settings stimulus and decision aids inputs can become available to the human at different times (e.g., the automation requires more information or processing time before it provides advice). Previous research has applied systems factorial technology to estimate the effect of decision aids on human processing capacity in tasks with varying decision-aid onset times (e.g., Yamani & McCarley, 2018). Although useful, such an approach does not make the distinctions provided by our model relevant to cognitive control and autonomy (i.e., between excitation, inhibition, and changes in thresholds). Future work could extend our model to situations with varying decision-aid onset times, potentially using approaches that combine evidence accumulation models with systems factorial technology (e.g., Bushmakin et al., 2017; Eidels et al., 2010).

Future work should also seek to situate our findings in the broader context of trust in automation. This study focused on reliability, which is known to affect trust which in turn affects automation reliance (e.g., Dzindolet et al., 2003). Our model appears to provide an objective measure of how variation in trust in automation, driven largely by changes in perceived reliability, affected decision making. However, trust in automation is a more complex construct than perceived reliability, affected by a range of human operator (e.g., affect, personality), automation performance (e.g., reliability, predictability), automation design (e.g., transparency, feedback) and environmental (e.g., workload, decision risk) factors

(Hoff & Bashir, 2015; Schaefer et al., 2016). It is possible that different types of trust in automation have different effects on the types of cognitive processes identified by our model, and future work should investigate this possibility.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462. https://doi.org/10.1207/s15327051hci1204_5

Avril, E., Valéry, B., Navarro, J., Wioland, L., & Cegarra, J. (2021). Effect of imperfect information and action automation on attentional allocation. *International Journal of Human–Computer Interaction*, *37*(11), 1063–1073. https://doi.org/10.1080/10447318.2020.1870817

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency." *Human Performance, Situation Awareness, and Automation: Current Research and Trends*, *1*, 54–59.

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. https://doi.org/10.1080/14639220500535301

Balakrishnan, J. D., MacDonald, J. A., Busemeyer, J. R., & Lin, A. (2002). Dynamic signal detection theory: The next logical step in the evolution of signal detection theory (Technical Report No. 248). Bloomington: Indiana University Cognitive Science Program.

Balfe, N., Sharples, S., & Wilson, J. R. (2018). Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors*, *60*(4), 477–495. https://doi.org/10.1177/0018720818761256

Barg-Walkow, L. H., & Rogers, W. A. (2016). The Effect of Incorrect Reliability Information on Expectations, Perceptions, and Use of Automation. *Human Factors*, *58*(2), 242–260. https://doi.org/10.1177/0018720815610271

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human factors, 59*(6), 881-900. https://doi.org/10.1177/0018720817700258

Bartlett, M. L., & McCarley, J. S. (2021). Ironic efficiency in automation-aided signal detection. *Ergonomics, 64*(1), 103-112. https://doi.org/10.1080/00140139.2020.1809716

Bates, D., Maechler, M., Bolker, B., & Walker., S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach*. Lawrence Erlbaum Associates.

Boag, R., Strickland, L., Heathcote, A., Neal, A., & Loft, S. (2019). Cognitive control and capacity for prospective memory in complex dynamic environments. *Journal of Experimental Psychology: General*, *148*(12), 2181–2206. https://doi.org/10.1037/xge0000599

Boag, R., Strickland, L., Loft, S., & Heathcote, A. (2019). Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition*, *191*, 103974. https://doi.org/10.1016/j.cognition.2019.05.011

Bowden, V. K., Griffiths, N., Strickland, L., & Loft, S. (2021). Detecting a Single Automation Failure: The Impact of Expected (But Not Experienced) Automation Reliability. *Human Factors*, 00187208211037188.

https://doi.org/10.1177/00187208211037188

Bowden, V. K., & Loft, S. (2016). Using memory for prior aircraft events to detect conflicts under conditions of proactive air traffic control and with concurrent task requirements. *Journal of Experimental Psychology: Applied*, *22*(2), 211–224. https://doi.org/10.1037/xap0000085

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Bushmakin, M. A., Eidels, A., & Heathcote, A. (2017). Breaking the rules in perceptual information integration. *Cognitive Psychology*, 95, 1–16. https://doi.org/10.1016/j.cogpsych.2017.03.001

Byrne, M. D., & Pew, R. W. (2009). A History and Primer of Human Performance Modeling. *Reviews of Human Factors and Ergonomics*, *5*(1), 225–263. https://doi.org/10.1518/155723409X448071

Calhoun, G. L., Ruff, H. A., Behymer, K. J., & Frost, E. M. (2018). Human-autonomy teaming interface design considerations for multi-unmanned vehicle control. *Theoretical Issues in Ergonomics Science*, *19*(3), 321–352. https://doi.org/10.1080/1463922X.2017.1315751

Cohen, M. S. (2000). A Situation Specific Model of Trust in Decision Aids. *Proceedings of Human Performance, Situation Awareness & Automation: User-Centered Design for the New Millennium*, 143–148.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The

role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, *13*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review, 17*(6), 763–771. https://doi.org/10.3758/PBR.17.6.763

Ferraro, J., Clark, L., Christy, N., & Mouloua, M. (2018). Effects of automation reliability and trust on system monitoring performance in simulated flight tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1232–1236. https://doi.org/10.1177/1541931218621283

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

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review*, *19*(2), 339–348. https://doi.org/10.3758/s13423-012-0216-z

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (2019). Dynamic models of choice. *Behavior Research Methods*, *51*(2), 961–985. https://doi.org/10.3758/s13428-018-1067-y

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*. https://www.frontiersin.org/article/10.3389/fnins.2014.00150

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hussein, A., Elsawah, S., & Abbass, H. A. (2020a). The reliability and transparency bases of trust in human-swarm interaction: Principles and implications. *Ergonomics*, *63*(9), 1116–1132. https://doi.org/10.1080/00140139.2020.1764112

Hussein, A., Elsawah, S., & Abbass, H. A. (2020b). Trust Mediating Reliability-Reliance Relationship in Supervisory Control of Human-Swarm Interactions. *Human Factors, 62*(8), 1237-1248. https://doi.org/10.1177/0018720819879273

Hutchinson, J., Strickland, L., Farrell, S., & Loft, S. (2022). The Perception of Automation Reliability and Acceptance of Automated Advice. *Human Factors*, 00187208211062985. https://doi.org/10.1177/00187208211062985

Jamieson, G., & Guerlain, S. (2000). *Operator interaction with model-based predictive controllers in petrochemical refining*. 172–177.

Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 7–24. https://doi.org/10.1177/1555343417737203

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies*, *96*, 290–303. https://doi.org/10.1016/j.trc.2018.07.001

Law, D. J., Pellegrino, J. W., Mitchell, S. R., Fischer, S. C., McDonald, T. P., & Hunt, E. B. (1993). Perceptual and cognitive factors governing performance in comparative arrival-time judgments. *Journal of Experimental Psychology: Human Perception and*

Performance, *19*(6), 1183–1199. https://doi.org/10.1037/0096-1523.19.6.1183

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package 'emmeans.'*

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive Research: Principles and Implications*, *7*(1), 1–21. https://doi.org/10.1186/s41235-022-00364-y

Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., & Bell, J. (2021). The impact of transparency and decision risk on human–automation teaming outcomes. *Human Factors*, 00187208211033445. https://doi.org/10.1177/00187208211033445

Loft, S., Bolland, S., Humphreys, M. S., & Neal, A. (2009). A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *Journal of Experimental Psychology: Applied*, *15*(2), 106–124. https://doi.org/10.1037/a0016118

Loft, S., Neal, A., & Humphreys, M. (2007). The development of a general associative learning account of skill acquisition in a relative arrival-time judgement task. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 938-959.

Ly, A., Boehm, U., Heathcote, A., Forstmann, B. U., Marsman, M., & Matzke, D. (2017). A flexible and efficient hierarchical bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 467–480). Wiley Blackwell.

Ly, A., Marsman, M., & Wagenmakers, E. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13. https://doi.org/10.1111/stan.12111

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology, 12*: 589585.

https://doi.org/10.3389/fpsyg.2021.589585

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Meyer, J., & Lee, J. D. (2013). Trust, reliance, and compliance. *The Oxford Handbook of Cognitive Engineering.*, 109–124. https://doi.org/10.1093/oxfordhb/9780199757183.001.0001

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *Selected Papers from the Second Cyberspace Conference on Ergonomi Cs, CybErg 1999*, *31*(3), 175–178. https://doi.org/10.1016/S0169-8141(02)00194-4

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science*, *1*(4), 354–365. https://doi.org/10.1080/14639220052399159

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

Musialek, B., Munafo, C. F., Ryan, H., & Paglione, M. (2010). Literature survey of trajectory predictor technology. *Federal Aviation Administration, William J. Hughes Technical Center, Tech. Rep*, 1–31.

National Transportation Safety Board. (2014). *Descent Below Visual Glidepath and Impact With Seawall, Asiana Airlines Flight 214, Boeing 777-200ER, HL7742, San*

Francisco, California, July 6, 2013 (Aircraft Accident Report NTSB/AAR-14/01). https://www.ntsb.gov/investigations/accidentreports/reports/aar1401.pdf

Noskievič, T., & Kraus, J. (2017). Air traffic control tools assessment. *MAD-Magazine of Aviation Development*, *5*(2), 6–10. https://doi.org/10.14311/MAD.2017.02.01

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

Pachella, R. G., & Pew, R. W. (1968). Speed-Accuracy Tradeoff in Reaction Time: Effect of Discrete Criterion Times. *Journal of Experimental Psychology*, *76*(1), 19–24. https://doi.org/10.1037/h0021275

Palada, H., Neal, A., Vuckovic, A., Martin, R., Samuels, K., & Heathcote, A. (2016). Evidence accumulation in a complex task: Making choices about concurrent multiattribute stimuli under time pressure. *Journal of Experimental Psychology: Applied*, *22*(1), 1. https://doi.org/10.1037/xap0000074

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226–1243. https://doi.org/10.1037/a0036801

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Rouder, J., Kumar, A., & Haaf, J. M. (2019). *Why most studies of individual differences with inhibition tasks are bound to fail*. https://doi.org/10.31234/osf.io/3cjr5

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors*, *49*(1), 76–87. ProQuest One Academic; SciTech Premium Collection. https://doi.org/10.1518/001872007779598082

Samms, C. (2010). Improved Performance Research Integration Tool (IMPRINT): Human Performance Modeling for Improved System Design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(7), 624–625. https://doi.org/10.1177/154193121005400701

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Sebok, A., & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human Factors*, *59*(2), 189–203. https://doi.org/10.1177/0018720816665201

Senders, J. W. (1983). *Visual sampling processes*. Lawrence Erlbaum.

Shah, S. J., & Bliss, J. P. (2017). Does Accountability and an Automation Decision Aid's Reliability Affect Human Performance in a Visual Search Task? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 183–187. https://doi.org/10.1177/1541931213601530

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Strickland, L., Heathcote, A., Bowden, V. K., Boag, R. J., Wilson, M. D., Khan, S., & Loft, S. (2021). Inhibitory cognitive control allows automated advice to improve accuracy while minimizing misuse. *Psychological Science*, *32*(11), 1768–1781. https://doi.org/10.1177/09567976211012676

Strickland, L., Loft, S., Remington, R. W., & Heathcote, A. (2018). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*, *125*(6), 851–887. https://doi.org/10.1037/rev0000113

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. P*sychological Science, 1*, 46-54.

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. https://doi.org/10.1080/14639220500370105

Wickens, C. D., Sebok, A., Li, H., Sarter, N., & Gacy, A. M. (2015). Using modeling and simulation to predict operator performance and automation-induced complacency with robotic automation: A case study and empirical validation. *Human Factors*, *57*(6), 959–975. https://doi.org/10.1177/0018720814566454

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. https://doi.org/10.1080/14639220110110306

Xu, X., & Rantanen, E. M. (2003, April). Conflict detection in air traffic control: A task analysis, a literature review, and a need for further research. *In Proceedings of the 12th*

International Symposium on Aviation Psychology (pp. 1289-1295). Wright State University Press Dayton, OH.

Yamani, Y., & McCarley, J. S. (2018). Effects of task difficulty and display format on automation usage strategy: A workload capacity analysis. *Human factors, 60*(4), 527-537. https://doi.org/10.1177/0018720818759356