Inhibitory cognitive control allows automated advice to improve accuracy while minimizing misuse

Luke Strickland[1], Andrew Heathcote[2,3], Vanessa Bowden[4], Russell J. Boag[5], Michael David Wilson[1], Samha Khan[2] & Shayne Loft[2]

[1] The Future of Work Institute,

Curtin University, Australia

[2] The School of Psychology,

The University of Tasmania, Australia

[3] The School of Psychology,

The University of Newcastle, Australia

[4] The School of Psychological Science,

The University of Western Australia, Australia

[4] Department of Psychology,

The University of Amsterdam, The Netherlands

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

This research was in part supported by an Australian Research Council Discovery Grant (DP160100575) awarded to Loft. The data and code associated with the manuscript are available at: https://github.com/lukestrickland/Automation_LBA

Abstract

Humans increasingly use automated decision aids. However, environmental uncertainty means that automated advice can be incorrect, creating the potential for humans to action incorrect advice or to disregard correct advice. We present a quantitative model of the cognitive process by which humans use automation when deciding whether aircraft would violate minimum separation. The model closely fitted the performance of twenty-four participants, whom each made 2400 conflict detection decisions (conflict vs non-conflict), either manually (with no assistance) or with the assistance of 90% reliable automation. When the decision aid was correct, conflict detection accuracy improved, but when the decision aid was incorrect, accuracy and response time were impaired. The model indicated that participants integrated advice into their decision process by inhibiting evidence accumulation toward the task response incongruent with that advice, thereby ensuring that decisions could not be made solely on automated advice without first sampling information from the task environment.

Statement of Relevance

In modern workplaces and industries such as aviation and healthcare, automated decision aids that recommend actions to humans are increasingly prevalent. However, automated advice can be wrong, creating the potential for humans to action incorrect advice or to disregard correct advice, with potentially catastrophic consequences. While prior research has identified several factors that impact how people use automation, there are currently no quantitative theories of the underlying cognitive mechanisms. We present an evidence-accumulation model of the cognitive process by which humans use automated advice to decide whether aircraft would violate minimum separation. The model provided a good fit to the observed effects (i.e., increased participant accuracy with correct advice, decreased accuracy/increased response time with incorrect advice) and indicated that advice from the decision aid was used to inhibit the task response incongruent with that advice. The model provides a tractable, quantitative theoretical framework for understanding how humans use automated advice.

Human interaction with automation has been a major area of inquiry for over 40 years (Bainbridge, 1983; Wiener & Curry, 1980). Decision aids that recommend actions to human operators are increasingly prevalent: in healthcare, decision aids support diagnoses and provide treatment advice; in air traffic control (ATC), decision aids advise controllers how to maintain aircraft separation; in defence, decision aids recommend how to coordinate unmanned vehicles; and in airports, decision aids support luggage inspection.

Decision aids benefit performance and reduce workload (Onnasch, Wickens, Li, & Manzey, 2014). However, the uncertainty inherent in complex work systems means automated advice can be incorrect. This creates the potential to erroneously action incorrect advice (*misuse*[1]; Lee & See, 2004), or disregard correct advice (*disuse*), although this is less likely with reliable automation (Wickens & Dixon, 2007). Automation misuse by experts can occur in work domains where incorrect decisions risk serious consequences, including aviation, healthcare, and process control.

Research has identified several design (e.g., reliability, transparency) and environmental (e.g., task demands, uncertainty) factors that impact automation use (Endsley, 2017; Parasuraman & Manzey, 2010). However, extant research has relied largely on verbal psychological theories, and tests of partial performance measures such as accuracy and mean response time (RT), with little progress towards quantitative theories of cognitive mechanisms. Quantitative human performance modelling is advantageous because it can provide theoretical insights, unify interpretation of seemingly disparate data, refine predictions, and predict performance when human in-the-loop testing is not feasible (Byrne & Pew, 2009; Farrell & Lewandowsky, 2010).

[1] As defined, "misuse" and "disuse" can occur even when the operator is using automation optimally, e.g., always following the recommendation of a decision aid that has a better chance of being right than they do, as if the recommendation happens to be wrong on a particular occasion that would be classified as misuse. However, such optimality considerations are not the focus of the current paper.

In the current paper, we apply a quantitative model of decision making (Boag, Strickland, Heathcote, Neal & Loft, 2019) to automation use in a simulated ATC conflict detection task. Conflict detection requires humans to decide whether aircraft will violate minimum separation standards in the future based on their altitude, speed, and relative distance from intersection (Loft, Bolland, Humphreys, & Neal, 2009). Decision aids are increasingly used to maximise airspace capacity. Conflict detection is particularly suitable to examine the cognitive dynamics of human-automation interaction, as decision uncertainty can trade-off with temporal pressure (Loft, Sanderson, Neal, & Mooij, 2007), and it is representative of other work contexts where operators make judgments about moving objects on displays (e.g., unmanned vehicle control, maritime surveillance).

A Model of Human Adaptation to Automation

Evidence accumulation models are the most successful class of model for understanding speeded binary choice decisions (Ratcliff & Smith, 2004). We model conflict detection with the linear ballistic accumulator (LBA; Brown & Heathcote, 2008), an evidence accumulation model where evidence for each possible decision accrues linearly and independently, with the first accumulator to reach threshold determining the decision made (Figure 1).

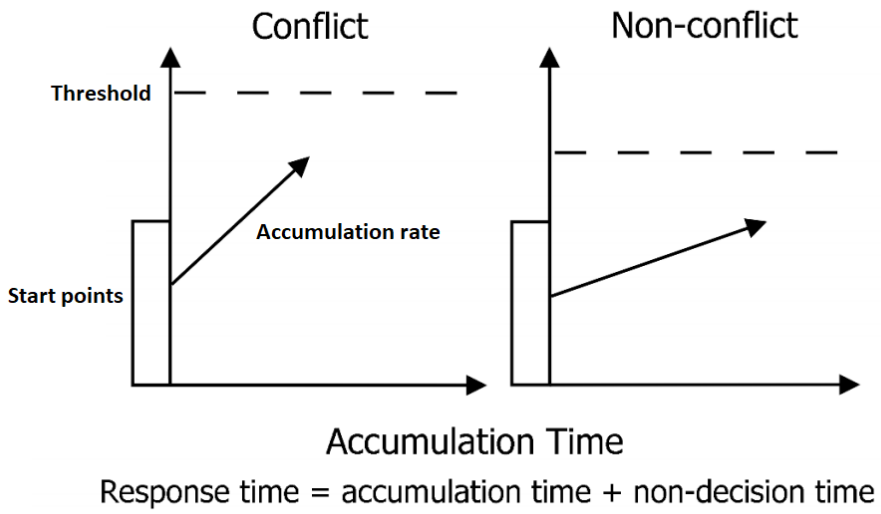

Figure 1. LBA model of conflict detection. Evidence for each accumulator begins at a start point, drawn from a uniform distribution, and increases at an accumulation rate, which is drawn from a normal distribution. The first accumulator to reach threshold determines the observed response.

The LBA provides estimates of the cognitive processes underlying observed performance. *Accumulation rates* index how fast evidence accrues towards each response. They are classified as either *matching* (i.e., accumulation toward the correct response, e.g., responding 'conflict' when aircraft are in conflict), or *mismatching* (i.e., accumulation toward the incorrect response). *Thresholds* index the amount of evidence required for each decision and reflect cognitive control. For example, raising thresholds increases accuracy, but at the cost of slower responding.

In the model, decision aids provide inputs to the decision process (automation inputs), integrated with task information (stimulus inputs) in a "feedforward" manner (Figure 2; see Boag et al., 2019; Strickland, Loft, Remington & Heathcote, 2018). Accumulation rates are simultaneously increased by excitation from stimulus inputs that match the response and decreased by inhibition from mismatching inputs, an assumption consistent with the finding that biased information affects decision making via a constant effect on rates (Hanks,

Mazurek, Kiani, Hopp & Shadlen, 2011). For example, processing stimulus inputs that match a conflict response (e.g., close predicted relative arrival times of aircraft; Loft, Neal, & Humphreys, 2007) excites the conflict accumulator and inhibits the non-conflict accumulator. Similarly, if a decision aid recommends a conflict decision, it excites the conflict accumulator and inhibits the non-conflict accumulator. The final accumulation rate is determined by summing the inhibition and excitation provided from both automation and stimulus inputs.

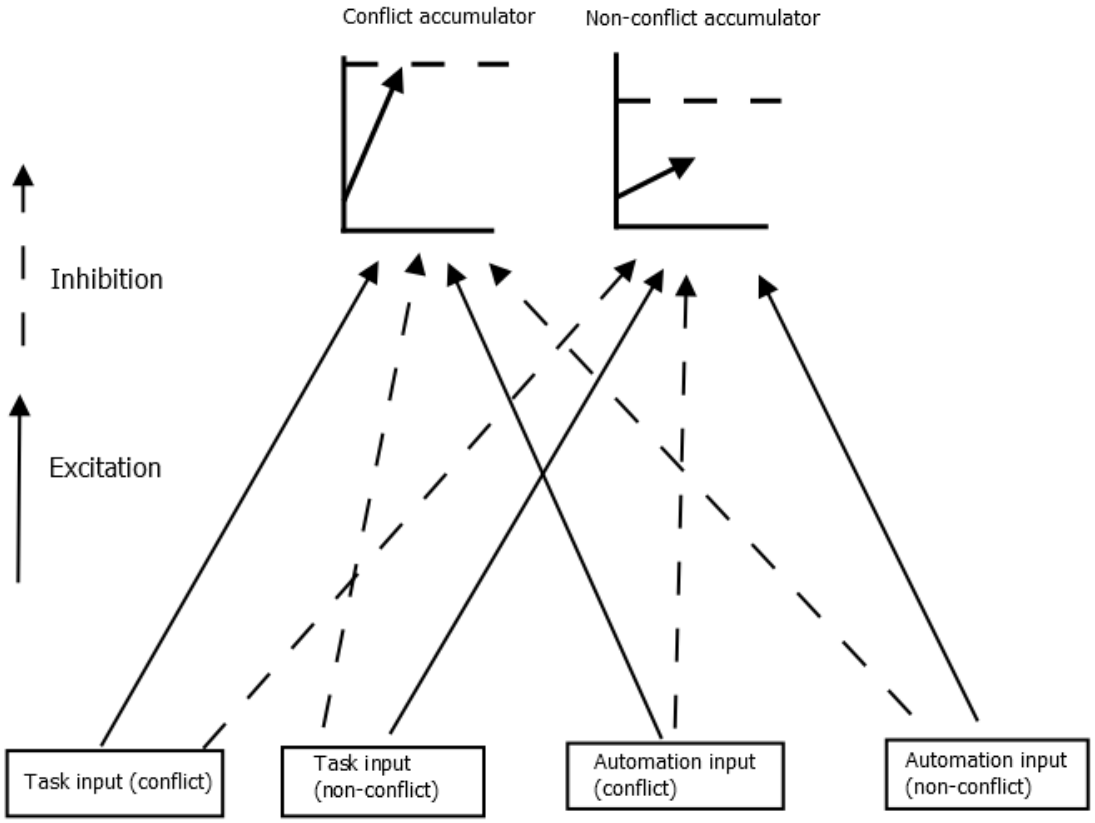

Figure 2. Impacts of automation inputs and stimulus inputs on evidence accumulation. Both stimulus inputs and automation inputs can potentially excite corresponding accumulators (solid lines), increasing accumulation rates, but also inhibit the opposing accumulator (dashed lines), decreasing accumulation rates.

The model can answer fundamental questions about automation use. Decision aids are typically provided with the expectation that humans will not solely rely on them. In model terms, over-reliance on automation occurs if excitation from the decision aid causes enough

accumulation to trigger a response without requiring excitation from stimulus inputs. Given the risks of over-reliance, humans may be reluctant to purely use automation excitation when accuracy is important. By contrast, inhibition of decisions that mismatch the decision aid cannot trigger a response without excitation from stimulus inputs but can increase the accuracy of decisions when automation inputs are correct. Thus, when automation is not perfectly reliable, inhibition may be preferable to encourage appropriate automation use while minimising misuse. In addition, humans might adjust response thresholds. For instance, when automation is less than 100% reliable, they could increase response thresholds to minimize noise in their own decision-making.

The Current Study

We use the model to quantify the excitation, inhibition, and threshold control underlying performance in an ATC conflict detection task with an automated decision aid. The automation had 90% accuracy, and based on Boag et al. (2019), the task parameters were expected to produce around 90% human manual accuracy, allowing us to study an approximately equal ability human-automation team.

There were two within-subjects conditions: “manual”, in which participants detected conflicts unassisted, and “automation”, in which a decision aid recommended a response. Participants were informed the automation was not perfectly reliable and were encouraged, with scoring and a financial incentive, to avoid complete reliance on automation. We expected that when the automation was correct, it would benefit accuracy. Thus, on “automation correct” trials, accuracy should be higher than for manual trials. By contrast, automation should impair performance when recommending an incorrect decision, decreasing accuracy relative to manual trials. RT could also be affected by automation. If automation causes excitation, responses that automation recommends should be faster compared with manual trials. If automation causes inhibition, non-recommended decisions should be slower

compared with manual trials.

Method

Participants

To ensure adequate measurement for cognitive modelling — where reliable inference depends on the number of trials per participant, rather than the number of participants (Smith & Little, 2018) — the experiment ran over two days, yielding 2400 total trials for each participant. We tested 27 participants, with the data from three excluded (one participant received their conditions in the wrong order; one had chance level performance during manual trials; one due to computer error). For the remaining 24 (14 female, 10 male), the mean age was 22.29 (range: 18-37). Participants were from a convenience sample of psychology students, and the University of Western Australia's community research pool. Participants received either course credit or $40 AUD. For all participants, an additional reward between $0 and $20 AUD was provided based on performance. The study was approved by the University of Western Australia's Human Research Ethics.

Design

Participants completed two sessions, each lasting approximately 90 minutes. In each, participants had one block with automated advice and one without (manual), with condition order counterbalanced across days. Each block contained 600 trials. The conflict detection task included two possible response key assignments, counterbalanced across participants, either 'f' for conflicts and 'j' for non-conflicts, or 'j' for conflicts and 'f' for non-conflicts.

Materials

ATC Conflict Detection Task. The conflict detection task (Fothergill, Loft & Neal, 2009) has previously been used to test expert controllers (Loft et al., 2009), balancing representativeness with experimental control. Figure 3 depicts the display. The sector was 180 nautical miles (nmi) by 112.5nmi. At the start of each trial, two aircraft appeared within

the circular light grey sector and flew on straight paths towards the intersection. Adjacent to each aircraft was a data block containing: the aircraft callsign (e.g., TVU740) and type (e.g., B737), flight level (e.g., 370 indicates 37,000 feet), and speed in knots (nmi per hour) divided by 10. A 10nmi by 20nmi scale was included on the left of the display, and a probe vector was attached to aircraft indicating their heading and predicted position in one minute.

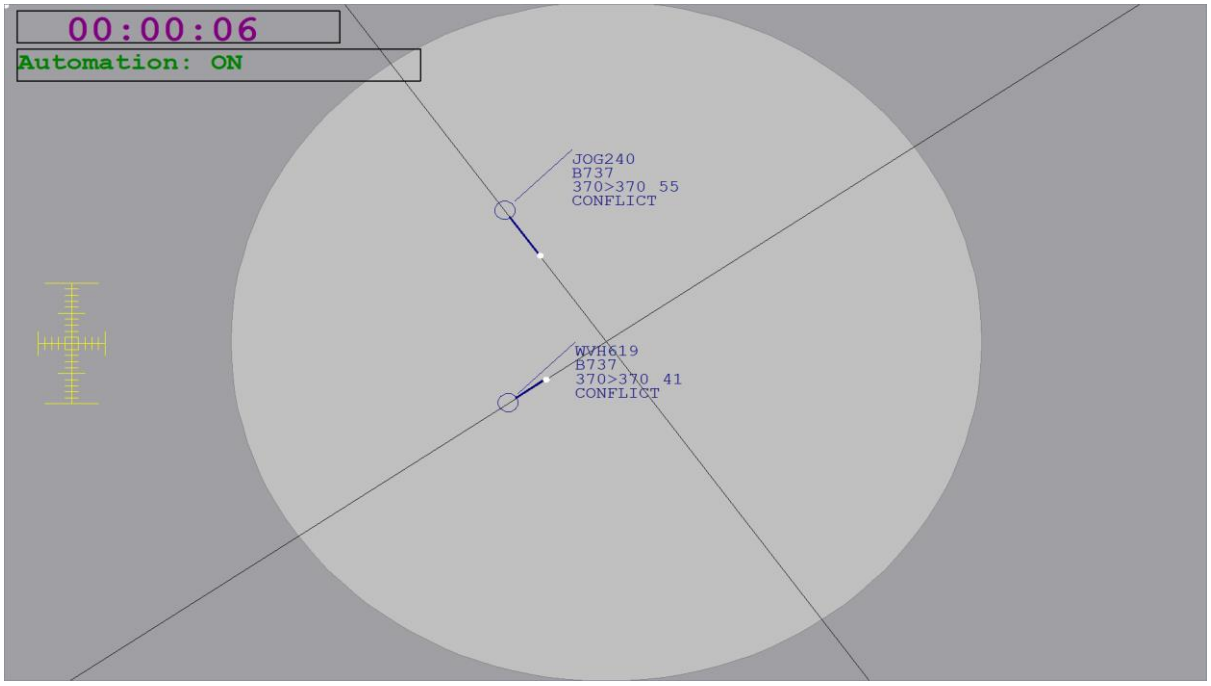

Figure 3. An example of the air traffic control simulator display. Next to each aircraft callsign (e.g., TVU740), aircraft type (e.g., B737), current and cleared altitude (e.g., 370 >370), and airspeed (e.g., 53) are displayed. Note that airspeed only shows two digits (e.g., 530 knots would be '530'). A trial countdown timer is displayed on the top left. The automated advice is placed under the data block of each aircraft. For example, in this screen capture, the decision aid is recommending a 'conflict' decision. In manual conditions, this was replaced with a string that had no special meaning, '########'.

Conflict detection stimuli. Participants judged whether aircraft pairs would conflict in the future (they had no control over aircraft). Aircraft pairs were in conflict if they would simultaneously travel within 5nmi (laterally) and 1,000ft (vertically). Altitude for all aircraft was fixed at 37,000 feet, and thus decisions were based on lateral separation. The spatial variables defining the aircraft pairs are included in Table 1. Conflict and non-conflict status

were created using the lateral distances of minimum separation (d_{min}). For conflict stimuli, d_{min} was drawn from uniform distribution U [0,1.5] nmi. For non-conflict stimuli, d_{min} was drawn from uniform distribution U [8.5,10] nmi. The angle of approach of one of the aircraft was randomly sampled between 0 and 360 degrees. The relative angle of approach between aircraft was fixed at 90 degrees. Other features were varied randomly to avoid instance-based learning (Bowden & Loft, 2016). Aircraft speeds were fixed randomly between 400 and 700 knots, and time to minimum separation between 120 and 210 seconds. On half the trials, the faster aircraft reached the intersection first, and on the other half the slower aircraft did.

Table 1

Ranges of the spatial variables defining aircraft stimuli. Distance of minimum separation is referred to as d_{min}, and time to minimum separation referred to as t_{min}.

Spatial variable	Lower	Upper	Units
d_{min} (conflicts)	0	1.5	nmi
d_{min} (non-conflicts)	8.5	10	nmi
Airspeed	400	700	knots
Direction of approach	0	360	degrees
t_{min}	120	210	seconds
First Pass Aircraft	0	1	0 = fastest, 1 = slowest

Automated decision aid. During training, participants performed the task with no decision aid. In automation conditions, advice was placed under the data block of each aircraft (Figure 3). The advice read 'CONFLICT', to recommend classifying the pair of aircraft as in conflict, or 'NON-CONF' (non-conflict). In manual conditions, a string

‘########’ was placed under the data block of each aircraft, with no special meaning.

Participants completed 2400 trials. In the automation condition, the decision aid failed on one randomly selected trial out of every 10, and equally often for conflicts and non-conflicts. This resulted in 60 automation failures on conflict trials (30 each session), and 60 on non-conflict trials. The non-failure stimuli in the automated condition for each day were matched to the manual condition in terms of aircraft speeds and distances from intersection. To minimize learning across matched aircraft pairs, presentation order and angle of approach were randomized for each condition (whilst maintaining a relative angle of 90 degrees between aircraft), and different callsigns were assigned. The automation failure stimuli in the automated condition were presented in the same trial positions as in the manual conditions. This provided 60 conflict and non-conflict ‘matched manual’ trials that had matched stimulus properties and trial positions across automated and manual blocks.

Automation Trust Questionnaire. After completing the experiment, participants rated their trust in the decision aid (supplementary materials). Participants rated six trust questions on a five-point scale from ‘strongly disagree’ to ‘strongly agree’.

Procedure

Experimental Procedure. Participants provided informed consent and then viewed training instructions followed by a demonstration that showed aircraft pairs with different d_{min} to help participants gauge whether aircraft pairs were in conflict. Subsequently, participants completed 40 training trials. After training, participants completed their first experimental block (either manual or automation). A financial reward was associated with accuracy, ranging from \$0 to \$20 AUD. The maximum reward for each block was \$5. In the manual condition, rewards were calculated by $\text{reward} = \frac{5}{600} \times (N_{\text{correct}} - N_{\text{incorrect}} - N_{\text{nonresponses}})$. In the automated condition, rewards were calculated by $\text{reward} = \frac{5}{600} \times (N_{\text{correct}} - 9 \times N_{\text{incorrect}_{\text{accept}}} - N_{\text{incorrect}_{\text{reject}}} - N_{\text{non-responses}})$. The cost of incorrectly

accepting automation was highly weighted so that participants could not receive a substantial reward for relying primarily on the automation. Participants were not informed of the precise weightings of the reward scheme, but were instructed that: "*Although the automation is highly reliable, it is not perfect. 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. Accepting the automated recommendation when it is wrong will result in a substantially greater reduction in your performance score and subsequent bonus."* In manual conditions, participants were informed that there was no automated advice, and instead just a string '########' which they should ignore. They were also instructed that "*Incorrect responses will reduce your performance score and subsequent bonus.*"

Participants took self-paced breaks between each block and also mid-block. After each block, they were presented accuracy feedback for that block. In the automation blocks, their feedback was broken down into the percentage of trials on which they incorrectly disagreed with the decision aid and the percentage of trials on which they incorrectly agreed with the decision aid. The end of the first block of each session was followed by instructions for the subsequent block. Participants returned for session two within 10 days of session one. The procedure for session two was the same as session one, except that session two did not include a demonstration of d_{min}, and after session two participants completed the trust questionnaire.

Trial Procedure. Trials began with an aircraft pair heading towards a common intersection. Participants had 8 seconds to respond. A trial completed once the participant responded, or once 8 seconds had elapsed. If participants submitted a correct response, the next trial began. If they submitted an incorrect response, or did not respond, then they received feedback. The feedback informed participants they were incorrect, and which

decision would have been correct (e.g. "*Incorrect! This pair was in conflict*"). Participants clicked an 'ok' button to proceed to the next trial.

Results

Trials were excluded from analysis if participants failed to respond (0.17% of trials) or responded very quickly (< 0.2 seconds; 0.03% of trials). We report analysis of accuracy and correct RTs with linear mixed effects models. We examined four factors – stimulus type (conflict/non-conflict), condition (automated /manual), automation accuracy (correct/ incorrect), and session (one/two). For manual conditions, 'automation incorrect trials' refer to trials that were matched to 'automation incorrect trials' in the automation condition; and 'automation correct trials' refer to the other manual trials. To examine accuracies, we fitted a generalized linear model with a probit link to response accuracy on every trial (either 0 or 1). To examine RTs, we fit a linear mixed effects model to the mean correct RTs of each participant. Significance tests of the factors in each model are tabulated in supplementary materials, as are follow-up contrasts. Our aim was to identify strong effects. Thus, our significance criterion was set at $p < .005$ (Benjamin et al., 2018). The standard errors reported in text use the Morey (2008) bias-corrected method for within-subjects designs.

Accuracy

Summaries of participant accuracies and mean RTs are displayed in Figure 4. There were main effects of stimulus type, session, condition, and automation accuracy on participant accuracy. Responses were slightly more accurate on conflict trials ($M = 0.87$, $SE = 0.04$) than non-conflict trials ($M = 0.86$, $SE = 0.04$). Accuracy was lower on session one ($M = 0.84$, $SE = 0.04$) than session two ($M = 0.88$, $SE = 0.03$). Condition and automation accuracy interacted. Accuracy was higher for trials where participants were provided correct automation ($M = 0.94$, $SE = 0.02$) than matched manual trials ($M = 0.89$, $SE = 0.01$), and lower when participants were provided incorrect automation ($M = 0.73$, $SE = 0.01$), than for

matched manual trials (M = 0.89, SE = 0.01). This suggests that participants used automation to their advantage, but when automation failed, it imposed a cost. However, importantly, accuracy on automation incorrect trials was far from floor, suggesting that participants did not rely on the automation entirely.

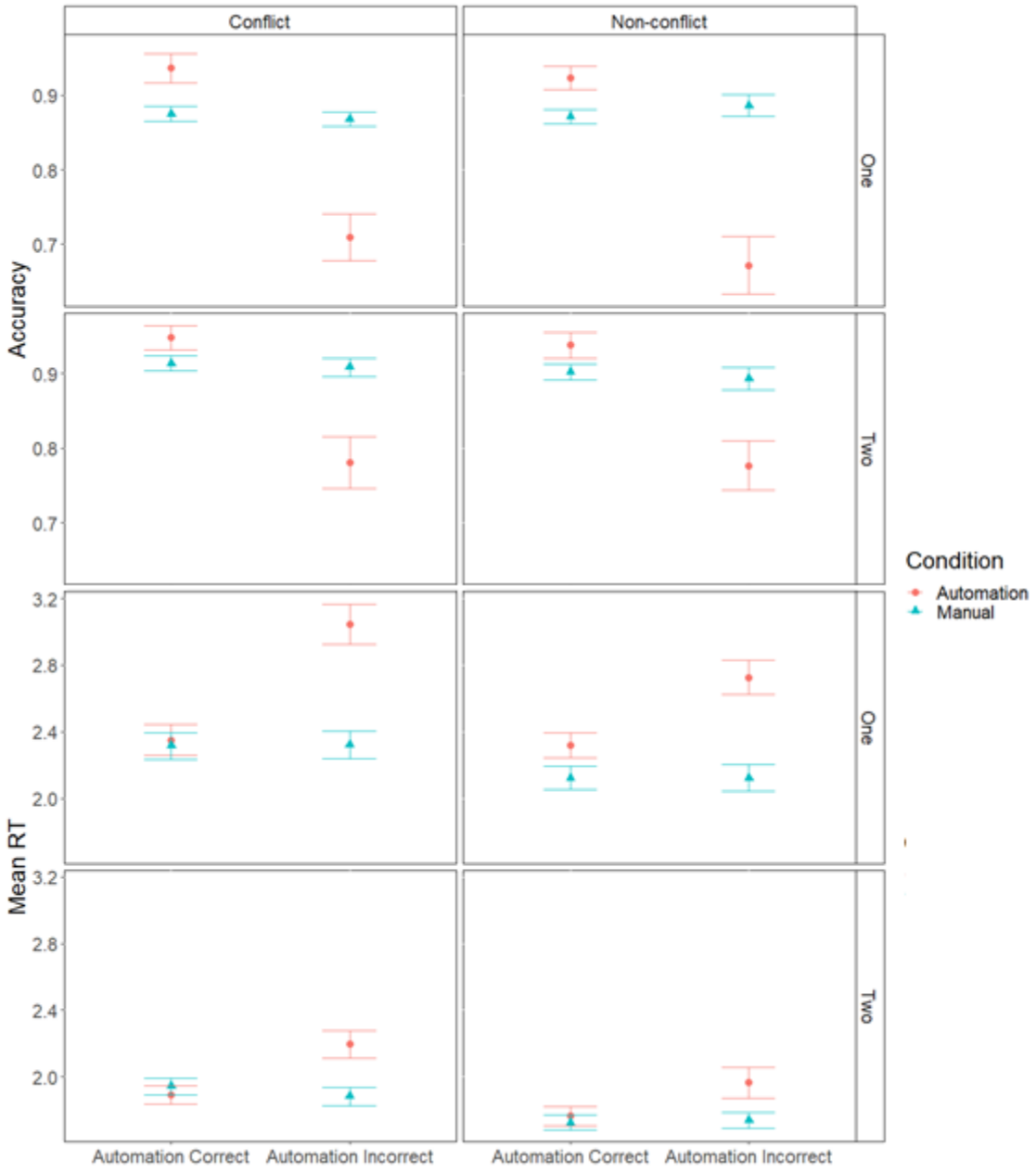

Figure 4. Summaries of performance. Each panel corresponds to one stimulus type, on one experimental session, for either participant response accuracy or mean RT. The error bars included were calculated using the Morey (2008) bias-corrected method for within-subjects error bars.

Response Times

There were main effects of stimulus type, session, condition, and automation accuracy on mean correct RTs. Correct responses were slower on conflict trials (M = 2.24, SE = 0.15) than non-conflict trials (M = 2.06, SE = 0.14). Condition and automation accuracy interacted. Correct RTs were similar on trials where participants were provided correct automation (M = 2.08 SE = 0.1) compared with matched manual trials (M = 2.03 SE = 0.09). By contrast, they were much slower on automation-incorrect trials (M = 2.48 SE = 0.15) than matched manual trials (M = 2.02 SE = 0.09). Condition and session interacted. In session one, RTs were slower in the automation condition (M = 2.61, SE = 0.13) than the manual condition (M = 2.22, SE = 0.09). In session two, RTs were slower in the automation condition (M = 1.95, SE = 0.09) than the manual condition (M = 1.82, SE = 0.06), though the magnitude of the difference was attenuated relative to session 1, and failed to reach the $p < .005$ threshold.

We conducted exploratory analyses of the correlations between the costs and benefits of automation. We examined whether the accuracy advantage provided by correct automation (*correct automation trial accuracy - matched manual trial accuracy*) was associated with the cost of automation on failure trials to accuracy (*matched failure trial accuracy - automation failure trial accuracy*) or RT (*automation failure trial correct RT - matched manual correct RT*). The increased accuracy provided by correct automation was positively correlated with the accuracy cost of automation on automation-incorrect trials, r (22) = 0.79, $p < .001$. The accuracy increase on automation-correct trials was positively correlated with the correct RT increase on automation-failure trials, r (22) = 0.42, p = .04, although this did not reach significance at $p < .005$. We report correlations between the three above measures and automation trust in the supplementary materials, although no significant associations were found.

In summary, when the decision aid was correct, participant accuracy was higher compared to manual trials. By contrast, when it was incorrect, accuracy was lower compared to manual trials. When the decision aid was correct, RTs were not impacted, but when it was incorrect RTs were slower compared to manual trials. These effects generally support an inhibition account of decision aid use, in which the response incongruent with the decision aid advice is inhibited. In the next section, we present LBA modelling to formally measure latent processes such as inhibition.

Model Results

Model Specification

We applied a two accumulator LBA (Figure 1), with each conflict and non-conflict decision assigned an accumulator. Evidence for each accumulator begins at a start point independently drawn from U [0, A]. It then accumulates linearly at a rate drawn from a normal distribution N (v, sv) truncated at 0, until one accumulator reaches its threshold b, determining the response. We estimate thresholds in terms of the positive quantity $B = b – A$. Total RT is determined by decision time plus non-decision time (i.e., the time to encode the stimulus and produce a motor response). To facilitate estimation, we only allowed one A parameter and one non-decision time parameter for each participant. The variability in mismatching accumulation rates was fixed at 1 as a scaling parameter. One sv parameter indexing variability in matching accumulation rates was estimated for each participant. Mean accumulation rates could vary by stimulus type, condition, and automation accuracy. We estimated separate thresholds for each accumulator, experimental condition, and session. Thresholds did not vary across stimulus type to avoid circularity (if the stimulus type were known, there would be no point in the decision process). However, a reviewer suggested participants might initially process the aid's advice without processing stimuli inputs, leading to an initial one step-change in evidence, which would be mathematically equivalent to a

threshold change. In supplemental materials, we explore a model that allows thresholds to adapt in response to the automation's recommendation. However, we did not find support for this mechanism, and thus we report the simpler model below.

Model Fit

We performed Bayesian parameter estimation using the Dynamic Models of Choice R Suite (Heathcote et al., 2019; see supplementary materials). This provides posterior samples estimating probability distributions of the model parameters. Figure 5 displays fits of the model to the data. The model closely fit the data, including the effects of automation on accuracy and RT.

Parameter Inference

For inference we created a group posterior distribution, by averaging the values of each posterior sample across participants. The values of the averaged model parameters are tabulated in supplementary materials. In the following sections, we examine the effect of automation on accumulation rate and threshold parameters. To test parameter differences, we calculate a one-tailed posterior *p*, corresponding to the proportion of posterior samples on which one parameter value was higher than another. We report the *p* value against whichever direction was closest to an observed effect (i.e., a *p* near 0 is evidence in favor of an effect). Many effects were 'significant' in the sense that $p < .005$. To estimate effect size, we report the mean of the parameter differences divided by the standard deviation, referred to as *Z*.

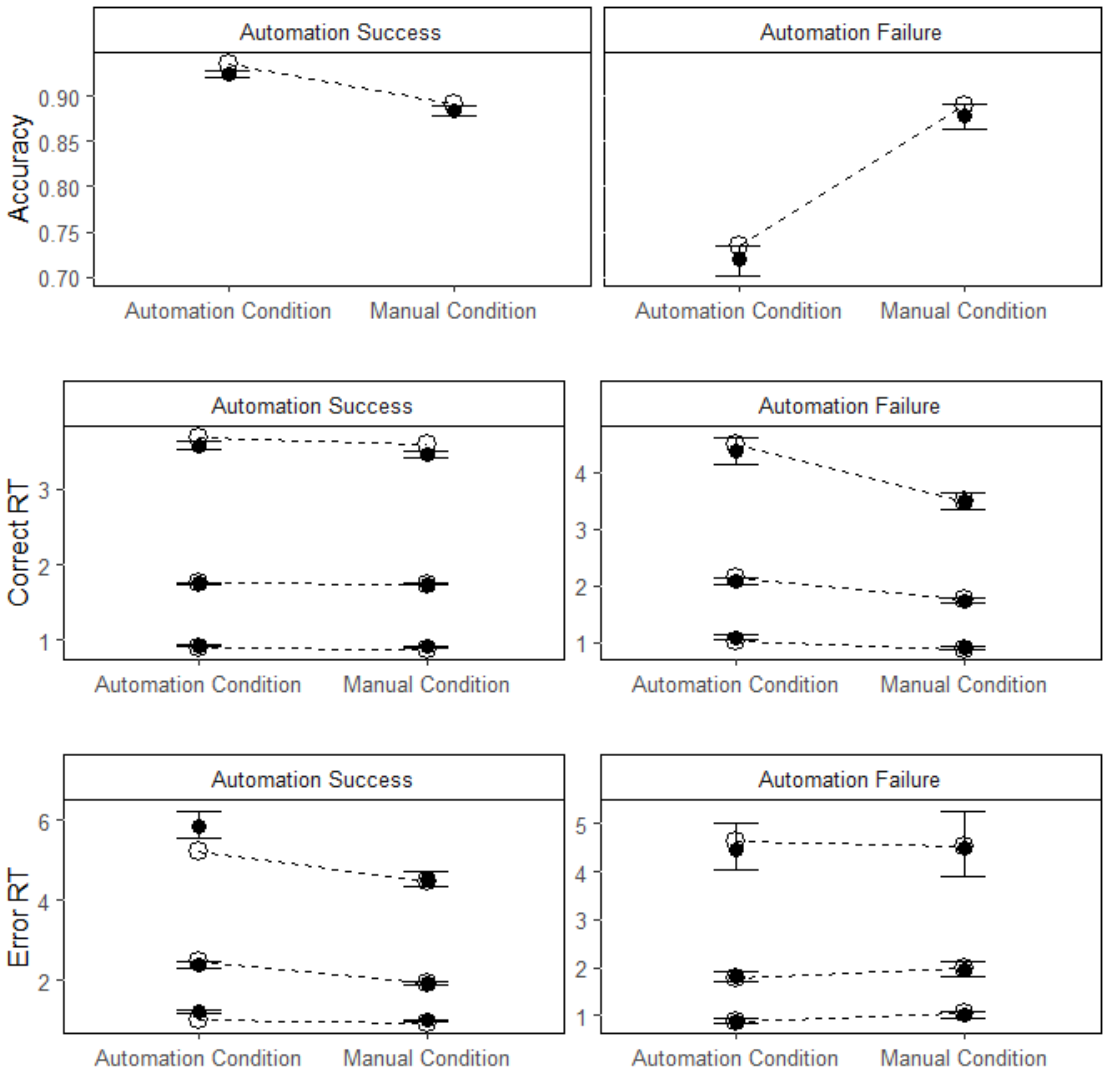

Figure 5. Posterior predictions of group performance. The model predictions correspond to the white circles, the posterior means correspond to the black shaded dots. The error bars display the 95% posterior credible intervals of the predictions. Three quantiles of response time (RT) are depicted, with 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.

Excitation and Inhibition

Accumulation rates are plotted in Figure 6. We compared accumulation rates on automation trials with accumulation rates on matched manual trials. Excitation is indicated by increased accumulation towards the accumulator that agrees with the decision aid (i.e.,

match). For example, on a conflict trial on which the automation recommends “conflict”, excitation would increase the conflict accumulation rate. Inhibition is indicated by reduced accumulation towards the accumulator that disagrees with the decision aid (i.e., mismatch). For example, for conflict trials on which the decision aid recommends “conflict”, inhibition would reduce accumulation in the non-conflict accumulator.

Table 2 presents statistical tests of excitation and inhibition. We found evidence of both. However, inhibition was much larger in magnitude. Further, several further analyses reinforce that inhibition was more relevant to automation use (supplementary materials). First, simulations from the fitted model indicated that inhibition was responsible for the majority of automation’s benefits to accuracy on automation correct trials, and for cost to accuracy and RT on failure trials. Second, exploration of individual differences indicated that inhibition was observed more consistent across participants than excitation, and was more responsible for automation benefits to accuracy. Third, inhibition was correlated across participants with accuracy improvements brought about by automation, whereas excitation was correlated with the accuracy costs of incorrect automation.

A final supplemental analysis included an additional model that allowed excitation and inhibition to vary over the “session” factor, to explore learning effects. This analysis indicated qualitatively similar patterns over experimental sessions, with a smaller (but still substantial) inhibition effect in session two, corresponding to the smaller effects in our conventional results.

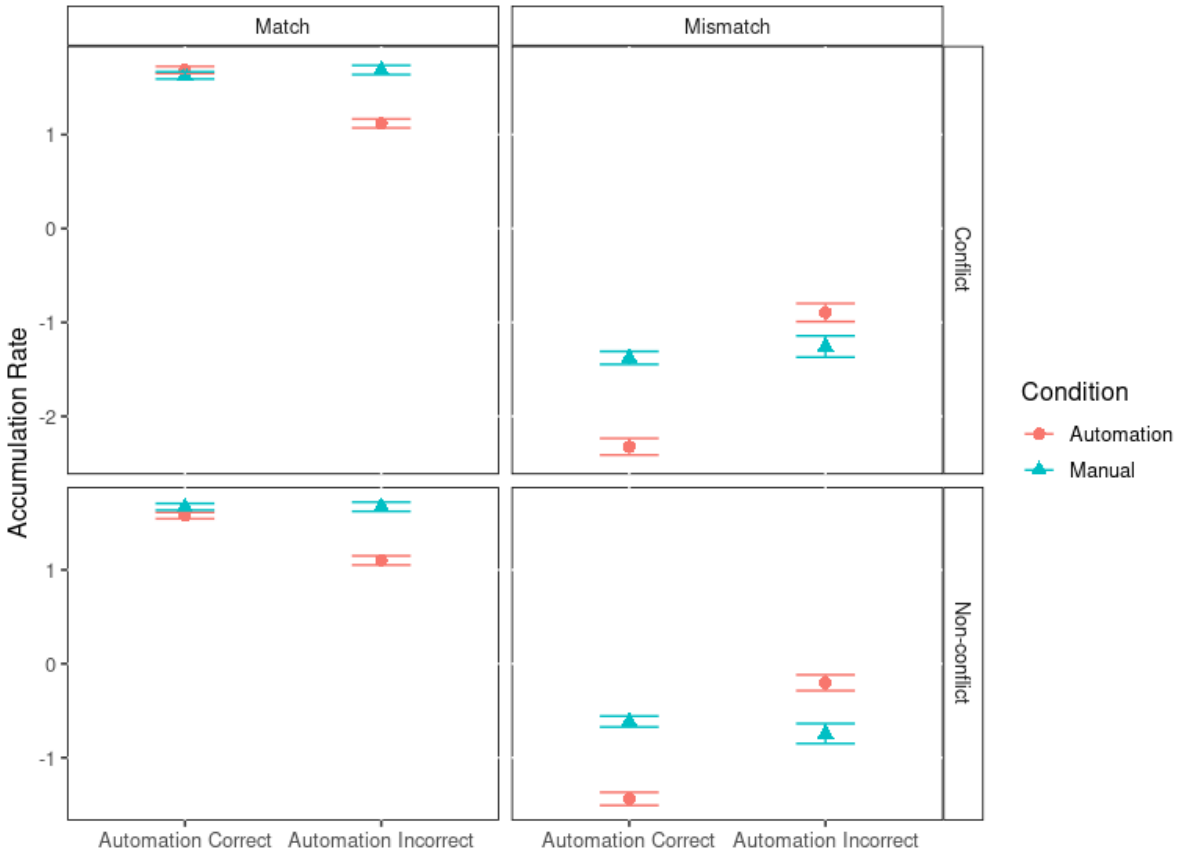

Figure 6. Estimates of accumulation rates. The shapes indicate the posterior means and the error bars correspond to the mean plus or minus the posterior standard deviation.

Table 2

Tests of automation-induced excitation and inhibition effects. We depict Z(p), where Z is the posterior mean of the parameter difference divided by its standard deviation, and p is the one-tailed posterior probability against their being an effect.

Trial Type	Excitation	Inhibition
Conflict Automation Success	2.13 (0.015)	11.41 (<.001)
Conflict Automation Failure	2.65 (0.003)	10.24 (<.001)
Non-conflict Automation Success	2.13 (0.015)	13.21 (<.001)
Non-conflict Automation Failure	2.65 (0.003)	9.8 (<.001)

Threshold effects

Threshold estimates are plotted in Figure 7. Statistical tests of differences in thresholds across automated and manual conditions are included in Table 3. Overall, automation had little effect on thresholds. In session two, both conflict and non-conflict thresholds were slightly higher in manual than automated conditions. Simulations suggest that these effects did not substantially affect performance (supplementary materials). Thus, automation primarily affected participants' evidence accumulation, with little evidence for shifts in speed-accuracy trade-off or bias.

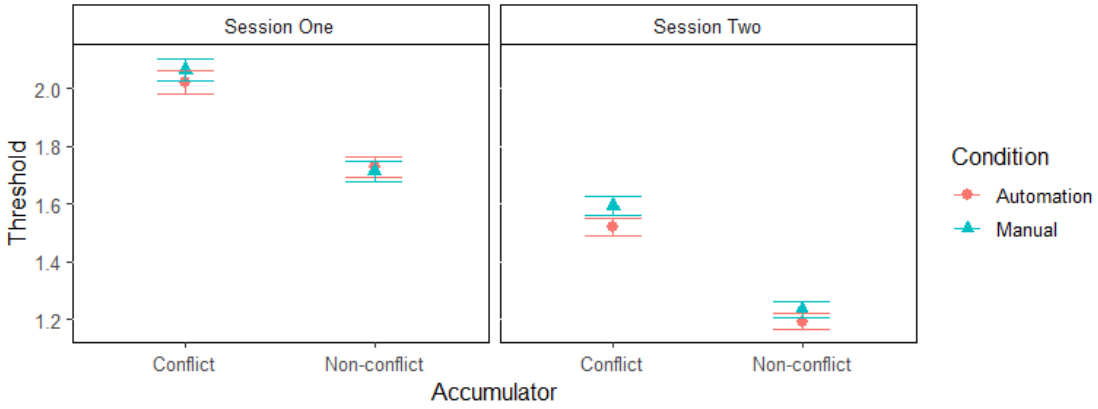

Figure 7. Estimates of thresholds. The shapes indicate the posterior means and the error bars correspond to the mean plus or minus the posterior standard deviation.

Table 3

Tests of differences in thresholds across automated and manual conditions. We depict Z(p), where Z is the posterior mean of the parameter difference divided by its standard deviation, and p is the one-tailed posterior probability against their being an effect.

Accumulator	Session one	Session two
Conflict	1.44 (0.073)	3.18 (<.001)
Non-conflict	-0.54 (0.299)	2.09 (0.019)

Discussion

We used a quantitative model to illuminate the cognitive processes by which humans use automation in an ATC conflict detection task. The automated decision aid increased accuracy when correct but decreased accuracy when incorrect. Correct RTs were longer when the decision aid was incorrect, demonstrating an RT cost from failed automation. Our evidence-accumulation model provided a good fit to the effects of automation use, indicating that advice from the decision aid was primarily integrated into the decision by inhibiting evidence towards the response incongruent with that advice.

Participants may have used inhibition to integrate the decision aid advice because this would increase accuracy without directly increasing the evidence (excitation) in either response accumulator, thereby avoiding the risk that decisions could be made solely based on the decision aid (i.e., complete reliance on automation inputs). Although we also found small excitation effects, it is unlikely that this was strong enough to trigger a response without sampling task information (stimulus inputs). Supplementary analyses, including simulations and exploration of individual differences, provided further support for the idea that inhibition can improve accuracy while minimizing misuse, whereas excitation can lead to risk of misuse.

The asymmetry observed between automation-induced inhibition and excitation may reflect a broader property of human information processing. In conflict tasks such as the Stroop task, participants are slower to respond to stimuli with incongruent dimensions (e.g., identifying that the word “red” is printed in green) than without (e.g., identifying the word “stage” is printed in green), referred to as interference (MacLeod, 1991; also see the picture-word interference task; Starreveld & La Heij, 2017). Participants can also be faster to identify stimuli with two congruent dimensions (e.g., identifying the word “green” is printed in green) than without, referred to as facilitation, however, this effect is much smaller (MacLeod,

1991). This asymmetry mirrors our results, in which decision aids cause strong inhibition, but only weak excitation. However, in our task, participants were aware that the decision aid was informative, and were encouraged to integrate it with stimulus inputs, whereas in Stroop they are requested to base decisions solely on one information source. Stroop conflict is attributed to interference from an automatically retrieved competing response, whereas in our model, inhibition arises due to mismatching inputs. Nonetheless, both our task and conflict tasks require participants to execute a response potentially cued by two conflicting sources of information, and thus it is reasonable to expect similarities in the underlying processes.

The success of our model in accounting for the effects of automation is a critical first step to moving beyond identifying disparate factors affecting automation use, and towards the identification and quantification of the cognitive mechanisms underlying automation use. Our tractable quantitative modelling framework has the potential to be able to be used to generalize and unify findings across the automation literature.

Practical Implications

To the extent that humans integrate decision aids into their decisions via inhibition, rather than excitation, incorrect decision aids are likely to slow RTs. Thus, in situations with high time pressure, providing imperfect decision aids may produce undesirable RT costs. However, in situations without time pressure, inhibition may boost accuracy, making decision aids desirable. Had we found that decision aids caused excitation, rather than inhibition, this would suggest decision aids can speed up performance and alleviate time pressure, but this was not the case. Given the apparent importance of inhibition for interacting with automation, inhibitory abilities may be a desirable quality to either train or select for in work contexts that require humans to interact with less than perfect decision aids. However, more work is needed to identify the boundary conditions under which inhibition is the mechanism underlying automation use. For example, our task did not require visual search, whereas

many dynamic display tasks do. In situations where automation can reliably direct visual attention, it seems likely it would improve rather than cost RT.

In many human factors studies, including in field settings, there may not be enough observations to directly apply our model. Fortunately, our findings highlighted a characteristic pattern in the manifest data that was associated with inhibition. Accurate decision aids benefitted accuracy, inaccurate decision aids reduced accuracy, and inaccurate automation slowed mean RT. Identifying this pattern provides a means to identify inhibition mechanisms in situations that are more difficult to cognitively model, such as high-fidelity task simulations and field studies.

A longer-term practical implication is the advancement of a human performance model of human-automation interaction, which could be used to predict performance and inform work design. With its dynamic decision 'front end' that predicts choice and RT, our model could provide a crucial link from performance data to cognitive architectures like *ACT-R* (Anderson & Lebiere, 2014), which in turn could be inputted to broader task network architectures like *IMPRINT* (Samms, 2010), to account for system-level work performance (Lebiere et al., 2005).

Future Directions

One key direction is to examine how the cognitive mechanisms underlying automation use change under different conditions. For example, the current automation was approximately equal to human ability. When using automation known to be more accurate than they are, humans may implement excitation-based strategies, possibly even to the extent that they routinely base their decisions solely on automation input. Similarly, the cognitive mechanisms underlying automation use may vary depending on the relative cost of errors when automation fails and succeeds. We instructed participants in a way that made the possibility of automation failure salient (incorrectly agreeing with automation was more

costly than incorrectly disagreeing), to encourage them not to rely entirely automation. If instead the cost of accepting faulty automated advice is low, the mechanisms underlying automation use may differ. Another direction is to extend our model to account for the effects of time on task and practice. Our supplemental analysis suggested that in participants' second experimental session, inhibition and subsequent behavioural effects, although remaining substantial, were reduced. This suggests it would be fruitful to pursue future studies that examine models of automation incorporating learning and adaptive processes.

References

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

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

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

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

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

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Farrell, S., & Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. https://doi.org/10.1177/0963721410386677

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

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience*, *31*(17), 6339–6352. http://doi.org/10.1523/JNEUROSCI.5613-10.2011

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

Lebiere, C., Archer, R., Warwick, W., & Schunk, D. (2005). *Abstract Integrating Modeling and Simulation into a General-Purpose Tool*. Proceedings of the 11th International Conference on Human Computer Interaction, Las Vegas, NV.

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

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. S. (2007). The development of a general associative learning account of skill acquisition in a relative arrival-time judgment task. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 938–959. https://doi.org/10.1037/0096-1523.33.4.938

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Human Factors*, *49*(3), 376–399. https://doi.org/10.1518/001872007X197017

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods in Psychology*, *4*(2), 61– 64. http://dx.doi.org/10.20982/tqmp.04.2.p061

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

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

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, *111*(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

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

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

Starreveld, P. A., & La Heij, W. (2017). Picture-word interference is a Stroop effect: A theoretical analysis and new empirical findings. *Psychonomic Bulletin & Review, 24*(3), 721-733. https://doi.org/10.3758/s13423-016-1167-6

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

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

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*(10), 995–1011. https://doi.org/10.1080/00140138008924809