




The Benefits and Costs of Low and High Degree of Automation

Monica Tatasciore^{}, Vanessa K. Bowden, Troy A. W. Visser^{},
Steph I. C. Michailovs^{}, and Shayne Loft, The University of Western
Australia, Perth, Australia

Objective: The objective of this study is to examine the effects of low and high degree of automation (DOA) on performance, subjective workload, situation awareness (SA), and return-to-manual control in simulated submarine track management.

Background: Theory and meta-analytic evidence suggest that as DOA increases, operator performance improves and workload decreases, but SA and return-to-manual control declines. Research also suggests that operators have particular difficulty regaining manual control if automation provides incorrect advice.

Method: Undergraduate student participants completed a submarine track management task that required them to track the position and behavior of contacts. Low DOA supported information acquisition and analysis, whereas high DOA recommended decisions. At a late stage in the task, automation was either unexpectedly removed or provided incorrect advice.

Results: Relative to no automation, low DOA moderately benefited performance but impaired SA and non-automated task performance. Relative to no automation and low DOA, high DOA benefited performance and lowered workload. High DOA did impair non-automated task performance compared with no automation, but this was equivalent to low DOA. Participants were able to return-to-manual control when they knew low or high DOA was disengaged, or when high DOA provided incorrect advice.

Conclusion: High DOA improved performance and lowered workload, at no additional cost to SA or return-to-manual performance when compared with low DOA.

Application: Designers should consider the likely level of uncertainty in the environment and the consequences of return-to-manual deficits before implementing low or high DOA.

Keywords: automation, submarine track management, situation awareness, workload, complacency

Address correspondence to Monica Tatasciore, The University of Western Australia, 35 Stirling Highway, Perth, Western Australia 6009, Australia; e-mail: monica.tatasciore@research.uwa.edu.au.

HUMAN FACTORS

Vol. 62, No. 6, September 2020, pp. 874–896

DOI: 10.1177/0018720819867181

Article reuse guidelines: sagepub.com/journals-permissions
Copyright © 2019, Human Factors and Ergonomics Society.

Technological developments in computer hardware and software have made it possible to automate many aspects of complex work systems, significantly improving workplace efficiency and safety (Sheridan, 2015; Vagia, Transeth, & Fjerdingen, 2016). Automation can be defined as “a device or system that accomplishes a function that was previously, or conceivably could be, carried out by a human operator” (Parasuraman, Sheridan, & Wickens, 2000, p. 287). Examples of automation include image-guided navigation tools in surgery, flight management systems in cockpits, aircraft separation assurance technology in air traffic control, and decision aids in unmanned vehicle control.

Whereas routine lower level tasks have typically been the first to be automated due to their operational predictability, with ongoing emphasis on maximizing system capacity, and the development of sophisticated machine-learning algorithms, automation can now begin to recommend or even execute high-level decisions for operators. The submarine control room is one area in which this type of “decision-level” automation is rapidly developing (Roberts, Stanton, & Fay, 2017). Submarine track management, for example, requires operators to coordinate information across multiple displays to create a tactical picture of the position and behavior of contacts in relation to the submarine (Ownship) and strategic landmarks (Kirschenbaum, 2011). A key question in this, and in similar work contexts (e.g., unmanned vehicle control, air traffic control), concerns the extent to which an operator can effectively use automated systems that recommend decisions.

Researchers have long recognized the potential costs associated with automation. These costs include reductions in operators’ understanding of

a task and their ability to anticipate future task events (situation awareness [SA]; Endsley, 1988) due to automation-induced complacency (Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997), and reductions in operators' ability to regain manual control after automation use (Kaber & Endsley, 2004; Parasuraman & Manzey, 2010). Notably, there is also some evidence that these costs increase as automation begins to assume higher level functions. For example, Onnasch, Wickens, Li, and Manzey (2014) reviewed 18 studies that varied in "degree of automation" (DOA)—an ordinal metric that ranked the level of work the automation was doing (Sheridan & Verplank, 1978) across four stages: information acquisition, information analysis, decision recommendation, and action execution (Parasuraman et al., 2000). Onnasch et al. (2014) found that as DOA increased, performance improved and workload decreased. However, SA and return-to-manual performance declined (for examples of specific studies that have shown this trade-off, see Kaber, Onal, & Endsley, 2000; Li, Wickens, Sarter, & Sebok, 2014; Manzey, Reichenbach, & Onnasch, 2012).

More recently, Chen, Visser, Huf, and Loft (2017) asked participants to monitor a submarine track management tactical display ("Surface Plot") that presented the location and heading of contacts in relation to the Ownship and landmarks, and a "waterfall" display that presented sonar bearings of contacts and how those bearings changed with time. Participants performed three tasks. The classification task required participants to classify contacts (hostile, friendly, etc.) based on how long they had spent within certain display regions. The closest point of approach (CPA) task required participants to monitor changes in contact heading to determine their CPA to Ownship. The dive task required participants to integrate contact location and heading information to determine when the submarine could safely dive. The simulation automated information acquisition and analysis stages of the classification and CPA tasks (i.e., relatively low DOA) by indicating to participants when contacts first entered display regions (to aid contact classification) and tracked when contacts made heading changes (to aid CPA detection).

Chen et al. (2017; Experiment 3; between-subjects design) demonstrated that low DOA resulted in benefits to classification performance (accuracy and response time [RT]) but not CPA performance, and did not reduce subjective workload, compared with when no automation was provided. In addition, participant SA was poorer when automation was used, as was performance on the non-automated dive task. The cost observed to the non-automated dive task with the use of automation is critical to further explore because this novel finding suggests that operators in complex work systems may find it difficult to maintain adequate performance on non-automated tasks that share information processing requirements with currently automated tasks. After low DOA was unexpectedly removed, costs to SA did not diminish, although there were no associated return-to-manual performance deficits.

With defense and other industries focused on developing high DOA that recommends decisions to operators (Endsley, 2017; U.S. Air Force, 2015), it is critical to further understand how high DOA systems can affect operators. Under conditions when automation is reliable, high DOA that recommends decisions to operators should further improve performance and reduce workload compared with low DOA (Onnasch et al., 2014). The key question concerns whether high DOA comes at increased cost to concurrent non-automated task performance and SA, or return-to-manual performance, compared with low DOA. The answer to this is critical for work design. If high DOA produces greater benefit at no extra cost, then it would be more desirable to employ than low DOA. If, however, high DOA produces greater benefit but at extra cost, whether high DOA is deployed would depend on factors such as the level of uncertainty in the environment or the operational consequences of reduced concurrent non-automated task performance, loss of SA, or return-to-manual performance deficits (Endsley, 2017; Wickens, Clegg, Vieane, & Sebok, 2015).

With these questions in mind, the current study began by examining the effects of low DOA and high DOA on operator performance, workload, SA, non-automated task performance,

and return-to-manual performance in submarine track management. Low DOA was identical to that used by Chen et al. (2017) and supported information acquisition and analysis by displaying when and for how long contacts were positioned in an area of interest (classification task) and by displaying contact heading changes (CPA task). High DOA not only provided the same information acquisition and analysis information but also made explicit recommendations to participants regarding when and what to classify contacts, and when a contact had made a CPA. The purpose of Experiment 1 was (a) to replicate the effects of low DOA on performance and SA that were demonstrated by Chen et al. (2017) when compared with no automation and (b) to examine whether high DOA produces benefits to performance and workload compared with no automation and low DOA, and whether high DOA increases costs to non-automated task performance, SA, or return-to-manual performance compared with no automation and low DOA. Our predictions regarding the effects of DOA are summarized in Table 1 and described in detail later.

AUTOMATED TASK PERFORMANCE

The classification and CPA tasks were automated for the low and high DOA conditions. Performance on these two tasks was assessed by accuracy and RT. Under routine states (when automation was reliably functioning), we expected higher classification accuracy and faster classification RT with increasing DOA. While Chen et al. (2017) did not show a benefit to CPA accuracy with the use of low DOA under routine states, we expected to find benefits to CPA accuracy with high DOA. Furthermore, Chen et al. found that participants made slower CPA decisions when using low DOA, which reflects that the automated track history allowed participants to detect CPAs well after they had occurred by detecting past heading changes. In contrast, high DOA should allow participants to make faster CPA decisions compared with both no automation and low DOA because it highlights CPA events at the actual time they occur.

Chen et al. (2017) found no return-to-manual deficits to the classification or CPA tasks when low DOA was removed. However, the

Onnasch et al. (2014) meta-analysis indicated that the negative consequences of automation are more likely with higher DOA. From a theoretical perspective, higher DOA that recommends decisions could reduce the perceived need to actively process raw information (e.g., contact position and heading) on the displays (complacency; Parasuraman & Manzey, 2010; Wickens, Sebok, Li, Sarter, & Gacy, 2015). Theory and evidence from the broader psychological science literature also predicts poorer understanding and retention of information when individuals passively process information rather than actively making decisions (e.g., the generation effect, Slamecka & Graf, 1978; the testing effect, Roediger & Karpicke, 2006; transfer of training, Blume, Ford, Baldwin, & Huang, 2010). We therefore expected to find deficits to performance on the classification and CPA tasks when high DOA was removed, as compared with the no automation and low DOA conditions.

NON-AUTOMATED TASK PERFORMANCE

The Chen et al. (2017) cost observed to the non-automated dive task (to both accuracy and RT) under routine states with the use of low DOA suggests that participants found it difficult to maintain performance on the non-automated task that shared information processing requirements with the automated tasks. That is, dive task performance was degraded because participants scrutinized contact location and heading information less closely when using automation (complacency). To the extent that complacency effects are heightened with high DOA as suggested by Onnasch et al. (2014), we would expect dive task performance to be further impaired with the use of high DOA. Furthermore, on the basis of Chen et al.'s (2017) and Onnasch et al.'s (2014) meta-analytic evidence, we expected return-to-manual deficits when high DOA was removed, as compared with the no automation and low DOA conditions.

WORKLOAD

Chen et al. (2017) did not find reduced subjective workload with low DOA during routine

TABLE 1: Predictions for Experiment 1

Task		Routine	Removal
Classification	Accuracy	None < Low < High (the higher the DOA, the better the accuracy)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	None > Low > High (the higher the DOA, the faster the decisions)	[None = Low] < High (slower decisions after high DOA removal)
CPA	Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	High < None < Low (benefits to RT with high DOA, slower decisions with low DOA)	[None = Low] < High (slower decisions after high DOA removal)
Dive	Accuracy	None > Low > High (the higher the DOA, the poorer the accuracy)	[None = Low] > High (lower accuracy after high DOA removal)
	RT	None < Low < High (the higher the DOA, the slower the decisions)	[None = Low] < High (slower decisions after high DOA removal)
Workload		[None = Low] > High (reduced workload with high DOA)	[None = Low] < High (higher workload after high DOA removal)
SA		None > Low > High (the higher the DOA, the poorer the SA)	None > Low > High (after removal, the higher the DOA, the poorer the SA)

Note. Routine = automation is reliable; Removal = after automation is removed; DOA = degree of automation; RT = response time; None = no automation; Low = low DOA; High = high DOA; CPA = closest point of approach; SA = situation awareness.

states, but on the basis of the Onnasch et al. (2014) meta-analysis, we expected reduced subjective workload with high DOA, as compared with the no automation and low DOA conditions. Chen et al. (2017) found no return-to-manual increase in workload when low DOA was removed, but on the basis of Onnasch et al. (2014), we expected to find increased subjective workload when high DOA was removed, as compared with the no automation and low DOA conditions.

SA

Chen et al. (2017) found reduced SA with low DOA during routine states, and based on Onnasch et al. (2014), we expected SA to be further impaired with high DOA. Chen et al. (2017) found reduced SA when low DOA was

removed, and on the basis of Onnasch et al. (2014), we expected to find that SA would be further impaired when high DOA was removed.

EXPERIMENT 1

Participants

Participants were 122 (86 females) undergraduate psychology students (age: *M* = 23 years, *SD* = 7.2) who took part for course credit and were randomly assigned to one of three conditions: no automation (*n* = 42), low DOA (*n* = 40), and high DOA (*n* = 40). This research complied with the American Psychological Association Code of Ethics and was approved by the Human Research Ethics Office at the University of Western Australia. Informed consent was obtained from each participant.

Design

A mixed design was used, where the between-subjects factor was condition (no automation, low DOA, high DOA) and the within-subjects factor was automation state (routine, automation removal). Automation condition was manipulated between-subjects so that each participant only experienced the unexpected automation failure once to ensure that there were no carryover effects (first-failure effect; see Merlo, Wickens, & Yeh, 2000). Participants completed three 27.5-min track management scenarios, each corresponding to different Australian coastal maps.

Simulated Submarine Track Management Task

The track management simulation (Figure 1) was developed based on a task analysis conducted with Royal Australian Navy Submariners (Chen et al., 2017). The tactical display, presented on the left monitor, showed a “bird’s eye” view of the area with concentric rings representing distance from the center point (Ownship). This tactical display presented the location and heading of contacts. The waterfall display, presented on the right monitor, showed contact bearings in relation to Ownship on the top horizontal axis, and how these bearings changed with time along the vertical axes. This information was displayed as vertical lines or “soundtracks,” which grew downward with time. Task load periodically varied with the number of contacts increasing (maximum of eight contacts) and decreasing (minimum of one contact) 3 times during each 27.5-min scenario.

The “Track Assist” automation interface was located at the bottom right of the tactical display and allowed participants to determine whether the automation was always on (fixed) or not available (none). During the third scenario, automation was unexpectedly removed (10.58-min into the 27.5-min scenario). When the automation was removed, a message appeared on the tactical display: “Attention. ENEMY SONAR detected. Track Assist turned off. Manual tracking required.” Participants were required to acknowledge this message by clicking an “ok” button. In the no automation condition, a message was presented at the same time that read:

“Attention. ENEMY SONAR detected. Keep vigilant and continue to track vessels.”

Classification task. Participants classified contacts depending on how long they spent within specific areas on the tactical display. A contact was a “Friendly” if it spent more than 2 continuous minutes within the area bounded by blue lines on the tactical display. A contact was a “Merchant” if it spent more than 2 continuous minutes within the “shipping lane” represented by two white parallel lines on the tactical display. A contact was a “Trawler” if it spent more than 2 continuous minutes in the shallow dark blue areas on the tactical display. A contact was an “Enemy” if in the first 4-min of its presentation, it had not spent at least 1 continuous minute in any classification zone. To track whether a contact had been in a given area for more than 2 min, participants could place horizontal lines on the top of each soundtrack on the waterfall display when a contact entered an area of interest. When this line reached the 2-min mark, a contact could be classified. To detect enemies, participants could place the horizontal line on the bottom of the soundtrack of any contact that had not crossed into an area of interest. Once this horizontal line reached 4-min, the contact could be classified as an enemy.

The contact classification task could be automated to either a low or high degree. For low DOA, horizontal lines were automatically placed on the soundtrack when a contact entered an area of interest. In addition, a horizontal line was automatically placed at the bottom of the soundtrack when it reached the 4-min mark to assist with classifying enemies. Participants still had to monitor the horizontal lines to see when they reached the 2-min mark (or 4-min mark for enemies) to classify contacts. When automation was removed in the third scenario, the existing horizontal lines on the waterfall display remained, but subsequent lines had to be manually entered.

High DOA was identical to low DOA, except that a square box with the recommended classification (i.e., f = Friendly, m = Merchant, t = Trawler, e = Enemy; see Figure 1) was attached to the horizontal lines on the soundtracks. In addition, when the horizontal line reached the 2-min mark, it flashed to notify participants that the contact had been in an area of interest for 2

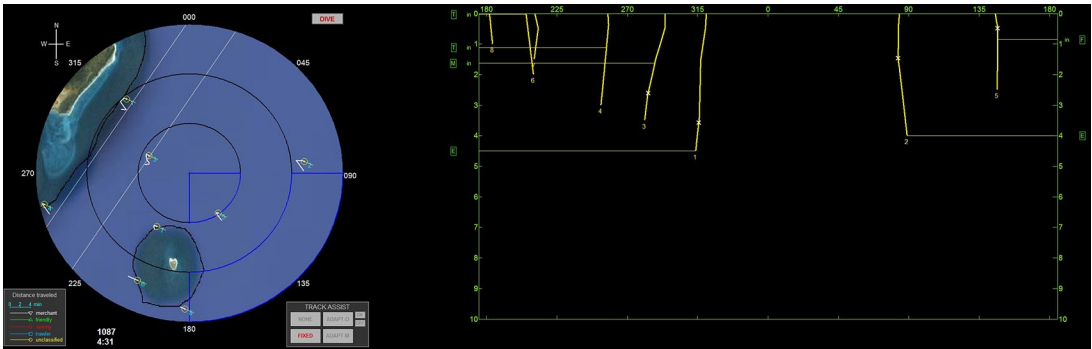


Figure 1. An example of submarine track management scenario. The display on the left is the tactical display which presents a bird's eye view of the area with concentric rings representing distance from the Ownship. The display on the right is the waterfall display, which provides the bearing of contacts in relation to the Ownship on the top horizontal axis, and how those bearings change with time along the vertical axes. These data are presented as soundtracks that grow downward with time. Eight contacts are displayed. Projecting from the center of each contact is a line which indicates the current heading of the contact. In this example, high DOA is active. On the tactical display, a track history is attached to each contact to reflect contact heading changes to assist with the CPA task. The track history will flash to notify participants when a CPA has occurred. Present on the waterfall display are horizontal lines which are automatically placed when a contact enters an area of interest. Attached to these lines are boxes which include the appropriate classification letter of a given contact. These lines will flash when a contact can be classified. The low DOA displays are the same to that shown in this figure, except there will be no boxes with the appropriate classification letter attached to the horizontal lines. In addition, the track history and the horizontal lines will not flash to signal a classification or CPA event with the use of low DOA. When no automation is provided, contacts will not have a track history for the CPA task, and horizontal lines must be placed manually for the classification task.

min. If the contact was an enemy, the horizontal line flashed at the 4-min mark. It remained the participants' task to then execute the classification task action (or not) after receiving the automated advice. When automation was removed, the existing horizontal lines and any classification letters remained, but did not flash in the future, and subsequent lines had to be manually entered.

CPA. The CPA is defined as the time at which a contact that was heading toward the Ownship turned away from the Ownship. Participants reported the time at which the CPA occurred by placing a cross on the corresponding soundtrack on the waterfall display. Each contact had one CPA per scenario. For the two automation conditions, the CPA task was automated to a low or high degree. For low DOA, each contact was presented with a track history, which reduced the need for participants to track which contacts made heading changes.

Participants were still required to interpret the track history to mark the timing of each CPA on the waterfall display. For high DOA, the track history also flashed to alert the participant when the contact had turned away from the Ownship. It was then the participant's responsibility to mark the appropriate CPA time on the waterfall display. When automation was removed, the existing track history for high DOA and low DOA remained on the tactical display but was not updated to reflect further contact movement.

Dive task. Participants were required to dive when (a) all contacts on the tactical display were heading in the same direction and (b) one contact was heading directly toward the Ownship. There were either 9 or 10 dive windows per scenario, and each dive window varied in duration between 10 and 30 s. Participants responded to dive windows by pressing the dive button. The dive task was not automated.

TABLE 2: The SAGAT Queries Used to Measure Participant SA

SA Level	SAGAT Queries		
1	Which vessel is currently in an X zone?	How many vessels are heading away from you?	How many vessels are currently facing the same direction?
	Is vessel X currently in an X zone?	Is vessel X heading away from you?	Are any vessels heading directly toward you? How many vessels are heading away from you?
2	Has any vessel been in an X zone for more than 1 min?	How many times has vessel X changed course?	Are any vessels heading in the same direction?
	How many vessels are currently in an X zone? Which vessel most recently crossed a classification boundary?	Has vessel X had any kinks in its soundtrack?	Which vessel is currently heading toward you?
3	Which unclassified vessel is most likely to be an X?	Which vessel would make a CPA if it turned to a heading of xxx?	Would vessel X head directly toward you if it turned to a heading of xxx?
	Could vessel X cross a boundary within 4-min time?	Would a CPA be made for vessel X if it turned to a heading of xxx?	

Note. SAGAT = Situation Awareness Global Assessment Technique; SA = situation awareness; CPA = closest point of approach.

Measures

SA. SA was measured using the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995). During each scenario, the simulation was paused 6 times, and the tactical and waterfall displays were blanked and replaced with SAGAT queries. Within each freeze, seven SAGAT queries were delivered. The first question always asked participants to mark a specific contact location on the tactical display, whereas the remaining six questions targeted knowledge necessary for the classification, CPA, and dive tasks, at the three levels of SA (Endsley, 1995). During each SAGAT freeze, all participants received the same six SA queries which were taken from the pool of queries presented in Table 2.

Workload. Two subjective measures of workload were used. The Air Traffic Workload Input Technique (ATWIT; Stein, 1985) was presented on the tactical display every minute

throughout the scenario. Participants had 10 s to select a workload rating between 1 and 10 (1–2 = very low, 3–5 = moderate, 6–8 = relatively high, and 9–10 = very high). The National Aeronautics and Space Administration Task Load Index (NASA-TLX; Hart & Staveland, 1987) was completed after each scenario. A NASA-TLX score was calculated by multiplying the ratings for each subscale of workload by its corresponding weighting, adding the values for all the subscales together, and then dividing the total by 15.

Procedure

The experiment duration was 3 hr. First, participants completed an 80-min training session. Training began with a 35-min audiovisual PowerPoint presentation that explained the task and included “learning checks.” Following this, participants viewed a narrated video of the simulation in which all tasks were demonstrated

without automation. Participants then completed a 27.5-min practice scenario with no automation. Participants who were in either the low or high DOA conditions watched a PowerPoint presentation that explained their automation. Participants then completed three 27.5-min experimental scenarios in their assigned automation condition. Each scenario contained unique contacts and the order of scenario maps was counterbalanced.

Results

The hit rates for the classification, CPA, and dive tasks were calculated as the number of correct task responses per scenario divided by the total number of task events. RTs were based on correct decisions only. A CPA was marked as correct if the cross was placed at any time 1.5 s before or after the actual CPA, so long as the cross was placed on the correct soundtrack. If placed outside this temporal range, then the cross was recorded as a false alarm. A parameter was estimated for false alarm rates as the exact number of contacts and events associated with making a false alarm was indeterminable (Chen et al., 2017). CPA false alarms were most likely to be made in response to contact course changes. The false alarm rate was therefore calculated to be the number of false alarms divided by the number of course changes, minus 24 (the total number of CPA events per scenario). CPA performance was then calculated by subtracting the CPA false alarm rate from the hit rate. Course changes were always required for a dive window to be open; hence, a dive false alarm was most likely to be made during a course change. As there are fewer dive windows than CPAs, as well as the rule that all contacts need to be heading in the same direction for a dive window, it was less likely that every course change could be mistaken for a dive window. Therefore, the dive false alarm rate was calculated by dividing the number of false alarms by half the number of course changes, minus the total number of dive windows (Chen et al., 2017). Dive task performance was then calculated by subtracting the dive false alarm rate from the hit rate.

The means and between-subjects 95% confidence intervals (CIs) for performance, subjective workload, and SA are presented in Table 3.

Data are separated into the time period automation was available (routine state: first two scenarios and first third of the last scenario) and the time period automation was removed (automation removal state: last two thirds of the last scenario). To test our predictions, we ran 3 Condition (no automation, low DOA, high DOA) \times 2 Automation State (routine, automation removal) mixed analyses of variance (ANOVAs) on the performance, workload, and SA variables. The between-subjects factor was condition and the within-subjects factor was automation state. The ANOVA results are summarized in Table 4. Significant main effects of condition, or interactions between condition and automation state, were followed up with tests of simple effects (reported in text). To do this, we ran one-way ANOVAs separately for the routine and automation removal states. We then followed significant one-way ANOVAs with post hoc *t* tests that compared the three conditions (no automation, low DOA, high DOA) to each other, correcting for family-wise error by reporting Bonferroni-corrected *p* values (the actual *p* value was multiplied by the number of comparisons for each dependent variable, which was three). Estimates of Cohen's *d* suggested we had a power of 0.82 to detect the medium-to-large effect sizes previously reported by Chen et al. (2017; Cohen, 1988).

Note that several main effects of automation state were found (see Table 4). Classification and CPA performance was poorer, and workload higher, after automation removal compared with routine states. Similar effects were reported by Chen et al. (2017) and are likely caused by the fact that for the low and high DOA conditions (which constitute two of the three conditions and thus 2/3 of the data), the removal state represented the time that automation was removed. For brevity, the main effects of automation state are not further discussed, and we focus on following up the main effect of condition and the interactions.

Automated Task Performance

Classification task. For classification accuracy, there was a main effect of condition and a Condition \times State interaction. The simple effect test revealed a significant difference between the

TABLE 3: Descriptive Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 1

Automation	Classification			CPA			Dive			SAGAT			Workload Rating		
	Hit	RT		Hit-FA	RT		Hit-FA	RT		Accuracy		ATWIT		NASA-TLX	
Routine state															
None	0.70 [0.62, 0.78]	29.40 [26.01, 32.91]		0.25 [0.18, 0.32]	18.70 [13.76, 23.64]		0.70 [0.64, 0.76]	9.08 [8.06, 10.09]		0.58 [0.54, 0.61]		4.83 [4.47, 5.18]		59.30 [54.87, 63.73]	
Low	0.77 [0.70, 0.85]	23.60 [20.64, 26.56]		0.30 [0.21, 0.40]	32.62 [25.53, 39.72]		0.55 [0.48, 0.63]	10.19 [8.98, 11.40]		0.50 [0.45, 0.54]		4.83 [4.44, 5.23]		61.35 [56.70, 66.00]	
High	0.90 [0.85, 0.95]	19.32 [16.30, 22.35]		0.74 [0.64, 0.84]	12.31 [10.07, 15.01]		0.57 [0.50, 0.65]	10.90 [9.84, 11.96]		0.52 [0.48, 0.56]		4.10 [3.70, 4.50]		54.02 [48.72, 59.33]	
Automation removal state															
None	0.68 [0.59, 0.78]	26.10 [20.09, 32.11]		0.30 [0.22, 0.38]	18.69 [12.65, 24.74]		0.73 [0.66, 0.80]	8.71 [6.99, 10.43]		0.51 [0.47, 0.56]		4.86 [4.41, 5.32]		57.68 [51.63, 63.74]	
Low	0.65 [0.56, 0.75]	31.12 [24.70, 37.54]		0.25 [0.17, 0.33]	30.89 [12.57, 49.21]		0.72 [0.64, 0.81]	9.77 [8.51, 11.03]		0.53 [0.49, 0.56]		5.92 [5.44, 6.41]		66.99 [61.44, 72.54]	
High	0.79 [0.72, 0.86]	23.36 [19.58, 27.14]		0.44 [0.36, 0.52]	18.37 [12.70, 24.05]		0.71 [0.62, 0.81]	9.86 [8.11, 11.61]		0.54 [0.49, 0.59]		5.24 [4.71, 5.77]		60.63 [55.13, 66.14]	

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach; SAGAT = Situation Awareness Global Assessment Technique; RT = response time; FA = false alarm; ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index.

TABLE 4: Inferential Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 1

Dependent Variable	Effect	<i>F</i>	<i>df</i>	<i>p</i>	η^2_p
Classification (Hit)	Condition	5.38	(1, 119)	.01*	.08
	State	26.69	(1, 119)	<.001*	.18
	Interaction	4.49	(1, 119)	.01*	.07
Classification (RT)	Condition	3.89	(1, 115)	.02*	.06
	State	3.46	(1, 115)	.07	.03
	Interaction	4.50	(1, 115)	.01*	.07
CPA (Hit-FA)	Condition	21.95	(1, 80)	<.001*	.27
	State	29.84	(1, 119)	<.001*	.20
	Interaction	31.25	(1, 119)	<.001*	.34
CPA (RT)	Condition	5.83	(1, 113)	.01*	.09
	State	0.56	(1, 113)	.46	.01
	Interaction	0.50	(1, 113)	.61	.01
Dive (Hit-FA)	Condition	1.53	(1, 119)	.22	.03
	State	40.63	(1, 119)	<.001*	.26
	Interaction	6.02	(1, 119)	.003*	.09
Dive (RT)	Condition	2.21	(1, 117)	.14	.04
	State	1.85	(1, 117)	.18	.02
	Interaction	0.19	(1, 117)	.83	.003
NASA-TLX	Condition	2.19	(1, 118)	.12	.04
	State	13.01	(1, 118)	<.001*	.10
	Interaction	7.03	(1, 118)	.001*	.11
ATWIT	Condition	3.31	(1, 117)	.04*	.05
	State	62.43	(1, 117)	<.001*	.35
	Interaction	14.39	(1, 117)	<.001*	.20
SAGAT (Accuracy)	Condition	1.04	(1, 117)	.36	.02
	State	0.14	(1, 117)	.71	.001
	Interaction	7.29	(1, 117)	.001*	.11

Note. RT = response time; CPA = closest point of approach; FA = false alarm; NASA-TLX = National Aeronautics and Space Administration Task Load Index; ATWIT = Air Traffic Workload Input Technique; SAGAT = Situation Awareness Global Assessment Technique.
**p* < .05.

conditions during routine states, $F(2, 119) = 9.04, p < .001, \eta^2 = .13$. During routine states, there was no difference in classification accuracy between the no automation and low DOA conditions, $t < 1$. However, participants provided high DOA made more accurate classifications than participants provided no automation, $t(80) = 4.43, p < .001, d = 0.99$, or low DOA, $t(78) = 2.84, p = .02, d = 0.64$. For automation removal states, the simple effect test indicated no

significant difference between the conditions, $F(2, 119) = 2.85, p = .06, \eta^2 = .05$. In summary, only high DOA benefited classification accuracy during routine states, and there were no return-to-manual deficits to classification accuracy following the use of low or high DOA.

For classification RT, there was a main effect of condition and a Condition \times State interaction. The simple effect test revealed a significant difference between the conditions during routine

states, $F(2, 118) = 10.66, p < .001, \eta^2 = .15$. During routine states, participants provided low DOA, $t(79) = 2.60, p = .03, d = 0.58$, or high DOA, $t(79) = 4.46, p < .001, d = 0.99$, made faster classifications than participants provided no automation. There was no difference in classification RT between participants provided high DOA and those provided low DOA, $t(78) = 2.04, p = .13$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 115) = 2.10, p = .13, \eta^2 = .04$. In summary, both low and high DOA benefited classification RT during routine states, and there were no return-to-manual deficits to classification RT following the use of low or high DOA.

CPA task. For CPA accuracy, there was a main effect of condition and a Condition \times State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 36.17, p < .001, \eta^2 = .38$. During routine states, there was no difference in CPA accuracy between the no automation and low DOA conditions, $t < 1$. Participants provided high DOA made more accurate CPA decisions than participants provided no automation, $t(80) = 7.95, p < .001, d = 1.76$, or low DOA, $t(78) = 6.38, p < .001, d = 1.47$. For automation removal state, the simple effect test revealed a significant difference between the conditions, $F(2, 119) = 6.63, p = .002, \eta^2 = .10$. There was no difference in return-to-manual CPA accuracy between the no automation and low DOA conditions, $t < 1$. Participants previously using high DOA made *more* accurate CPA decisions than participants using no automation, $t(80) = 2.57, p = .04, d = 0.57$, or low DOA, $t(78) = 3.58, p = .003, d = 0.78$. In summary, high DOA benefited CPA accuracy during both routine states and after automation was removed.

For CPA RT, there was a main effect of condition but no Condition \times State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 16.10, p < .001, \eta^2 = .22$. During routine states, participants provided low DOA made slower CPA decisions than those provided no automation, $t(78) = 3.29, p = .01, d = 0.73$, or high DOA, $t(77) = 5.46, p < .001, d = 1.22$. The difference in CPA RT for the high

DOA condition compared with the no automation condition did not reach significance, $t(79) = 2.24, p = .08$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 115) = 1.58, p = .21, \eta^2 = .03$. In summary, low DOA impaired CPA RT during routine states, and there were no return-to-manual deficits for CPA RT following the use of low or high DOA.

Non-Automated Task Performance

For dive task accuracy, there was a Condition \times State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 119) = 5.27, p = .01, \eta^2 = .08$. During routine states, participants provided low DOA, $t(80) = 3.02, p = .01, d = 0.68$, and participants provided high DOA, $t(80) = 2.71, p = .02, d = 0.60$, made poorer dive decisions than participants provided no automation. Dive accuracy was not further degraded by the use of high compared with low DOA, $t < 1$. For automation removal state, the simple effect test revealed no significant difference between the conditions, $F(2, 119) = 0.03, p = .97, \eta^2 = .00$. In summary, dive task accuracy was poorer with the use of low or high DOA, but there were no return-to-manual deficits for dive accuracy following the use of low or high DOA.

For dive task RT, there was no main effect of condition or Condition \times State interaction; thus, no follow-up simple effect analyses were conducted.

Workload

For the ATWIT subjective workload measure, there was a main effect of condition and a Condition \times State interaction. The simple effect test revealed a significant difference between conditions during routine states, $F(2, 117) = 4.94, p = .01, \eta^2 = .08$. During routine states, there was no significant difference in ATWIT ratings between the no automation and low DOA conditions, $t < 1$. Participants provided high DOA reported lower ATWIT ratings than participants provided no automation, $t(78) = 2.64, p = .02, d = 0.59$, or low DOA, $t(78) = 2.64, p = .03, d = 0.90$. For automation removal state, the simple effect test revealed a significant difference

between the conditions, $F(2, 119) = 4.95, p = .01, \eta^2 = .08$. After automation removal, participants previously using low DOA made higher ATWIT ratings than participants provided no automation, $t(80) = 3.23, p = .01, d = 0.71$. There was no significant difference in ATWIT ratings between the low DOA and high DOA conditions, $t < 1$, or the high DOA and no automation conditions, $t < 1$.

For NASA-TLX, there was a Condition \times State interaction. However, the simple effect tests revealed no significant difference between the conditions during routine states, $F(2, 118) = 2.52, p = .09, \eta^2 = .04$, or during automation removal states, $F(2, 118) = 2.83, p = .06, \eta^2 = .05$.

In summary, high DOA reduced workload during routine states, and workload was increased after low DOA removal, but only as measured by ATWIT. Although the effects trended in the same direction (Table 3), they did not reach significance when workload was measured by the NASA-TLX.

SA

For SA, there was a Condition \times State interaction. The simple effect test revealed a significant difference between the conditions during routine states, $F(2, 117) = 5.20, p = .01, \eta^2 = .08$. During routine states, participants provided low DOA made less accurate SAGAT responses than participants provided no automation, $t(78) = 3.10, p = .01, d = 0.66$. There was no difference in SAGAT accuracy between the low DOA and high DOA conditions, $t < 1$, or the high DOA and no automation conditions, $t(78) = 2.35, p = .07$. For automation removal state, a simple effect test on SAGAT accuracy revealed no significant difference between the conditions, $F(2, 119) = 0.50, p = .61, \eta^2 = .01$. In summary, SAGAT accuracy was poorer with the use of low DOA, and there were no return-to-manual deficits for SAGAT following the use of low or high DOA.

Discussion

The aim of Experiment 1 was to examine the impact of low DOA and high DOA on performance, workload, and SA both during routine states and after automation removal. Our predictions regarding the effects of DOA were summarized in Table 1. These predictions were

based on the findings of a previous experiment in this task domain (Chen et al., 2017) and on the Onnasch et al. (2014) meta-analysis. Our findings are summarized in Table 5.

The use of low DOA in Experiment 1 benefited classification RT, but no other automated task performance metrics, compared with when no automation was provided. Furthermore, workload was not reduced with the use of low DOA, and workload increased (as measured by ATWIT) when low DOA was removed compared with no automation. There were also costs to dive task accuracy and SA with the use of low DOA compared with no automation during routine states, but these costs disappeared after the automation was removed. Other than the fact that we did not find a benefit to classification accuracy with low DOA, these findings for the low DOA condition compared with the no automation condition replicate Chen et al. (2017).

The use of high DOA benefited classification (accuracy/RT), CPA (accuracy), and lowered workload (as measured by ATWIT, but not the NASA-TLX) compared with the use of no automation. The use of high DOA also benefited classification (accuracy), CPA (accuracy/RT), and lowered workload (as measured by ATWIT) compared with the use of low DOA. The use of high DOA did not cost SA compared with no automation, but did impair dive task accuracy (but no more than when compared with low DOA). Contrary to our predictions made on the basis of the Onnasch et al. (2014) meta-analysis, high DOA removal did not cost classification/CPA performance, workload, SA, or dive task performance compared with no automation or low DOA.

EXPERIMENT 2

In Experiment 1, the use of high DOA provided several benefits to automated task performance and workload, without costs to SA, dive task performance, or return-to-manual control, when compared with the use of low DOA. It is evident that high DOA is superior to low DOA, at least in the context of simulated submarine track management, and therefore we did not further test low DOA in Experiment 2.

The removal of automation in Experiment 1 can be linked to the type of automation failure

TABLE 5: Summary of Findings From Experiment 1

Task	Routine	Matches Prediction	Removal	Matches Prediction
Classification				
Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	Partial	None = Low = High (no RTM effects)	Partial
RT	None > [Low = High] (faster decisions with either low or high DOA)	Partial	None = Low = High (no RTM effects)	Partial
CPA				
Accuracy	[None = Low] < High (benefits to accuracy with high DOA)	Yes	[None = Low] < High (higher accuracy after high DOA removal)	Partial
RT	[High = None] < Low (slower decisions with low DOA)	Partial	None = Low = High (no RTM effects)	Partial
Dive				
Accuracy	None > [Low = High] (poorer accuracy with either low or high DOA)	Partial	None = Low = High (no RTM effects)	Partial
RT	None = Low = High (no difference in RT)	No	None = Low = High (no RTM effects)	Partial
Workload				
(ATWIT)	[None = Low] > High (reduced workload with high DOA)	Yes	[None < Low] = High (higher workload after low DOA removal)	No
Workload (NASA-TLX)	None = Low = High (no difference in workload)	Partial	None = Low = High (no RTM effects)	Partial
SA	[None > Low] = High (poorer SA with low DOA)	Partial	None = Low = High (no RTM effects)	No

Note. Routine = automation is reliable; Removal = after automation is removed; Gray shading = observed result matches predicted result; None = no automation; Low = low DOA; High = high DOA; DOA = degree of automation; RTM = return to manual; RT = response time; CPA = closest point of approach. ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index; SA = situation awareness.

that Wickens, Clegg, et al. (2015) referred to as “automation gone,” in which automation is removed. Wickens, Clegg, et al. (2015) also noted that automation may not be removed, but rather begin to provide incorrect information—a condition they referred to as “automation wrong.” In comparing these two types of failures, Wickens, Clegg, et al. (2015) found that operators had more difficulty detecting and compensating for automation wrong failures

than automation gone failures. Here, in Experiment 2, we aim to test whether these two types of automation failures have differential effects when individuals use high DOA.

Participants completed three scenarios, and during the last scenario, the automation was either removed (automation gone) or incorrect (automation wrong). The automation gone and automation wrong conditions were identical during routine states in that high DOA provided decision

recommendations for the classification and CPA tasks. However, the conditions differed when the automation was removed/failed. For the automation gone condition, the automation was removed as in Experiment 1. For the automation wrong condition, the automation started providing incorrect advice for the classification task. Participants from both automation conditions were instructed to report as soon as they noticed that automation was providing wrong advice.

Of particular interest was whether participants in the automation wrong condition would notice the automation failure, and if they did how long it would take them to do so. We were also interested in whether there would be any performance deficits immediately following the automation failure. Specifically, we examined performance immediately following the automation failure by analyzing classification performance (the task on which the automation was providing incorrect recommendations) on the first three classification events after the automation failure. To the extent that participants take some time to detect that the automation is providing incorrect classification recommendations, we predicted that classification accuracy on the first three events after the automation failure could be poorer, and classification RT slower, for the automation wrong condition compared with the no automation condition. In contrast, we did not expect to see classification accuracy or RT deficits immediately following the automation failure for the automation gone condition compared with the no automation condition because participants were notified that automation was no longer available, and the evidence to date from Chen et al. (2017) and the current Experiment 1 suggests that participants should be able to regain manual control relatively quickly under these circumstances.

In addition to these aforementioned novel analyses, we expected to replicate the benefits to automated task performance and workload, and deficits to the dive task, for the two high DOA conditions compared with the no automation condition during routine states. The importance of establishing the robustness of psychological effects has received much recent attention (e.g., Pashler & Wagenmakers, 2012) and is particularly vital when the resulting knowledge could

be used by practitioners in safety-critical work settings (Jones, Derby, & Schmidlin, 2010). The replication is also important because unlike Experiment 1, in Experiment 2 participants who were using high DOA were instructed that although automation was highly reliable, it may not be perfect, which may decrease the extent to which they trust and rely on the automated recommendations (see Lee & See, 2004). Automation removal state in Experiment 2 was defined as the time after automation was removed for the automation gone condition (as in Experiment 1), and as the time period after participants detected the automation failure for the automation wrong condition. Note that correctly reporting the automation failure resulted in the automation being disengaged (but participants were not informed that this would occur). Based on Experiment 1 results, we did not expect return-to-manual deficits during the automation removal state for the high DOA conditions compared with the no automation condition.

An additional goal of Experiment 2 was to examine how participants rated the importance of each of the three tasks. In Experiment 1, we found significant costs to the dive task that replicated those reported by Chen et al. (2017). We wanted to investigate the possibility that the dive task deficit was due to participants placing less importance on their performance on the dive task compared with the other two tasks due to the fact that the dive task was the only task that was not automated.

Method

Participants. Participants were 120 (70 females) undergraduate psychology students (age: $M = 21.7$, $SD = 6.17$) who participated for course credit and were randomly assigned to one of three conditions: no automation ($n = 40$), automation gone ($n = 40$), and automation wrong ($n = 40$).

Simulated submarine track management task. The simulation was identical to the no automation and high DOA conditions from Experiment 1 with the following exceptions. For the automation gone and automation wrong conditions, the automation was unexpectedly removed or failed during the last scenario at 10.38, 10.48, or 10.88 min into the 27.5-min

scenario. The message provided to participants in the automation gone condition was identical to that used for the high DOA condition in Experiment 1. In the automation wrong condition, the automation started providing incorrect advice for the classification task. Specifically, the horizontal lines were placed either 30 s too early or late on the soundtracks, and the recommended classification was incorrect (e.g., if the contact was an enemy, the classification letter presented next to the line was f, t, or m).

The Track Assist interface was modified to include a “fail” button. This button was available from the beginning of each scenario. Participants (in both automation conditions) were instructed to click this button if they believed the automation was providing wrong advice. When clicked, a message appeared saying “Automation failure detected. Track Assist turned off. Manual tracking required” and participants had to acknowledge that they had read this message by clicking the “ok” button. If the fail button was clicked when the automation was functioning correctly, a message appeared saying “Automation has not failed” and the automation continued operating as usual.

Measures and procedure. Workload and SA measures were identical to Experiment 1. Participants rated the perceived importance of each task on a 5-point Likert-type scale (1 = *not at all important* to 5 = *extremely important*) after the last scenario. The training was the same as in Experiment 1, but there was an additional instruction specifying that although the automation was highly reliable, it may not be perfect, and participants were instructed how to report an automation failure. Each participant completed three scenarios in their assigned condition, and the order of scenarios was counterbalanced.

Results

The mean RT to detect the automation failure by participants in the automation wrong condition was 174.13 s; 95% CI = [125.64, 222.61]. As seen in Figure 2, 50% of the participants in the automation wrong condition had not reported the automation failure 173.87 s after the failure occurred. Three participants in the automation wrong condition did not detect the automation failure at all.

Classification Task Performance Immediately Following the Automation Failure

We analyzed performance on the first three classification events after the automation failure. The classification accuracy and RT data for these three classification events are presented in Figure 3. To test our predictions, we ran 3 Condition (no automation, automation gone, automation wrong) \times 3 Classification Event (first event after failure, second event after failure, third event after failure) mixed ANOVAs on classification accuracy and on classification RT. The between-subjects factor was condition and the within-subjects factor was classification event. We planned to follow-up significant main effects of condition, or interactions between condition and classification event, with tests of simple effects separately (with Bonferroni corrections), comparing the three conditions on each classification event.

For classification accuracy, it was predicted that performance immediately after the automation failure would be poorer for the automation wrong condition compared with the no automation condition, but there would be no difference in performance between the automation gone and no automation conditions. A mixed ANOVA on classification accuracy revealed a main effect of classification event, $F(2, 234) = 3.21, p = .04, \eta_p^2 = .03$, but no main effect of condition, $F(1, 117) = 0.55, p = .587, \eta_p^2 = .01$, and no interaction effect $F(2, 234) = 2.24, p = .07, \eta_p^2 = .04$.

For classification RT, it was predicted that RT would be slower immediately after the automation failure for the automation wrong condition compared with the no automation condition, but that there would be no difference in RT between the automation gone and no automation conditions. A mixed ANOVA on classification RT revealed no main effect of classification event, $F(2, 128) = 1.62, p = .20, \eta_p^2 = .03$; no main effect of condition, $F(1, 64) = 2.09, p = .13, \eta_p^2 = .06$; and no interaction effect, $F(2, 128) = 0.98, p = .42, \eta_p^2 = .03$.

In brief, there were no reliable differences in classification accuracy or RT between the conditions for the three classification events after

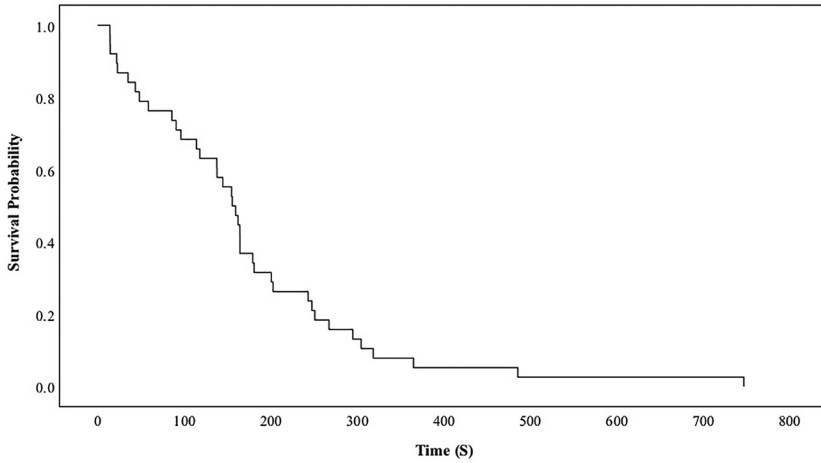


Figure 2. A Kaplan–Meier survival analysis representing the time (in seconds) taken for participants in the automation wrong condition to detect the automation wrong failure.

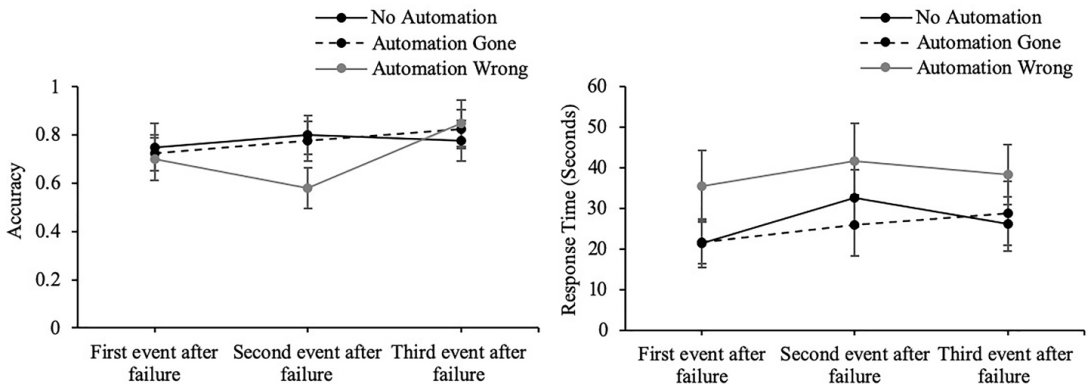


Figure 3. Classification accuracy (left graph) and RT (right graph) for the first three classification events after the automation failure, as a function of automation condition. Error bars represent 95% between-subjects confidence intervals.

the automation failure. Thus, performance immediately after the automation failure was not poorer for the automation wrong condition compared with the no automation condition. As predicted, there was no difference in performance between the automation gone and no automation conditions.

Routine and Automation Removal States

We then conducted analyses to replicate the results from Experiment 1. The data were

separated into the time period automation was available (routine state: first two scenarios and one third of the last scenario) and time period automation was removed (automation removal: two thirds of the last scenario for the automation gone condition, and from the time point, automation was turned off by the participant for the automation wrong condition). The three participants in the automation wrong condition who never reported the automation failure were excluded from the analyses reported in the following text.

TABLE 6: Descriptive Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 2

Automation	Classification		CPA		Dive		SAGAT	Workload Rating	
	Hit	RT	Hit-FA	RT	Hit-FA	RT	Accuracy	ATWIT	NASA-TLX
Routine state									
None	0.73	30.82	0.38	21.09	0.72	9.88	0.53	4.98	61.54
	[0.65, 0.81]	[26.37, 35.28]	[0.30, 0.46]	[15.27, 26.92]	[0.67, 0.77]	[9.09, 10.66]	[0.48, 0.58]	[4.57, 5.39]	[57.87, 65.21]
High	0.94	21.41	0.80	11.97	0.59	9.46	0.51	4.11	52.09
	[0.90, 0.96]	[19.91, 22.92]	[0.74, 0.86]	[10.44, 13.50]	[0.55, 0.64]	[8.73, 10.19]	[0.49, 0.54]	[3.81, 4.40]	[48.79, 55.38]
Automation removal state									
None	0.72	28.61	0.28	22.18	0.77	9.88	0.54	5.19	61.66
	[0.62, 0.82]	[23.37, 34.05]	[0.19, 0.36]	[15.63, 28.73]	[0.69, 0.85]	[7.93, 11.84]	[0.49, 0.59]	[4.77, 5.61]	[57.73, 66.78]
High	0.77	31.80	0.36	19.95	0.77	9.60	0.53	5.50	58.06
	[0.72, 0.82]	[28.26, 35.34]	[0.31, 0.41]	[14.54, 25.36]	[0.72, 0.83]	[8.61, 10.59]	[0.50, 0.57]	[5.15, 5.86]	[54.62, 61.51]

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach; SAGAT = Situation Awareness Global Assessment Technique; RT = response time; FA = false alarm; ATWIT = Air Traffic Workload Input Technique; NASA-TLX = National Aeronautics and Space Administration Task Load Index.

The automation gone and automation wrong conditions were identical during routine states in that they provided high DOA for the classification and CPA tasks, and identical at automation removal in that all participants knew the automation was no longer available (removed, or detected as having failed and removed). On this basis, we combined the data from the two high DOA conditions from Experiment 2. The means and between-subjects 95% CIs for performance, workload, and SA are presented in Table 6.

We compared the two high DOA conditions to the no automation condition with 2 Condition (high DOA, no automation) \times 2 Automation State (routine, automation removal) mixed ANOVAs with the between-subjects factor as condition and the within-subjects factor as automation state. The inferential statistics from the ANOVAs are summarized in Table 7. Significant main effects of condition, and interactions between condition and automation state, were followed by comparisons of simple effects (t tests) conducted separately for the routine state and the automation removal state, and are presented in text. We had a power of 0.82 to detect a medium-to-large

effect size (Cohen, 1988). As in Experiment 1, several main effects of automation state were found. They are reported in Table 7 but for brevity are not further discussed.

Automated Task Performance

Classification task. For classification accuracy, there was a main effect of condition and a Condition \times State interaction. During routine states, participants provided high DOA made more accurate contact classifications compared with participants provided no automation, $t(118) = 5.89, p < .001, d = 1.02$. After automation removal, there was no difference in classification accuracy between the conditions, $t < 1$. For classification RT, there was a Condition \times State interaction. During routine states, participants provided high DOA made faster classifications than those provided no automation, $t(118) = 4.99, p < .001, d = 0.86$. After automation removal, there was no difference in classification RT between the conditions, $t < 1$. These findings for the classification task replicate Experiment 1.

CPA task. For CPA accuracy, there was a main effect of condition and a Condition \times State

TABLE 7: Inferential Statistics for Performance, Subjective Workload, and Situation Awareness by Condition and Automation State in Experiment 2

Dependent Variable	High Degree of Automation vs. No Automation				
	Effect	<i>F</i>	<i>df</i>	<i>p</i>	η_p^2
Classification (Hit)	Condition	11.24	(1, 115)	.001*	.09
	State	33.50	(1, 115)	<.001*	.23
	Interaction	23.99	(1, 115)	<.001*	.17
Classification (RT)	Condition	2.22	(1, 112)	.14	.02
	State	6.34	(1, 112)	.02*	.05
	Interaction	18.21	(1, 112)	<.001*	.14
CPA (Hit-FA)	Condition	37.87	(1, 115)	<.001*	.25
	State	111.30	(1, 115)	<.001*	.49
	Interaction	43.20	(1, 115)	<.001*	.27
CPA (RT)	Condition	3.63	(1, 112)	.06	.03
	State	5.24	(1, 112)	.02*	.05
	Interaction	2.31	(1, 112)	.13	.02
Dive (Hit-FA)	Condition	2.55	(1, 115)	.11	.02
	State	39.40	(1, 115)	<.001*	.26
	Interaction	11.11	(1, 115)	.001*	.09
Dive (RT)	Condition	0.55	(1, 114)	.46	.01
	State	0.03	(1, 114)	.87	<.001
	Interaction	0.02	(1, 114)	.88	<.001
NASA-TLX	Condition	6.37	(1, 118)	.01*	.05
	State	7.00	(1, 118)	.01*	.06
	Interaction	6.46	(1, 118)	.01*	.05
ATWIT	Condition	1.13	(1, 115)	.29	.01
	State	55.29	(1, 115)	<.001*	.33
	Interaction	29.65	(1, 115)	<.001*	.21
SAGAT (Accuracy)	Condition	0.19	(1, 115)	.67	.002
	State	0.50	(1, 115)	.48	.004
	Interaction	1.00	(1, 115)	.76	.001

Note. RT = response time; CPA = closest point of approach; FA = false alarm; NASA-TLX = National Aeronautics and Space Administration Task Load Index; ATWIT = Air Traffic Workload Input Technique; SAGAT = Situation Awareness Global Assessment Technique.

**p* < .05.

interaction. During routine states, participants provided high DOA made more accurate CPA task decisions compared with participants provided no automation, $t(118) = 8.28, p < .001, d = 1.63$. After automation removal, there was no difference in CPA accuracy between the conditions, $t < 1$. For CPA RT, there was no main effect of condition or Condition \times State interaction. These findings for the CPA task replicate Experiment 1.

Non-Automated Task Performance

For dive task accuracy, there was a Condition \times State interaction. During routine states, participants provided high DOA made less accurate dive task decisions compared with participants provided no automation, $t(118) = 3.37, p = .001, d = 0.68$. After automation removal, there was no difference in dive accuracy between the conditions, $t < 1$. For dive RT, there was no main

TABLE 8: Descriptive Statistics for Task Importance Ratings by Condition in Experiment 2

Automation	Classification	CPA	Dive
None	4.20 [3.92, 4.48]	2.98 [2.65, 3.30]	3.75 [3.39, 4.11]
Gone	4.18 [3.89, 4.46]	3.08 [2.75, 3.40]	3.60 [3.26, 3.94]
Wrong	4.15 [3.86, 4.44]	3.05 [2.73, 3.37]	3.48 [3.11, 3.84]

Note. The 95% between-subjects confidence intervals are presented in brackets. CPA = closest point of approach.

effect of condition or Condition \times State interaction. These findings for the dive task replicate Experiment 1.

Workload

For ATWIT, there was a Condition \times State interaction. During routine states, participants provided high DOA made lower ATWIT ratings compared with participants provided no automation, $t(118) = 3.44, p = .001, d = 0.67$. After automation removal, there was no difference in ATWIT ratings between the conditions, $t < 1$. These findings for the ATWIT replicate Experiment 1. In addition, for NASA-TLX, the main effect of condition and Condition \times State interaction reached significance. During routine states, participants provided high DOA made lower NASA-TLX ratings compared with participants provided no automation, $t(118) = 3.54, p = .001, d = 0.71$. After automation removal, there was no difference in NASA-TLX ratings between the conditions, $t < 1$.

SA

There was no main effect of condition or Condition \times State interaction, replicating Experiment 1.

Task Importance

The task importance ratings are presented in Table 8. A 3 Condition (no automation, automation gone, automation wrong) \times 3 Task Type (classification, CPA, dive) mixed ANOVA on task importance ratings revealed a main effect of task type, $F(2, 234) = 47.22, p < .001, \eta_p^2 = .29$. Participants rated the classification task as being more important than the CPA task, $t(119) = 10.57, p < .001, d = 1.20$, and the dive task, $t(119) = 5.20, p < .001, d = 0.56$. In addition, the dive task was rated as being more important than the CPA task, $t(119) = 4.35, p < .001, d = 0.55$.

There was no main effect of condition and no interaction. Thus, there were no differences in dive task importance ratings between the conditions, suggesting that participants in the automated conditions did not place less importance on their performance on the dive task compared with participants not provided automation.

GENERAL DISCUSSION

In Experiment 1, we examined the effects of low and high DOA on performance, workload, SA, non-automated task performance, and return-to-manual performance. Low DOA provided information acquisition and analysis support. High DOA provided decision recommendation support while still requiring participants to execute task actions. Participants completed two tasks that were supported by the automation (classification and CPA), and one task that was not supported by automation (dive). In Experiment 2, when automation failed, it was either removed completely (automation gone condition), as in Experiment 1, or started providing incorrect advice for the classification task (automation wrong condition). We examined whether participants would notice the automation wrong failure and if so, how long it would take them to do so. We also examined whether there would be any performance deficits immediately following the automation failure on the classification task. Furthermore, in Experiment 2, we expected to replicate the benefits to automated task performance and workload, and the costs to non-automated task performance, with the use of high DOA that were found in Experiment 1.

The Benefits and Costs of Low and High DOA

We found little evidence of a benefit in using low DOA. Only one of the four automated task performance metrics (classification

RT) improved, and there was no reduction in workload. In addition, there were costs to dive task accuracy and SA with the use of low DOA compared with no automation during routine states. These findings largely replicate those reported by Chen et al. (2017).

Compared with the use of low DOA, the use of high DOA benefited three automated task performance metrics (classification accuracy, and CPA accuracy/RT). Participants also reported lower workload with the use of high DOA compared with low DOA and no automation. In addition, although high DOA did cost non-automated task performance compared with no automation, the extent of this cost was not larger than that for the low DOA condition. Also, in Experiments 1 and 2, after the automation was no longer available (removed, or detected as having failed and removed), it was not more difficult for participants previously using high DOA to regain manual control. Therefore, participants were able to effectively return-to-manual control after knowing that automation was no longer available. Overall, we have found evidence that under some conditions, it is possible that moving from a low DOA to a high DOA can provide a “free lunch,” that is, it can increase the benefits of automation without further increasing the costs (see Wickens, 2018).

At first glance, our findings of increased benefits without further costs when using high compared with low DOA seem inconsistent with the Onnasch et al. (2014) meta-analytic finding that the negative consequences of automation are more likely, the higher the DOA. However, close inspection of Onnasch et al. indicates that their meta-analysis contained substantial variance in effect size between studies, with trends and effects ranging from strong to weak or even reversed. This variance is likely due to the variability in the nature of the tasks used across studies included in the Onnasch et al. meta-analysis. In addition, many of the studies in the Onnasch et al. meta-analysis used relatively fast evolving tasks such as air traffic control and unmanned vehicle control. In contrast, a key feature of submarine track management is the very slow pace in which contacts move on the display. High DOA may have indeed reduced the extent to which participants actively processed raw infor-

mation (e.g., contact position and heading) (complacency; Parasuraman & Manzey, 2010), but the effect of this may have been attenuated by the slow pace of the task that allowed sufficient time for participants to recover. Furthermore, Onnasch et al. suggest that the negative consequences of automation were the strongest when DOA moved from supporting information acquisition and analysis to also supporting action selection/action execution. It is worth noting that in the current study, high DOA did not cross the boundary between decision recommendation and action execution because participants were still required to execute the final task action.

The benefits to automated task performance and workload for the high compared with low DOA and no automation conditions (Experiment 1), and for the high DOA conditions compared with the no automation condition (Experiment 2), were reasonably consistent. In Experiments 1 and 2, there were also clear and consistent costs to dive task accuracy during routine states with the use of low and high DOA compared with the no automation condition. Even with the reduction in workload with the use of high DOA compared with no automation, performance on the dive task degraded. It would have been reasonable to expect that the reduced workload associated with the use of high compared with no automation should have provided the operator with the additional cognitive capacity to more effectively manage the non-automated dive task (Manzey et al., 2012; Rovira, McGarry, & Parasuraman, 2007). Nevertheless, reduced workload with high DOA would only have benefited dive task performance to the extent that the spared capacity was directly allocated toward scrutinizing contact location and heading information. It seems that participants who were provided with high DOA for classification and CPA tasks scrutinized contact location and heading information less closely than participants who were not provided with automation (complacency), and the dive task likely suffered because it also required assessment of contact location/heading information. To further test this explanation, future research could examine performance on a non-automated task that is independent of the automated tasks. Performance on an independent

non-automated task should be the same or if not better for those who receive high DOA compared with those who receive no automation, due to the spare cognitive capacity from the reduction in workload with the use of automation. Note that in Experiment 2, we ruled out the possibility that the dive task deficit could be due to participants provided with automation placing less importance on the dive task compared with the two other automated tasks.

In Experiment 2, we predicted that classification performance immediately after the high DOA failure would be poorer for the automation wrong condition compared with the no automation condition. However, this prediction was not supported. Interestingly, although it took participants on average 3 min to report the automation wrong failure, they were still able to correctly classify contacts immediately after the failure as successfully as the no automation condition. In post hoc analyses, at each of the three classification events immediately after the failure, we split participants in the automation wrong condition according to whether they had reported the automation failure or not. There was still no significant difference in classification accuracy or RT on the three classification events immediately after the automation failure for the subset of participants who had not yet reported the failure, compared with the no automation condition or automation gone condition. This suggests that participants may have become suspicious about the accuracy of the automation and started to make their own manual classification decisions, but decided to allow some time to clarify and ensure that the automation was not performing accurately before they formally reported the failure.

PRACTICAL IMPLICATIONS AND CONCLUSIONS

A key question in complex work systems is to what extent decision recommendation automation can be effectively used. The results of the current study suggest that automation that recommends decisions can be effectively used and, in the current context, was superior to a low DOA that only provided information acquisition and analysis support. Specifically, automation that recommended decisions leads to performance

and workload benefits without any costs to SA or return-to-manual performance, compared with automation that provided information acquisition and analysis support. Although the current study used a simulation of submarine track management, the findings of this work are also relevant to other work contexts, particularly those involving slowly evolving contexts that require operators to monitor demanding perceptual displays (e.g., maritime surveillance).

Automating tasks can improve operator performance and reduce workload, but accidents have occurred because human operators have been unprepared to take over when manual control is required. If automation fails, the operator's ability to resume manual control is critical. In the current study, although participants were able to regain manual control, it took on average 3 min for them to detect that automation was providing incorrect advice. As discussed, the fact there was no decrement to classification performance immediately following the automation failure suggests at least some participants who had not indicated that there was a failure were suspicious that automation may have not been performing accurately and were making manual classification decisions. Nonetheless, in a context where there is more time pressure (e.g., unmanned vehicle control, air traffic control) to make a manual decision, such a delay could be catastrophic. Future research could examine whether operators would still take as long to formally register an automation wrong failure in a faster updating task. Before implementing automation that recommends decisions, designers should carefully consider the level of uncertainty in the environment (i.e., the chance that automation may be incorrect) and the operational consequences of a loss of SA or return-to-manual performance deficits.

The simulated submarine track management task used in the current study was designed based on a task analysis conducted with Royal Australian Navy Submariners. Accordingly, the current experiments have external validity as they represent a typical example of a work context that requires operators to monitor demanding perceptual displays. That said, we are aware of the potential problems in generalizing from novice participants to expert operators as there

are undoubtedly differences in their cognitive skills and motivation. Future research could examine how expert submariners are affected differently by DOA and task type in the current simulated submarine track management task. There is, however, evidence to suggest that our results with novice participants can validly inform practical issues in operational contexts. A study by Loft et al. (2016) found relatively consistent results across novice participants using the current simulated submarine track management task and expert submariners using real submarine combat systems. In addition, Onnasch et al. (2014) found that expertise did not moderate the benefits and costs of automation; thus, benefits and costs were as statistically likely to occur for experts as they were for novice participants.

In conclusion, the automated system that recommended decisions was effectively utilized by participants in the current context and appeared to be superior to the automated system that supported information acquisition and analysis. Automation that recommends decisions is appropriate in contexts where the consequences of an automation failure are not serious enough to outweigh the benefits. However, designers should be cautious and consider the level of uncertainty in the environment and the consequences of a loss of SA or return-to-manual performance deficits before implementing decision-aiding automation. In contexts where return to manual performance is of serious concern, operators should be kept involved in the action selection and execution stages.

ACKNOWLEDGMENT

This research was supported by Discovery Grant DP160100575 awarded to Loft from the Australian Research Council.

KEY POINTS

- With the ongoing emphasis on developing high degree of automation (DOA) that can recommend decisions to operators, it is critical to further understand how high DOA systems affect the human operator.
- In a simulated submarine track management task, high DOA that provided decision recommendations provided benefits to performance and workload,

without additional costs to SA or non-automated task performance, compared with low DOA.

- There were no return-to-manual deficits when participants had knowledge that low DOA or high DOA was disengaged.
- Participants using high DOA took on average 3 min to notice that automation was providing incorrect recommendations, but there was no deficit to performance immediately following the automation failure.
- Designers should consider the level of uncertainty in the environment and the consequences of a loss of situation awareness (SA) or return-to-manual deficits before implementing decision-aiding automation.

ORCID iDS

Monica Tatasciore  <https://orcid.org/0000-0001-7290-0225>

Troy A. W. Visser  <https://orcid.org/0000-0003-3960-2263>

Steph I. C. Michailovs  <https://orcid.org/0000-0002-6767-6692>

REFERENCES

- Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*, 1065–1105. doi:10.1177/0149206309352880
- Chen, S. I., Visser, T. A. W., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*, 240–262. doi:10.1037/xap0000126
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *32*, 97–101.
- Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*, 65–84. doi:10.1518/001872095779049499
- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, *59*, 5–27. doi:10.1177/0018720816681350
- Hart, S. G., & Staveland, L. E. (1987). Development of NASA-TLX: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, The Netherlands: Elsevier.
- Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors*, *52*, 586–595. doi:10.1177/0018720810384394
- Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*, 113–153. doi:10.1080/1463922021000054335

- Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *10*, 409–430.
- Kirschenbaum, S. S. (2011). Expertise in the submarine domain: The impact of explicit display on the interpretation of uncertainty. In K. L. Mosier & U. M. Fischer (Eds.), *Informed by knowledge: Expert performance in complex situations* (pp. 189–199). New York, NY: Psychology Press.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. doi:10.1518/hfes.46.1.50_30392
- Li, H., Wickens, C. D., Sarter, N., & Sebok, A. (2014). Stages and levels of automation in support of space teleoperations. *Human Factors*, *56*, 1050–1061. doi:10.1177/0018720814522830
- Loft, S., Morrell, D. B., Ponton, K., Braithwaite, J., Bowden, V., & Huf, S. (2016). The impact of uncertain contact location on situation awareness and performance in simulated submarine track management. *Human Factors*, *58*, 1052–1068. doi:10.1177/0018720816652754
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*, 57–87. doi:10.1177/1555343411433844
- Merlo, J. L., Wickens, C. D., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. In *Proceedings of the 4th Annual Army Federated Laboratory Symposium* (pp. 27–31). College Park, MD: Army Research Federated Laboratory Consortium.
- 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*, 476–488. doi:10.1177/0018720813501549
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*, 381–410. doi:10.1177/0018720810376055
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced “complacency.” *The International Journal of Aviation Psychology*, *3*, 1–23. doi:10.1207/s15327108ijap0301_1
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253. doi:10.1518/001872097778543886
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *30*, 286–297.
- Pashler, H., & Wagenmakers, E. J. (2012). Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253
- Roberts, A., Stanton, N. A., & Fay, D. (2017). The command team experimental test-bed phase two: Assessing cognitive load and situation awareness in a submarine control room. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (pp. 427–437). Berlin, Germany: Springer.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. doi:10.1111/j.1745-6916.2006.00012.x
- 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.
- Sheridan, T. B. (2015). Automation. In D. A. Boehm-Davis, F. T. Durso, & D. L. John (Eds.), *APA handbook of human systems integration* (pp. 449–465). Washington, DC: American Psychological Association.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge: Man-Machine Systems Laboratory, Massachusetts Institute of Technology.
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.
- Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe*. Atlantic City, NJ: Federal Aviation Administration.
- U.S. Air Force. (2015). *Autonomous horizons*. Washington, DC: U.S. Air Force Office of the Chief Scientist.
- Vagia, M., Transeth, A. A., & Fjerdingen, S. A. (2016). A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, *53*, 190–202. doi:10.1016/j.apergo.2015.09.013
- Wickens, C. D. (2018). Automation stages & levels, 20 years after. *Journal of Cognitive Engineering and Decision Making*, *12*, 35–41. doi:10.1177/1555343417727438
- Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors*, *57*, 728–739. doi:10.1177/0018720815581940
- 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*, 959–975. doi:10.1177/0018720814566454
- Monica Tatasciore is a master’s student enrolled in the Doctor of Philosophy and Master of Industrial and Organizational Psychology program at the University of Western Australia.
- Vanessa K. Bowden is a lecturer at the University of Western Australia. She received her PhD in psychology in 2012 from the University of Western Australia.
- Troy A. W. Visser is an associate professor at the University of Western Australia. He received his PhD in cognitive systems in 2001 from the University of British Columbia.
- Steph I. C. Michailovs is a postdoctoral research fellow at the University of Western Australia. She received her PhD in psychology in 2019 from the University of Western Australia.
- Shayne Loft is an associate professor at the University of Western Australia. He received his PhD in psychology in 2004 from the University of Queensland.

Date received: November 13, 2018

Date accepted: July 3, 2019