Should We Just Let the Machines Do It? The Benefit and Cost of Action Recommendation and Action Implementation Automation

Monica Tatasciore, **Vanessa K. Bowden**, **Troy A. W. Visser**, and **Shayne Loft**,
The University of Western Australia, Australia

Objective: To examine the effects of action recommendation and action implementation automation on performance, workload, situation awareness (SA), detection of automation failure, and return-to-manual performance in a submarine track management task.

Background: Theory and meta-analytic evidence suggest that with increasing degrees of automation (DOA), operator performance improves and workload decreases, but SA and return-to-manual performance declines.

Method: Participants monitored the location and heading of contacts in order to classify them, mark their closest point of approach (CPA), and dive when necessary. Participants were assigned either no automation, action recommendation automation, or action implementation automation. An automation failure occurred late in the task, whereby the automation provided incorrect classification advice or implemented incorrect classification actions.

Results: Compared to no automation, action recommendation automation benefited automated task performance and lowered workload, but cost nonautomated task performance. Action implementation automation resulted in perfect automated task performance (by default) and lowered workload, with no costs to nonautomated task performance, SA, or return-to-manual performance compared to no automation. However, participants provided action implementation automation were less likely to detect the automation failure compared to those provided action recommendations, and made less accurate classifications immediately after the automation failure, compared to those provided no automation.

Conclusion: Action implementation automation produced the anticipated benefits but also caused poorer automation failure detection.

Application: While action implementation automation may be effective for some task contexts, system designers should be aware that operators may be less likely to detect automation failures and that performance may suffer until such failures are detected.

Keywords: automation, situation awareness, workload, return-to-manual control, submarine track management

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

HUMAN FACTORS
2022, Vol. 64(7) 1121–1136
DOI:10.1177/0018720821989148
Article reuse guidelines: sagepub.com/journals-permissions

Humans increasingly need to interact with automation designed to improve workplace efficiency and safety. Automation is defined as “a device or system that accomplishes a function that was previously, or conceivably could be, carried out by a human operator” (Parasuraman et al., 2000, p. 287). Examples include diagnostic decision aids in health care, conflict detection automation in air traffic control, and recommender systems in unmanned vehicle control. Reliable automation usually exceeds operator manual performance and can reduce operator workload (Onnasch et al., 2014). However, automation can also reduce operators’ understanding of a task, their ability to anticipate future events (situation awareness [SA]; Endsley, 1988), and impair return-to-manual performance if automation unexpectedly fails.

Costs and benefits of automation vary systematically depending on the degree of automation (DOA). DOA orders the level of automation support (complete manual control to complete autonomy; Sheridan & Verplank, 1978) by four stages of information processing: information acquisition, information analysis, action recommendation, and action implementation (Parasuraman et al., 2000). Thus, higher levels and later stages of automation increase the DOA. A meta-analysis by Onnasch et al. (2014) found that increased DOA improves performance and reduces workload, but costs SA and return-to-manual performance (labeled the “lumberjack effect”). Onnasch et al. (2014) also identified a “critical boundary” where DOA begins to support action recommendation, after which costs of automation are more likely to drastically increase.

We recently published two papers examining this lumberjack effect in simulated submarine track management (Chen et al., 2017; Tatasciore et al., 2020). Here, participants monitored a “surface plot” display (which showed the location and heading of contacts in relation to the participant’s submarine, referred to as “Ownship”) and a “waterfall” display (which showed contact bearings over time) in order to perform tasks including contact classification (defined by time spent within display regions), closest point of approach (CPA; detecting when contacts turned away from Ownship), and deciding when Ownship should dive.

Tatasciore et al. (2020) automated the classification and CPA tasks using a “low” or “high” DOA. Low DOA supported information acquisition and analysis by displaying how long contacts spent within display regions (classification) and by tracking contact heading changes (CPA). High DOA provided this same information, but also gave action recommendations regarding contact classification type and CPA occurrence. The dive task was never automated. Automation failed late in the experiment, either by shutting off with a message to participants to resume manual control (automation “gone”) or by providing incorrect contact classification information (automation “wrong”; Wickens et al., 2015). Low DOA participants experienced only an automation gone failure, whereas high DOA participants experienced either an automation gone or automation wrong failure. Participant detection of an automation wrong failure resulted in the automation being removed and manual control resumed.

Under routine states (when automation was reliably functioning), Tatasciore et al. (2020) found that high DOA benefited classification and CPA performance and reduced workload, compared to low DOA and no automation. However, during routine states, both low and high DOA impaired nonautomated dive task performance compared to no automation. Only low DOA impaired SA compared to no automation. Following an automation gone failure, low DOA participants experienced increased workload, but return-to-manual performance for both low and high DOA participants was comparable to no automation. High DOA participants who experienced the automation wrong failure took ~3 min to detect it. When the automation wrong failure was detected and high DOA removed, return-to-manual performance was comparable to no automation. There was also no difference in classification performance immediately after the automation wrong failure for high DOA compared to no automation. It was concluded that high DOA produced superior benefits compared to low DOA, at no extra cost.

While these findings may initially seem inconsistent with the lumberjack effect, it is notable that the majority of studies in the Onnasch et al. (2014) meta-analysis used relatively fast evolving tasks such as air traffic control, unmanned vehicle control, and driving. In contrast, emulating submarine control room operational settings (Kirschenbaum, 2011; Roberts et al., 2017), contacts and their tracks in the track management task move very slowly on operators’ displays. Thus, while high DOA may have reduced the extent to which participants monitored raw information (reflected by poorer dive task performance), the slow pace of the task might have allowed participants to retain SA and the ability to return-to-manual performance.

If this conjecture is accurate, it raises the question of whether it would be better to provide action implementation automation (full DOA), rather than just action recommendation automation (high DOA), in slowly evolving task environments such as submarine track management. Under routine states, full DOA would yield perfect performance on automated tasks and should reduce workload more than high DOA. If this could be achieved without further costs to nonautomated task performance, SA, or greater impairment when automation fails compared to high DOA, full DOA has the potential to provide a “free lunch” (Wickens, 2018)—that is, superior benefits when compared to high DOA, without extra costs.

To determine whether this is possible, the current study examined the effects of high and full DOA on automated task performance, workload, nonautomated task performance, SA, automation failure detection, and return-to-manual performance. High DOA replicated Tatasciore et al. (2020), while full DOA provided the same action recommendations, but also implemented these actions for the classification and CPA tasks. Our aims were to examine whether: (a) full DOA reduces workload and increases costs to nonautomated task performance, SA, or return-to-manual performance compared to high DOA; (b) full DOA reduces the speed or accuracy of detecting automation wrong failures compared to high DOA; (c) full DOA impairs classification performance after an automation failure compared to high DOA and no automation; and (d) the extent to which the impact of high DOA reported by Tatasciore et al. (2020), when compared to no automation, could be replicated.

PREDICTIONS

Predictions are outlined in Table 1 and detailed below.

Automated Task Performance

Compared to no automation, high DOA should yield benefits to classification and CPA

TABLE 1: Predictions Regarding the Effects of DOA as a Function of Automation State

Task	Performance During Automation Working (Routine State)	Performance Immediately After Automation Failure	Performance After Automation Failure Detected (Removal State)
Classification			
Accuracy	None < High < Full* (the higher the DOA, the better the accuracy)	[None = High] > Full (poorer accuracy immediately after failure with full DOA)	[None = High] > Full (lower accuracy after full DOA removal)
RT	None > High > Full* (the higher the DOA, the faster the decisions)	[None = High] < Full (slower RT immediately after failure with full DOA)	[None = High) < Full (slower decisions after full DOA removal)
CPA			
Accuracy	None < High < Full* (the higher the DOA, the better the accuracy)		[None = High] > Full (lower accuracy after full DOA removal)
RT	None > High > Full* (the higher the DOA, the faster the decisions)		[None = High] < Full (slower decisions after full DOA removal)
Dive			
Accuracy	[None = Full] > High (poorer accuracy with high DOA)		None = High = Full (no RTM effects)
RT	None = High = Full (no difference in RT)		None = High = Full (no RTM effects)
Workload	None > High > Full (the higher the DOA, the lower the workload)		[None = High] < Full (higher workload after full DOA removal)
SA	[None = High] > Full (poorer SA with full DOA)		[None = High] > Full (poorer SA after full DOA removal)

Note. CPA = closest point of approach; DOA = degree of automation; Full = full DOA; High = high DOA; RT = response time; RTM = return-to-manual; Routine = reliable automation; Removal = point after the automation failure is detected by participant and the automation subsequently removed; None = no automation; SA = situation awareness.
*Note that by definition, classification and CPA will be perfect with the use of full DOA when automation is reliable during routine states.

accuracy and response times (RTs) similar to those reported by Tatasciore et al. (2020). Following automation removal (i.e., after the automation failure is detected and automation removed), we expected no return-to-manual costs to the classification or CPA tasks for high DOA based on Tatasciore et al. (2020). However, as per the lumberjack effect, we expected return-to-manual costs to the classification and CPA tasks for full DOA, given that we are further crossing the lumberjack critical boundary (Onnasch et al., 2014).

Nonautomated Task Performance

Replicating Tatasciore et al. (2020), we expected a cost to nonautomated dive task accuracy during routine states for high DOA compared to no automation. However, we expected no difference in dive task performance between full DOA and no automation, and better dive task accuracy for full DOA than high DOA, because the dive task is the sole manual task when full DOA is available. Furthermore, based on Tatasciore et al. (2020), we did not expect return-to-manual costs following high or full DOA removal.

Workload

On the basis of Tatasciore et al. (2020), we expected reduced workload with increased DOA and no return-to-manual costs to workload when high DOA was removed. However, given that we are further crossing the lumberjack critical boundary (Onnasch et al., 2014), we expected workload to be higher when full DOA was removed compared to no automation, and when high DOA was removed.

Situation Awareness

Replicating Tatasciore et al. (2020), we expected no cost to SA for high DOA compared to no automation before or after automation removal. Based on Onnasch et al. (2014), we expected SA to be lower for full DOA compared to no automation and high DOA both before and after automation removal.

Automation Failure Detection and Performance Immediately After the Failure

Tatasciore et al. (2020) examined performance on the first three classification events immediately after both the automation gone and automation wrong failures and found no classification deficits on these events for high DOA compared to no automation. However, based on Onnasch et al. (2014), we expected the first three classification decisions to be less accurate after the failure for full DOA compared to high DOA and no automation. We also expected the automation failure to be detected less often or more slowly for participants using full DOA compared to high DOA due to increased workload and reduced SA.

METHOD

Participants

One hundred and twenty-three (70 women, 53 men) psychology students (age: M = 21.15 years, SD = 5.18) from The University of Western Australia (UWA) volunteered for course credit and provided informed consent. Participants were randomly assigned to no automation (N = 40), high DOA (N = 40), or full DOA (N = 43) conditions. Research complied with the American Psychological Association Code of Ethics and was approved by the UWA Human Research Ethics Office.

Design

A mixed design was used, with the between-subjects factor being automation condition (no automation, high DOA, full DOA) and the within-subjects factor being automation state (routine, automation removal).

Simulated Submarine Track Management Task

The simulation comprised of the surface plot and waterfall displays (Figure 1). The surface plot presented a top-down view of the area with concentric rings representing distance from Ownship, and contact location and heading information. The waterfall display comprised a horizontal axis presenting contact bearings. Vertical lines (soundtracks) were presented on the waterfall display and grew downwards to show contact bearings and changes in bearings over time. At the bottom of the surface plot was the “Track Assist” automation interface, which indicated whether the automation was activated.

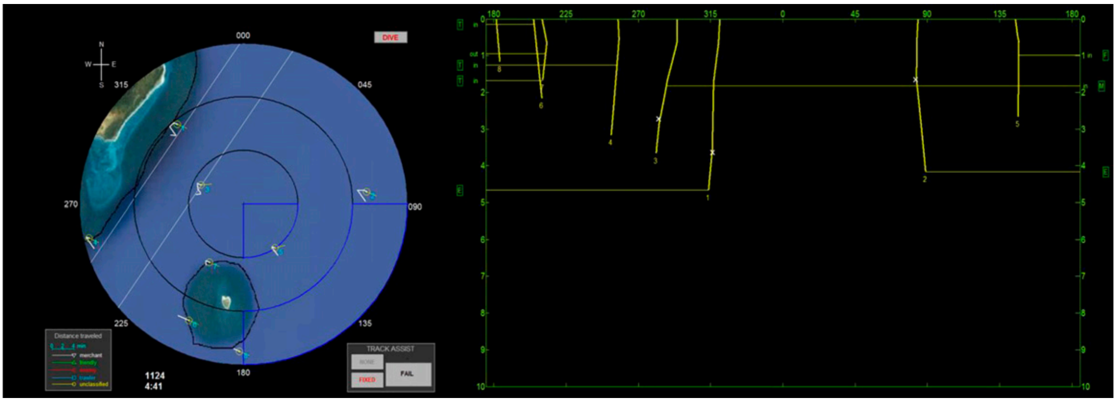

Figure 1. An example scenario with high DOA active. The left display is the surface plot (bird's-eye view of the area), and the right display is the waterfall. Presented on the waterfall display are soundtracks, which signify bearing changes over time. Up to eight contacts are displayed. Projecting from the center of each contact on the surface plot is a line that indicates the current heading of the contact. Also attached to each contact on the surface plot is a track history to represent contact heading changes (CPA task). Presented on the waterfall display are horizontal lines that are automatically placed to signify when a contact has entered an area of interest (classification task). Attached to these lines are boxes with the appropriate classification letter inside. The full DOA display looks identical, except the automation implements all classification and CPA task actions. When no automation is provided, there are no track history lines, and horizontal lines are placed manually. *Note*. CPA = closest point of approach; DOA = degree of automation.

It also included a "FAIL" button that participants were instructed to click if they believed the automation had failed. If clicked after the automation failure, the message "Automation failure detected. Track Assist turned off. Manual tracking required" was presented. If clicked when automation was functioning reliably, it read "Automation has not failed." Task-load varied cyclically during each scenario, starting from a minimum of one, increasing to a maximum of eight, and then decreasing again to minimum three times during each scenario.

Classification task. Contacts were classified according to how long they spent within specific regions on the surface plot. A contact was a "Merchant" (located within the two white parallel lines), "Friendly" (located within the area bounded by blue lines), or "Trawler" (located within the shallow dark blue area) if it spent two continuous minutes within these specified regions on the surface plot (Figure 1). A contact was an "Enemy" if it had not spent at least one continuous minute in any classification-relevant region during the first 4 min of its presentation. To assist with tracking how long contacts spent within regions, participants could manually place a horizontal line at the top of each soundtrack on the waterfall display when a contact entered a region. The contact was classifiable when this line reached the 2-min mark. To detect enemies, participants could place a horizontal line on the bottom of the soundtrack of contacts that had not entered a classification-relevant region. When this line reached 4 min, the contact could be classified as an enemy.

When high DOA was available, horizontal lines were automatically placed on the soundtracks when contacts entered classification-relevant regions. To assist with classifying enemies, a horizontal line was automatically placed at the bottom of the soundtrack when it reached the 4-min mark. In addition, a square box with a letter signifying the recommended classification was attached to each horizontal line (f = Friendly, m = Merchant, t = Trawler, e = Enemy). The horizontal line flashed to notify participants when a contact could be classified, and participants were responsible for executing the classification task manually. Full DOA was identical to high DOA, except immediately after the horizontal line flashed, the automation implemented the classification task action. After the automation (full DOA) or participant (high DOA) implemented the classification, the contact soundtrack and the contact icon on the surface plot changed color (Friendly = green, Merchant = white, Trawler = blue, Enemy = red).

When automation failed, incorrect classification advice was provided for subsequently presented contacts. Specifically, the horizontal lines were placed either 30 s too early or late on the soundtracks, and the recommended classification was incorrect (e.g., merchants assigned classification letter f, t, or e). Additionally, for full DOA, the automation implemented the incorrect classification and the subsequent change in contact soundtrack and icon color was incorrect. The type of misclassification (i.e., incorrect recommended or implemented classification, and incorrect placement of horizontal timelines—either 30 s too early or late) assigned to each contact presented after the automation failure was random.

CPA. A CPA occurred when a contact that was heading toward Ownship subsequently turned away. Each contact had one CPA. Participants reported CPAs by marking a cross on the contact soundtrack on the waterfall display. With high DOA, each contact had a track history marked on the surface plot, minimizing the need for participants to track heading changes. The track history also flashed to notify participants when a CPA occurred and continued flashing until participants marked the CPA. Full DOA was identical to high DOA, except the automation immediately marked the CPA.

Dive task. The dive task was never automated. Participants were required to click the dive button on the surface plot when all contacts on the surface plot were heading in the same direction, and one contact was heading directly toward Ownship. Each scenario contained 9–10 dive windows of variable duration (10–30 s).

Measures

Situation awareness. SA was measured using the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995). The simulation was paused, and both displays blanked and replaced with seven SAGAT queries, six times during each scenario. The first query required participants to mark one of the contacts location on the surface plot. The next six queries targeted underlying information necessary for performing classification, CPA, and dive tasks. During a given SAGAT pause, all conditions received the same seven queries. Queries were taken from a pool of queries (Table 2).

Workload. The Air Traffic Workload Input Technique (ATWIT; Stein, 1985) was presented on the surface plot once every minute. Participants had 10 s to rate their workload from 1 to 10 (1–2 = very low, 3–5 = moderate, 6–8 = relatively high, 9–10 = very high). The National Aeronautics and Space Administration Task Load Index (NASA-TLX; Hart & Staveland, 1987) was completed after each scenario.

Procedure

The experiment took 3 hr. First, participants completed 80 min of training (an audio-visual PowerPoint presentation explaining the tasks and measures, a 10-min narrated video showing the simulation with no automation, and a 27.5-min practice scenario with no automation). Following this, participants in the high and full DOA conditions watched a training presentation explaining automation. They were notified that although the automation was highly reliable, it may not be perfect, and were instructed to report any automation failures by clicking the fail button. Participants then completed three 27.5-min scenarios. Each scenario contained different maps and unique contacts. Map order was counterbalanced. For high and full DOA, the automation unexpectedly provided incorrect advice during the last scenario (10.38, 10.48, or 10.88 min into the scenario). The automation continued to provide incorrect advice until the failure was reported. Once a failure was reported, the automation was removed, and manual performance on the classification and CPA tasks resumed.

RESULTS

The hit rates for each task were calculated as the number of correct task responses per scenario divided by the total number of task events. RTs

TABLE 2: SAGAT Queries Used to Measure Participant SA

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

Note. CPA = closest point of approach; SA = situation awareness; SAGAT = Situation Awareness Global Assessment Technique.
Retrieved from "The benefits and costs of low and high degree automation", Tatasciore et al. (2020).

were based on correct responses. For the CPA task, a response was correct if the cross was marked on the correct soundtrack at any time 1.5 s before or after the actual CPA event. If the cross was marked outside of this range, it was recorded as a false alarm. The exact number of contacts and events related to making a CPA false alarm was indeterminable. However, we reasoned that because a CPA false alarm was most likely to occur following contact course changes, a reasonable estimate of the false alarm rate was the number of false alarms divided by the number of contact course changes, minus the total number of CPA events (see Chen et al., 2017; Tatasciore et al., 2020). CPA accuracy was calculated by subtracting the CPA false alarm rate from the hit rate. For the dive task, a false alarm was most likely during a contact course change, as contact course changes were a prerequisite for a dive window to be initiated. However, given that there were fewer dive windows than CPA events, and the rule that all contacts need to be heading in the same direction, it was unlikely that every course change would be mistaken as a dive window. Therefore, we calculated the dive false alarm rate as the number of false alarms divided by half the number of contact course changes, minus the total number of dive windows (Chen et al., 2017; Tatasciore et al., 2020). Dive accuracy was calculated by subtracting the dive false alarm rate from the hit rate.

Table 3 presents means and 95% confidence intervals for performance, workload, and SA, separated into routine state (first two scenarios and one third of last scenario when automation functioned as expected) and automation removal state (after automation failure was detected by participants and disengaged).

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

Automation	Classification		CPA		Dive		SAGAT Accuracy	Workload	
	Hit	RT	Hit-FA	RT	Hit-FA	RT		ATWIT	NASA-TLX
					Routine state				
None	.75 [.68, .82]	30.09 [26.84, 33.33]	.35 [.28, .42]	17.01 [12.74, 21.27]	.78 [.70, .87]	8.86 [7.96, 9.76]	.58 [.54, .62]	5.29 [4.94, 5.64]	62.31 [58.10, 66.52]
High	.93 [.88, .97]	19.23 [16.96, 21.51]	.80 [.70, .89]	11.82 [10.19, 13.44]	.60 [.53, .68]	10.62 [9.47, 11.77]	.52 [.48, .55]	4.13 [3.72, 4.54]	56.37 [51.59, 61.15]
Full					.68 [.62, .75]	9.99 [8.85, 11.14]	.56 [.51, .60]	3.88 [3.49, 4.27]	51.67 [46.33, 57.02]
					Automation removal state				
None	.75 [.68, .83]	28.84 [24.56, 33.11]	.32 [.24, .40]	19.76 [11.37, 28.16]	.83 [.77, .89]	8.00 [7.15, 8.85]	.55 [.50, .60]	5.28 [4.76, 5.79]	61.15 [56.20, 66.10]
High	.74 [.66, 82]	34.41 [28.51, 40.32]	.34 [.25, .43]	13.50 [10.07, 16.94]	.62 [.52, .72]	11.22 [9.04, 13.40]	.55 [.50, .59]	5.59 [5.10, 6.08]	60.14 [54.36, 65.93]
Full	.85 [.81, .89]	34.09 [29.80, 38.38]	.38 [.30, .47]	15.96 [12.42, 19.49]	.76 [.65, .87]	10.10 [7.52, 12.69]	.53 [.47, .58]	5.27 [4.71, 5.84]	58.43 [52.73, 64.14]

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

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

Dependent Variable	Effect	*F*	*df*	*p*	η_p^2
Dive (Hit-FA)	Condition	7.25	(1, 107)	.001*	.12
	State	3.44	(1, 107)	.07	.03
	Condition × State	.34	(1, 107)	.71	.01
Dive (RT)	Condition	5.29	(1,102)	.01*	.09
	State	.00	(1, 102)	.99	<.01
	Condition × State	.60	(1, 102)	.55	.01
NASA-TLX	Condition	2.58	(1, 108)	.08	.05
	State	12.84	(1, 108)	<.001*	.11
	Condition × State	7.38	(1, 108)	<.001*	.12
ATWIT	Condition	3.02	(1, 106)	.05*	.05
	State	67.22	(1, 106)	<.001*	.39
	Condition × State	18.08	(1, 106)	<.001*	.25
SAGAT (Accuracy)	Condition	.81	(1, 104)	.45	.02
	State	2.22	(1, 104)	.14	.01
	Condition × State	3.31	(1, 104)	.04*	.02

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

To test our predictions for the automated tasks (classification, CPA), we conducted separate analyses for routine and automation removal states. The full DOA condition was not included in routine state analyses because the tasks were performed by the automation and were 100% accurate (zero variance). We compared classification and CPA performance for high DOA and no automation during routine states by conducting *t*-tests, and for full DOA, high DOA, and no automation during the automation removal state using one-way analyses of variance (ANOVAs). Significant one-way ANOVAs were followed with post-hoc *t*-tests comparing the three conditions to each other, corrected for family-wise error by reporting Bonferroni *p* values (*p* values multiplied by three, the number of comparisons for each dependent variable).

To test our predictions for nonautomated dive task performance, workload, and SA, we ran Automation Condition (no automation, high DOA, full DOA) × Automation State (routine, automation removal) mixed ANOVAs (Table 4). Significant main effects and interactions were followed up as described above.

To test our predictions regarding the association between automation condition and automation failure detection accuracy and RT, we ran a χ^2 test and a nonparametric Mann–Whitney U test, respectively. Furthermore, to test our predictions regarding classification performance (accuracy and RT) immediately after the automation failure, we ran Automation Condition (no automation, high DOA, full DOA) × Classification Event (first, second, third event after failure) mixed ANOVAs.

Estimates of Cohen's *d* indicate we had a power of 0.82 to detect the medium-to-large effect sizes previously reported by Cohen (1988); Tatasciore et al. (2020).

Automated Task Performance

Classification. During routine states, high DOA participants made significantly more accurate, $t(78) = 4.42$, $p < .001$, $d = .99$, and faster,

$t(78) = 5.54$, $p < .001$, $d = 1.24$, classifications than no automation participants. For automation removal states, there was no difference in accuracy, $F(2,107) = 2.81$, $p = .07$, $\eta_p^2 = .05$, or RTs, $F(2,107) = 1.75$, $p = .18$, $\eta_p^2 = .03$.

In summary, high DOA benefited classification accuracy and RT during routine states, and there were no return-to-manual costs for either high or full DOA.

CPA. During routine states, high DOA participants made more accurate, $t(78) = 7.56$, $p < .001$, $d = 1.69$, and faster, $t(78) = 2.30$, $p = .02$, $d = .52$, CPA decisions than no automation participants. For automation removal states, there was no difference in accuracy, $F < 1$, or RTs, $F(2,102) = 1.15$, $p = .32$, $\eta_p^2 = .02$.

In summary, high DOA benefited CPA accuracy and RT during routine states, and there were no return-to-manual costs for either high or full DOA.

Nonautomated Task Performance

Dive task. Following up the significant main effect of Condition (Table 4) showed that during routine states, high DOA participants made less accurate dive decisions than no automation participants, $t(78) = 3.25$, $p = .003$, $d = .73$. No other comparisons were significant ($p > .17$, $d < .46$). For automation removal states, following up the significant main effects of Condition showed that high DOA participants made significantly slower, $t(73) = 2.93$, $p = .01$, $d = .68$, and less accurate, $t(75) = 3.60$, $p = .003$, $d = .82$, dive decisions than no automation participants. No other accuracy or RT comparisons were significant ($t < 1.95$, $p > .16$).

In summary, high DOA lead to poorer dive task performance during both routine and automation removal states. However, there were no comparable decrements to dive task performance for the full DOA condition during either routine or automation removal states.

Workload

Following up on the significant Condition × State interactions for ATWIT and NASA-TLX showed that during routine states, there was a significant difference between the automation conditions, $F(2,120) = 15.50$, $p < .001$, $\eta_p^2 = .21$. High DOA participants, $t(78) = 4.33$, $p < .001$, $d = .97$, and full DOA participants, $t(81) = 5.41$, $p = .003$, $d = 1.19$, reported lower workload on the ATWIT than no automation participants, but there was no difference between full and high DOA, $t < 1$. For the NASA-TLX, there was also a significant difference between the automation conditions, $F(2,120) = 5.01$, $p = .01$, $\eta_p^2 = .08$. Full DOA participants reported lower workload than no automation participants, $t(81) = 3.13$, $p = .01$, $d = .69$. No other comparisons were significant, ($t < 1.90$, $p > .17$). For automation removal states, there was no difference between conditions for either the ATWIT or NASA-TLX, $F < 1$.

In summary, full DOA reduced workload during routine states compared to no automation as measured by both ATWIT and NASA-TLX, while high DOA reduced workload as measured by ATWIT but not the NASA-TLX. Neither DOA condition showed evidence for return-to-manual workload increases after automation removal.

Situation Awareness

Following up on the significant Condition × State interaction for SAGAT showed that during routine states, the difference between the conditions approached significance, $F(2,120) = 2.84$, $p = .06$, $\eta_p^2 = .05$. For automation removal states, there was no significant difference between conditions, $F < 1$.

In summary, there were no significant costs to SA with the use of high or full DOA during routine or automation removal states.

Automation Failure Detection and Performance Immediately After the Failure

The association between the type of automation and the successful detection of the automation wrong failure was significant, $\chi^2 = 5.58$, $p = .02$ (Full DOA: 76.67% vs. High DOA: 95.0%). However, there was no difference in the time taken to correctly detect the automation wrong failure between the full DOA ($M = 237.79$s; CI [153.70, 321.88]) and high DOA ($M = 192.91$s; CI [118.30, 267.52]) conditions, $U = 540.00$, $z = 1.00$, $p = .32$, $r = .12$.

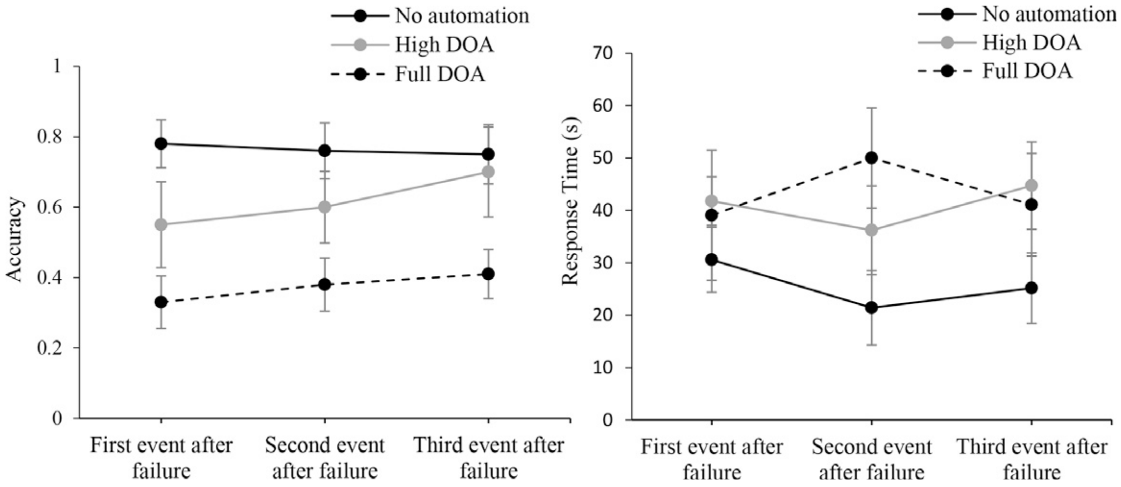

Figure 2. Classification accuracy (left graph) and RT (right graph) for the first thre events after the automation failure, as a function of condition. Error bars represent 95% within-subject confidence intervals. *Note*. DOA = degree of automation; RT = response time.

We analyzed performance on the three classification events immediately following the automation failure (Figure 2). For classification accuracy, there was a main effect of Condition, $F(2,119) = 12.25, p < .001, \eta_p^2 = .17$, but no main effect of Classification Event and no interaction ($p > .36, \eta_p^2 < .02$). Follow-up tests indicated that full DOA participants were less accurate than no automation participants, $t(80) = 4.43$, $p < .001$, $d = 1.00$. No other comparisons were significant ($t < 2.17, p > .08$). For classification RTs, there were no significant main effects or interactions, ($p > .30, \eta_p^2 < .06$).

We also re-ran these analyses including only the subset of participants in the high ($N = 24$) and full DOA ($N = 19$) conditions that detected the automation failure prior to the first post-failure classification event. For classification accuracy and RT, there were no main effects of Condition, Classification Event, and no interaction ($p > .11, \eta_p^2 < .11$). Thus, there was no significant difference in performance on the three classification events immediately after the failure for participants that had detected the automation failure in the full DOA condition compared to the no automation condition.

In summary, classification accuracy immediately after the automation failure was poorer for the full DOA compared to no automation condition. However, poorer performance was largely driven by participants who had not yet detected the automation failure.

DISCUSSION

Meta-analytic evidence suggests that increased DOA improves operator performance and reduces workload, but costs SA and return-to-manual performance, particularly when DOA crosses the critical boundary to action recommendation (Onnasch et al., 2014). In submarine control rooms, representations of contacts and their tracks evolve very slowly on displays (Kirschenbaum, 2011; Roberts et al., 2017). Given our past research outcomes (Chen et al., 2017; Tatasciore et al., 2020) and that the majority of studies included in the Onnasch et al. (2014) meta-analysis used relatively fast evolving tasks, we wanted to explore whether we could achieve what Wickens (2018) referred to as a “free lunch.” That is, the benefits of action implementation automation without costs. To test this, we examined the effects of high (action recommendation) and full (action implementation) DOA on automated task performance, subjective workload, nonautomated task performance, SA, and return-to-manual performance in a submarine track management task. We also examined whether DOA influenced the ability to detect automation wrong failures, and whether there were differences in classification performance immediately after automation failures across the full DOA, high DOA, and no automation conditions (Table 5 for a summary of findings).

As predicted, full DOA lowered participant workload compared to no automation. Additionally, full DOA did not cost dive task performance or SA compared to high DOA or no automation. That said, although there was no cost to dive task performance, it remains possible that full DOA participants were just as complacent as those with high DOA (Parasuraman & Manzey, 2010), but were possibly able to overcome this using the additional cognitive resources freed up by the full DOA. In contrast to the findings of the "lumberjack effect" reported by Onnasch et al. (2014), participants provided full DOA were able to return-to-manual performance after knowing that automation was removed. We suggest that the slow nature of the submarine track management task may have allowed participants in the full DOA condition time to recover following automation removal.

In terms of automation failure, participants were less likely to detect automation wrong failures when using full DOA compared to high DOA. Additionally, we found that classification accuracy immediately after the automation failure was poorer in the full DOA condition compared to the no automation condition (post-hoc analyses indicated that this was largely driven by participants who had not yet detected the automation failure). This decrement to contact classification may indicate that participants using full DOA may not have been monitoring the classification and CPA tasks as frequently or closely as those with high DOA, increasing the probability that they missed the automation contact classification failure because their attention was focused elsewhere (possibly focused on the dive task as discussed above).

Finally, comparisons between high DOA and no automation largely replicated Tatasciore et al. (2020), thereby confirming the robust benefits of high DOA in track management (Jones et al., 2010; Pashler & Wagenmakers, 2012). We found that high DOA benefited classification and CPA accuracy and RT, and lowered workload compared to no automation. There were also no return-to-manual performance costs for the classification and CPA tasks, or to workload following removal of high DOA. However, as in Tatasciore et al. (2020), during routine states high DOA did degrade dive task performance, and in this current study this cost also continued after automation removal.

Practical Implications, Limitations, and Conclusions

The current data indicate that action implementation automation could be used effectively to benefit task performance and reduce workload, without costs to nonautomated dive task performance, SA, or return-to-manual performance after automation failure detection. However, participants provided action implementation automation were significantly less likely to detect automation failures compared to those provided action recommendation automation. Almost 25% of action implementation participants did not detect automation failures even after 17 min. The consequence of this was that, until the automation failure was detected, classification performance was poorer for those using action implementation automation compared to no automation. In complex work environments, if automation fails, the operator's ability to promptly detect the failure is critical. This is particularly true in faster evolving task environments, or contexts with more time pressure to make a manual decision (e.g., air traffic control, unmanned vehicle control), where delays in noticing automation failures could be catastrophic (e.g., collision between the USS McCain and Tanker Alnic MC; National Transportation Safety Board, 2017). However, there are slowly evolving work domains, such as process control and nuclear power plants, in which operational parameters neglected due to undetected automation failures could quickly become irretrievably problematic or economically costly (e.g., restarting power plants; Muir & Moray, 1996).

TABLE 5: Summary of Findings Regarding the Effects of DOA as a Function of Automation State

Task	Performance During Automation Working (Routine State)	Matches Prediction	Performance Immediately After Automation Failure	Matches Prediction	Performance After Automation Failure Detected (Removal State)	Matches Prediction
Classification						
Accuracy	None < High < Full* (the higher the DOA, the better the accuracy)	Yes	[None = High] > Full (poorer accuracy immediately after failure with full DOA)	Yes	None = High = Full (no RTM effects)	Partial
RT	None > High > Full* (the higher the DOA, the faster the decisions)	Yes	None = High = Full (no difference in RT immediately after failure)	Partial	None = High = Full (no RTM effects)	Partial
CPA						
Accuracy	None < High < Full* (the higher the DOA, the better the accuracy)	Yes			None = High = Full (no RTM effects)	Partial
RT	None > High > Full* (the higher the DOA, the faster the decisions)	Yes			None = High = Full (no RTM effects)	Partial
Dive						
Accuracy	[None = Full] > High (poorer accuracy with high DOA compared to no automation)	Yes			[None = Full] > High (poorer accuracy after high DOA removal compared to no automation)	Partial
RT	None = High = Full (no difference in RT)	Yes			[None = Full] < High (slower decisions after high DOA removal compared to no automation)	Partial
Workload						
ATWIT	None > [High = Full] (reduced workload with high or full DOA compared to no automation)	Partial			None = High = Full (no RTM effects)	Partial

(Continued)

TABLE 5 (Continued)

Task	Performance During Automation Working (Routine State)	Matches Prediction	Performance Immediately After Automation Failure	Matches Prediction	Performance After Automation Failure Detected (Removal State)	Matches Prediction
NASA-TLX	[None = High] > Full] (reduced workload with full DOA)	Partial			None = High = Full (no RTM effects)	Partial
	None = High = Full (no difference in SA)	Partial			None = High = Full (no RTM effects)	Partial

Note. CPA = closest point of approach; DOA = degree of automation; Full = full DOA; High = high DOA; None = no automation; Removal = point after the automation failure is detected by participant and the automation subsequently removed; Routine = reliable automation; RT = response time; RTM = return-to-manual; SA =situation awareness.
Grey Shading = observed result matches predicted result.
*Note that by definition classification and CPA performance was perfect with the use of full DOA when automation was reliable during routine states.

The simulated submarine track management task used in the current experiment was designed in consultation with Royal Australian Navy Submariners. As such, the current experiment has external validity as it is broadly representative of work environments that require monitoring of demanding perceptual displays. Thus, although we concentrated on track management, our findings are relevant to other work contexts, particularly those involving slowly evolving situations that require monitoring of multiple demanding displays (e.g., maritime surveillance). That said, there are potential issues in generalizing from novice participants to experts as there are undeniable differences in their experience, motivation, and cognition (for discussion see Jamieson & Skraaning, 2020). Furthermore, we have not directly tested whether discrepancies between the current findings and those outcomes predicted by the Onnasch et al. (2014) meta-analysis are due to the slow nature of the submarine track management task, or some other factor(s). Finally, we did not introduce participants to automation transitions during the practice scenario as we did not want participants to expect that automation would necessarily fail (first-failure effect; see Merlo et al., 2000). It may be the case that the immediate return-to-manual deficit observed when action implementation automation failed would be reduced if participants had practice with automation-manual transitions (see Zhang et al., 2019).

In conclusion, in our experiment, action implementation automation did not produce significant costs to nonautomated task performance, SA, or return-to-manual performance, compared to action recommendation automation or no automation. However, participants provided action implementation automation were less likely to notice automation failures compared to those provided action recommendation automation, and until they did so, were more likely to make inaccurate decisions compared to those provided no automation. This suggests that if automation failure detection could be improved, action implementation automation has the potential to provide the proverbial “free lunch” (Wickens, 2018) in complex task environments.

ACKNOWLEDGMENTS

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

KEY POINTS

- Theory and meta-analytic evidence suggest that with increasing degrees of automation, operator performance improves and workload decreases, but situation awareness and return-to-manual performance can decline.
- In a slowly evolving simulated submarine track management task, relative to no automation, action recommendation automation benefited automated task performance and workload, but with costs to nonautomated task performance.
- Action implementation automation improved automated task performance and lowered workload, with no costs to nonautomated task performance, situation awareness, or return-to-manual performance compared to no automation.
- Participants provided action implementation automation were poorer at detecting automation failures compared to participants provided action recommendation automation, and made less accurate decisions during the period immediately after the automation failure compared to participants provided no automation.
- Action implementation automation may be effective for some task contexts, but system designers should be aware that operators may be less likely to detect automation failures and that performance may suffer until such failures are detected.

ORCID iDs

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

Vanessa K. Bowden https://orcid.org/0000-0001-5553-3400

Shayne Loft https://orcid.org/0000-0002-5434-0348

REFERENCES

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. https://doi.org/10.1037/xap0000126

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Lawrence Erlbaum.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*, 97–101. https://doi.org/10.1177/154193128803200221

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*, 65–84. https://doi.org/10.1518/001872095779049499

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). Elsevier Science Pub. Co.

Jamieson, G. A., & Skraaning, G. (2020). The absence of degree of automation trade-offs in complex work settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*, 516–529. https://doi.org/10.1177/0018720819842709

Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*, 586–595. https://doi.org/10.1177/0018720810384394

Kirschenbaum, S. S. (2011). Expertise in the submarine domain: The impact of explicit display on the interpretation of uncertainty. In K. L. Mossier & U. M. Fischer (Eds.), *Infomed by knowledge: Expert performance in complex situations* (pp. 189–199). Psychology Press.

Merlo, J. L., Wickens, C. D., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. *Proceedings of the 4th Annual Army Federated Laboratory Symposium*, 27–31.

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

National Transportation Safety Board. (2017). *Collision between US Navy destroyer John S McCain and tanker Alnic MC Singapore strait, 5 miles northeast of Horsburgh lighthouse*. https://www.marinha.mil.br/dpc/sites/www.marinha.mil.br.dpc/files/MAR1901.pdf

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. 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: The Journal of the Human Factors and Ergonomics Society*, *52*, 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, *30*, 286–297. https://doi.org/10.1109/3468.844354

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. https://doi.org/10.1177/1745691612465253

Roberts, A. P. J., Stanton, N. A., & Fay, D. (2017). Land ahoy! Understanding submarine command and control during the completion of inshore operations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 1263–1288. https://doi.org/10.1177/0018720817731678

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators (Technical Report)*. Man Machine Systems Laboratory, MIT.

Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe*. FAA.

Tatasciore, M., Bowden, V. K., Visser, T. A. W., Michailovs, S. I. C., & Loft, S. (2020). The benefits and costs of low and high degree of automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*, 874–896. https://doi.org/10.1177/0018720819867181

Wickens, C. (2018). Automation stages & levels, 20 years after. *Journal of Cognitive Engineering and Decision Making*, *12*, 35–41. https://doi.org/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: The Journal of the Human Factors and Ergonomics Society*, *57*, 728–739. https://doi.org/10.1177/0018720815581940

Zhang, B., de Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 285–307. https://doi.org/10.1016/j.trf.2019.04.020

Monica Tatasciore is a PhD and Master's student enrolled in the Doctor of Philosophy and Master of Industrial and Organisational Psychology programs 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

Shayne Loft is a professor at The University of Western Australia. He received his PhD in 2004 from the University of Queensland.

Date received: July 6, 2020
Date accepted: December 16, 2020