Journal of Experimental Psychology: Applied

1076-898X/17/$12.00 http://dx.doi.org/10.1037/xap0000126

Optimizing the Balance Between Task Automation and Human Manual Control in Simulated Submarine Track Management

Stephanie I. Chen and Troy A. W. Visser
The University of Western Australia

Samuel Huf
Defence Science and Technology Group, Perth, Western Australia

Shayne Loft
The University of Western Australia

Automation can improve operator performance and reduce workload, but can also degrade operator situation awareness (SA) and the ability to regain manual control. In 3 experiments, we examined the extent to which automation could be designed to benefit performance while ensuring that individuals maintained SA and could regain manual control. Participants completed a simulated submarine track management task under varying task load. The automation was designed to facilitate information acquisition and analysis, but did not make task decisions. Relative to a condition with no automation, the continuous use of automation improved performance and reduced subjective workload, but degraded SA. Automation that was engaged and disengaged by participants as required (adaptable automation) moderately improved performance and reduced workload relative to no automation, but degraded SA. Automation engaged and disengaged based on task load (adaptive automation) provided no benefit to performance or workload, and degraded SA relative to no automation. Automation never led to significant return-to-manual deficits. However, all types of automation led to degraded performance on a nonautomated task that shared information processing requirements with automated tasks. Given these outcomes, further research is urgently required to establish how to design automation to maximize performance while keeping operators cognitively engaged.

Keywords: adaptable automation, adaptive automation, submarine track management, situation awareness, workload

The term *automation* has been defined as a system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out by a human operator (Parasuraman, Sheridan, & Wickens, 2000). Automation that supports our workplaces is designed to alleviate the requirement for humans to control tasks, to increase system capacity. Examples include flight management systems in cockpits, robotics in manufacturing and undersea exploration, military command and control automation, radar plotting tools in seaboard navigation, image-guided navigation tools in medical surgery, and separation assurance tools in air traffic control. Technological advancement and its potential economic benefits mean that there is a continuing trend toward requiring humans to deal with larger amounts of information in more complex work environments using increasingly capable, highly automated systems (Bindewald, Miller, & Peterson, 2014; Sheridan, 2015).

Researchers have long recognized the potential safety issues associated with automation (Parasuraman, Molloy, & Singh, 1993; Rasmussen & Rouse, 1981), particularly under conditions where it is continually used by operators (static automation/function allocation). To the extent that static automation is reliable and trusted (Wickens & Dixon, 2007), the automation of a task will exceed human manual performance and reduce operator workload, particularly as the degree of automation (DOA) moves across a critical boundary from "acquiring and analyzing" information to "recommending task actions" (Onnasch, Wickens, Li, & Manzey, 2014). However, part of the reason static automation reduces workload is that operators process less of the raw information related to the task being automated (automation-induced complacency; Parasuraman & Manzey, 2010; Wickens, Sebok, Li, Sarter, & Gacy, 2015). Automation-induced complacency can compromise the operator's understanding of their task environment and consequently their situation awareness (SA; Endsley, 1988). A loss of SA is problematic when automation is highly, but not perfectly, reliable because this creates a need for infrequent and unpredictable operator intervention. Research indicates it can be challenging under these conditions for operators to regain manual control of previously automated tasks (Manzey, Reichenbach, & Onnasch, 2012;

Stephanie I. Chen and Troy A. W. Visser, School of Psychological Science, The University of Western Australia; Samuel Huf, Defence Science and Technology Group, Perth, Western Australia; Shayne Loft, School of Psychological Science, The University of Western Australia.

This research was supported by Discovery Grant DP160100575 from the Australian Research Council awarded to Shayne Loft. We thank Janelle Braithwaite for her programming of the submarine simulation.

Correspondence concerning this article should be addressed to Shayne Loft, School of Psychological Science, University of Western Australia, Crawley WA 6009, Australia. E-mail: shayne.loft@uwa.edu.au

Onnasch et al., 2014). Several accidents have occurred in industry where humans have failed to adequately regain manual control (e.g., the grounding of the Royal Majesty off the coast of Nantucket in 1995; Air France 447, which nosedived 38,000 feet into the Atlantic in 2009).

In this article, three experiments are presented, using simulations of submarine track management, that examine the extent to which automation could be designed to benefit performance while ensuring that individuals can maintain SA and regain manual control of previously automated tasks. Our application domain, submarine track management, requires submariners to coordinate the output from the submarine's passive sonar, with other sensors, to compile a coherent tactical picture of the position and behavior of contacts in relation to their own vessel (Ownship) and to strategic landmarks (Kirschenbaum, 2011). Track management is similar to an increasing array of work contexts that require operators to monitor computer screens that display abstract features of dynamic situations occurring outside of the operator's actual physical perceptual experience. Examples include air traffic control, unmanned vehicle control, and air battle management. In these work settings, experts typically remain in charge of making task decisions based upon abstract display information, but increasingly, automation is provided to facilitate information acquisition and analysis.

The first objective of our research was to examine the extent to which static automation could benefit performance and reduce workload, without degrading operator SA or return-to manual performance. In addition, we were motivated by the practical concern that there are situations in which operators are required to perform multiple interdependent tasks, only some of which may be automated. We reasoned that operators in complex work systems may find it more difficult to perform nonautomated tasks that share information processing requirements with concurrently automated tasks if they are processing less of the raw information related to the tasks being automated (automation-induced complacency). To our knowledge, this possibility has never been tested, and we did so by examining whether the use of static automation impaired performance on an interdependent nonautomated task.

The second objective of our research was to examine whether the possible risks of automation to participant SA, interdependent nonautomated task performance, and return-to-manual performance, could be minimized by designing automation to be used intermittently rather than continuously, depending on perceived (*adaptable* automation; Scerbo, 2001) or objective (*adaptive* automation; (Kaber & Riley, 1999; Parasuraman, Mouloua, & Molloy, 1996; Scerbo, 1996) task demands. The rationale for implementing adaptable and adaptive automation is to engage automation only when task demands rise, in order to facilitate performance and decrease workload. Automation is subsequently disengaged when task demands decrease, to encourage operators to update their SA, and maintain their performance on nonautomated tasks by continuing to adequately attend to the displayed task information relevant to the tasks being intermittently automated.

Submarine Track Management and the Degree of Automation

Personnel on a submarine usually do not have visual contact with the various vessels, landmarks or other navigational objects (known as contacts) that may be located around them. Consequently, they must rely on information gathered from the submarine sensors to create their own "view" of their surrounding area. The submarine command team comprises a number of departments including navigation, communications, and sensor and weapons systems, all coordinated by the watch leader. The track manager uses information from outstations, including passive sonar and the periscope, to create a tactical picture of the position and behavior of contacts (their bearing, range, course, and speed) relative to the submarine and to strategic landmarks. The track manager effectively acts as the information manager, informing the command team's decision making with respect to maneuvering the submarine during missions (Kirschenbaum, 2011; Stanton, 2014).

The core tasks of the submarine track manager simulated in the current experiments were developed based on observations and interviews with Royal Australian Navy submariners in real track management combat systems. Participants in the current study worked with two displays (see Figure 1). The left monitor presented a tactical display of the area of operations, including strategic landmarks, contacts currently detected within the operational area, and the Ownship represented in the center of the tactical display. The right display presented a sonar time-bearing plot (referred to as a *waterfall display*), representing the bearing of each contact on the tactical display in relation to Ownship and how those bearings have changed with time. Participants used these displays to complete three tasks. The contact "classification" task required participants to judge how long each contact spent inside landmarks on the tactical display, to identify a contact as friendly, merchant and so forth. A second task required participants to monitor changes in contact heading (course) to determine the closest point of approach (CPA) of contacts to the Ownship. Finally, the "dive" task required participants to integrate contact location and heading information to determine when the submarine could safely dive.

The degree of automation implemented in the simulation was derived from a recent classification system defining "more versus less automation" by Onnasch et al. (2014). Automation can do more or less "work" (*levels of automation*; Sheridan & Verplank, 1978) at each of the following four stages of human information processing (*stages of automation*; Parasuraman et al., 2000): information acquisition, information analysis, decision making, and action execution. Increasing levels within a stage and/or implementing automation at later stages increases the degree of automation (DOA; Onnasch et al., 2014). Automation in the submarine control room, and in other settings such as air traffic control, is typically designed to help the operator acquire and analyze information, but not necessarily to make task decisions or execute task actions. Similarly, automation in the current experiments was designed to track when contacts first entered strategic areas of interest or when contacts made heading changes, but task decisions and their execution were left to manual control.

Will a Low Degree of Static Automation Produce Benefits and/or Costs?

The potential risks of automation, such as complacency, loss of vigilance, and loss of SA are well documented (Parasuraman & Riley, 1997; Parasuraman & Wickens, 2008). The conventional wisdom has been that the more automation is applied, the greater

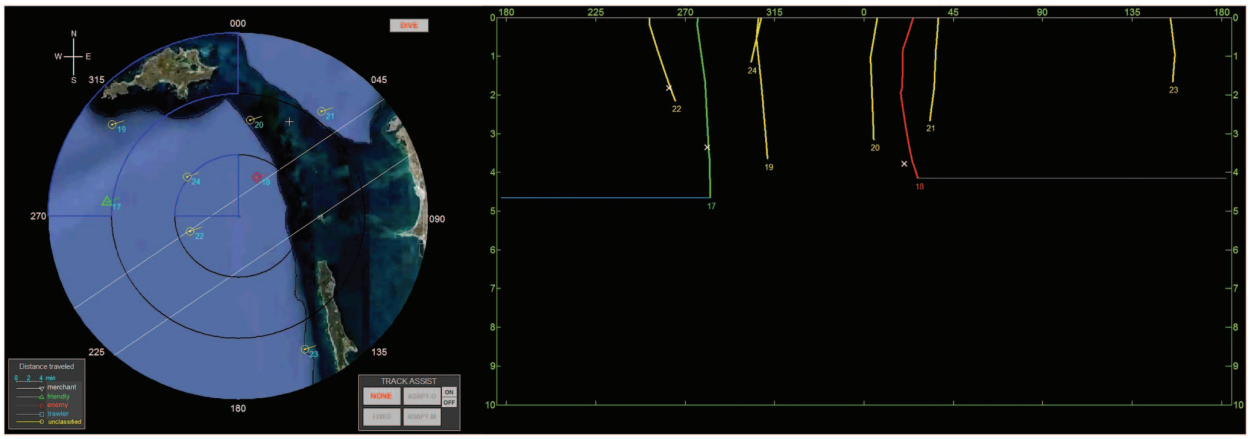

Figure 1. An example submarine track management simulation scenario (Chen, Loft, Huf, & Visser, 2014). The display on the left is the tactical display which represents the area of operations, with Ownship located at the center of the tactical display. On the tactical display, the concentric rings indicate the distance from Ownship. The rings extend in 5-km increments. The parallel white (lighter) lines indicate a shipping lane. The two friendly sectors are bounded by the blue (thicker) lines. The fishing areas are darker in color, which depicted shallower waters more likely to have accessible fish. Note that the other two Australian coastal maps used different physical arrangements of these strategic zones in their area of operations. Eight contacts are displayed in Figure 1. The leader lines projecting from the center of each contact icon indicates the current heading of each contact, and the contacts are numbered. The red (darker) diamond icon for Contact 18 indicates a contact that has been classified as an enemy. Contact 17 is a green (darker) triangle because it has been classified as a friendly. Contacts represented by yellow circles have yet to be classified. The screen on the right is the waterfall display, which is a sonar time-bearing plot of the history of each contact shown on the tactical display. The waterfall display provided the bearing of contacts (along the top horizontal axis) in relation to Ownship, and indicated how those bearings changed with time (on the vertical axis). The waterfall display presented this data as vertical lines ("sound tracks"), which "grew" down with time, representing track history. Each contact track on the waterfall display is numbered, and the color matches the classification given to the contact. Participants (or the automation) could place horizontal lines on the waterfall display when the contact entered an area of interest or to track how long they have been visible. In Figure 1, the participant has attempted to indicate when several closest points of approach (CPAs) had occurred by marking the corresponding sound track of that contact with a cross on the waterfall display. Participants clicked the Dive button to signal the Ownship to dive. When a contact abrupts off the screen the track for that contact terminates from both displays. The automation was referred to as *track assist*. The automation interface (shown at the bottom right of the tactical display) allowed participants to verify the automation condition in which they were currently operating. Participants click the ON and OFF buttons to control the automation (track assist) in the adaptable automation condition. From "Static and Adaptable Automation in Simulated Submarine Track Management," by S. Chen, S. Loft, S., Huf, and T. Visser, 2014, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*, p. 2281, Thousand Oaks, CA: SAGE. Copyright 2014 by SAGE Publications. See the online article for the color version of this figure.

the cost to SA and, potentially, to return-to-manual performance. Consistent with this, a meta-analysis conducted by Onnasch et al. (2014) found that the benefits of static automation to routine performance and workload, during periods when automation functioned reliably, increased with greater DOA but that SA and performance when participants had to resume manual control decreased with greater DOA. Specifically, the costs of automation to SA and return-to-manual performance were more likely to occur when the DOA crossed the boundary from information acquisition and analysis to action selection and implementation. Onnasch et al. (2014) concluded that when automation moves across this "critical" boundary, the operator is relieved of some or all aspects of choosing an action and this greatly reduces the extent to which they monitor the raw situation information processed by the automation (also see Li, Wickens, Sarter, & Sebok, 2014).

On the basis of these findings, we reasoned that providing our participants with a relatively low degree of static automation (information acquisition/analysis) in simulated track management could benefit performance and reduce workload compared with when no automation was provided (Experiments 1, 2, and 3). Moreover, our logic was that engaging participants to make task decisions, and execute associated actions, should encourage them at least partially to attend to the raw information related to the tasks being automated. Thus, we predicted that a relatively low DOA might not harm operator SA (Experiments 1, 2, and 3) or return-to-manual performance (Experiment 3), compared with conditions without automation.

In addition, we examined the impact of static automation on nonautomated task performance (the dive task). On the one hand, the anticipated reduction in workload from static automation use

should provide the operator with the spare cognitive capacity to more effectively manage other nonautomated tasks (Loft, Smith, & Bhaskara, 2011; Loft, Smith, & Remington, 2013; Manzey et al., 2012; Rovira, McGarry, & Parasuraman, 2007; Sethumadhavan, 2009). On the other hand, we designed our nonautomated dive task such that it required the assessment of information also necessary to complete the classification and CPA tasks (i.e., tracking the location and the heading of contacts). Thus, performance on the interdependent dive task could be degraded if participants scrutinized contact location and heading information less closely (automation-induced complacency; Parasuraman & Manzey, 2010; Wickens et al., 2015) when using static automation compared with no automation. To our knowledge, this is the first research to have actively examined the interaction between automation and performance on an interdependent nonautomated task.

Adapting Automation to Keep the Operator Engaged

A potential way of mitigating the costs of static automation is to periodically reallocate automated tasks to human manual control, in order to increase the extent to which operators update their SA by attending to the raw information sources feeding the automation (Farrell & Lewandowsky, 2000; Parasuraman, Galster, Squire, Furukawa, & Miller, 2005; Wickens et al., 2015). Such a work system, designed to engage and disengage automation according to the perceived or objective task demands placed on the operator, has recently been described as one of the "more important ideas in the history of human factors/ergonomics" (Hancock et al., 2013, p. 11). Readers may question how intermittent automation usage could possibly yield benefits equal to the continuous use of automation. The answer is that, in many operational settings, it is common for task demands to rise and fall periodically (Loft, Sanderson, Neal, & Mooij, 2007; Remington, Folk, & Boehm-Davis, 2012). Thus, when task demands fall, automation could be reduced so that the operator completes tasks manually, keeping the operator engaged, facilitating better SA, and bolstering competent task completion. The key question is how to balance timely engagement of automation when task demands rise, in order to maximize performance, with disengagement when demands decrease, to encourage the participant to update their SA.

There are several potential ways to trigger the engagement and disengagement of automation. An automation trigger might be based on an operator performance decrement threshold (e.g., Calhoun, Ward, & Ruff, 2011), a physiological indicator of operator workload/stress (e.g., Wilson & Russell, 2007), or a secondary task measure of workload (Kaber & Riley, 1999). This is known as *adaptive automation*. Although evidence suggests that these triggers can be beneficial, a potential problem is that they are all reactive in that they only engage automation after performance degrades or after workload is elevated. A better system would be proactive, engaging automation prior to such problems arising. We reasoned that operators themselves might be well placed to proactively adapt automation. Thus, in Experiments 1 and 2, we allowed participants to decide when to engage and disengage automation—a condition referred to as *adaptable automation* (Scerbo, 2001).

Notionally, if adaptable automation can be engaged before problems arise, operator performance and workload should benefit. Also, the process of deciding when to use automation might help maintain SA, thereby reducing return-to-manual deficits. On the other hand, it is possible that deciding when to engage/disengage automation may impair performance by increasing cognitive load (Kaber & Riley, 1999). It is also possible that operators are not always accurately able to monitor their task demands/performance on a moment-by-moment basis and thus will have some difficulty forming effective strategies for invoking automation (i.e., limitations in operator metacognition; Flavell, 1979; Osman, 2010). Indeed, the evidence to date has been mixed. Sauer, Nickel, and Wastell (2013) found that adaptable automation facilitated performance and reduced workload compared with static automation, but was no better than adaptive automation. In contrast, several other studies have reported that adaptable automation decreased performance and increased workload compared with adaptive automation based on physiological triggers (Bailey, Scerbo, Freeman, Mikulka, & Scott, 2006) or performance (Kidwell, Calhoun, Ruff, & Parasuraman, 2012).

These mixed results may have another explanation. Prior research examining adaptable automation shares two notable limitations. First, in several studies, automation engaged by the operator was automatically disengaged after a short interval (in as little as 10 s; Bailey et al., 2006; Kidwell et al., 2012). Performance and workload might be more likely to improve if the operator could also decide when to disengage automation. Second, no studies have made comparisons between adaptable and no-automation conditions or measured the potential costs of using adaptable automation to SA, interdependent nonautomated tasks, or return-to-manual performance. Thus, a clearer indication of the utility of adaptable automation is needed to demonstrate that adaptable automation can benefit performance and workload compared with no-automation conditions, and without costs to SA, interdependent nonautomated tasks, or return-to-manual performance.

Experiment 1

In Experiment 1, we examined the effectiveness of static and adaptable automation in a simulated submarine track management task. Task load (number of displayed contacts) was systematically varied. Participants monitored two adjacent displays to perform three tasks (see Figure 1). The contact classification task required participants to judge how long each contact spent inside certain geospatial boundaries, which defined contacts as friendly, merchant and so forth The CPA task required participants to mark the closest point of approach of contacts to the Ownship. Under some conditions, these tasks were supported by a relatively low degree of automation (information acquisition/analysis), which eliminated the need for participants to monitor when contacts first entered an area of interest (classification task) or to monitor contact heading (CPA task). The dive task, which was not supported by automation, required participants to be aware of the location and heading of all contacts, both in relation to other contacts, and to the Ownship.

Participants completed three 27.5-min scenarios, each corresponding to different Australian coastal maps. A within-subjects design was used: Each participant completed a no-automation, a static automation, and an adaptable automation condition. SA was measured using the Situation Present Assessment Method (SPAM; Durso & Dattel, 2004). The SPAM measure has been shown to predict performance in simulations of track management, without

interfering with primary task performance (Loft, Bowden, et al., 2015). To measure subjective workload, the Air Traffic Workload Input Technique (ATWIT) was administered several times during each scenario, and the National Aeronautics and Space Administration Task Load Index (NASA-TLX) was administered after each scenario. We were mindful of participants potentially overusing the automation to make their experience in the experiment easier. To prevent this, we limited the time the automation could be used to 10 min for each scenario (on the basis of results from a pilot experiment). Participants were informed of this limit and were encouraged to use automation only when necessary to maintain performance.

Method

Participants. Participants were 38 (15 females) undergraduate psychology students (M = 19.68 years, SD = 3.56) who volunteered to take part in the experiment for course credit.

Simulated submarine track management task. Participants performed a simulated track management task (see Figure 1). The left tactical display was a bird's eye view of the area and displayed concentric range rings indicating distance from the center point, that is, the location of Ownship. The location and heading of the contacts was presented on the tactical display. The right waterfall display provided the bearing of contacts (along the top horizontal axis) in relation to Ownship and indicated how those bearings changed with time (on the vertical axis). The waterfall display presented this data as vertical lines ("sound tracks"), which "grew" down with time, representing track history. The number of contacts on the displays (task load) periodically increased (peaking at eight contacts) and decreased three times (plateauing at one) during each 27.5-min scenario.

The automation was referred to as *track assist*. The automation interface at the bottom right of the tactical display (illustrated in Figure 2a) allowed participants to verify the automation condition in which they were currently operating. In the adaptable condition, participants engaged and disengaged automation by clicking the ON and OFF button. A countdown clock was provided that indicated the available automation time remaining.

Contact classification task. Participants classified contacts according to their movement on the tactical display. Contacts were classified after they had spent more than two continuous minutes within a specific area of the tactical display. A contact was a "friendly" if it spent more than 2 continuous min within the sectors on the tactical display bounded by the blue lines. A contact was a "merchant" if it spent more than 2 continuous min on the tactical display within the "shipping lane" denoted by the two parallel white lines. A contact was a "trawler" if it spent more than 2 continuous min on the tactical display within the "shallow" dark blue areas. A contact was an "enemy" if, in the first 4 min of its presentation, it had not spent at least 1 continuous min in any classification zone. As shown in Figure 1, to help determine whether a contact had spent more than 2 min in an area of interest, participants were advised that they could place horizontal lines on the waterfall display when the contact entered an area of interest. Once that line reached the 2-min indicator on the waterfall display, the contact could be classified as friendly, merchant or trawler. Participants could also identify enemy contacts by noting that a contact with no horizontal lines on the waterfall display had reached the 4-min indicator. Note that the term *enemy* (a common, succinct, and easily recognized vernacular) was used instead of *possible hostile*, the correct term used by submariners.

A relatively low DOA was used, below the 'critical boundary' suggested by Onnasch et al. (2014), supporting the acquisition/analysis human information processing stage. The automation reduced the need for participants to determine when contacts entered an area of interest on the tactical display by placing horizontal timing lines automatically on the waterfall display when a contact entered an area of interest. However, the automation served no further function. The participant still had to monitor the subsequent behavior of the contacts after they entered the area to ensure they continuously remained in that area during the designated time period before the horizontal blue line reached the 2-min mark on the waterfall display (or departed in under 1min if the contact was a potential enemy), to make the classification decision. This DOA also required participants to remember which horizontal blue line on the waterfall display was associated with which contact on the tactical display. When automation was disengaged, the horizontal lines on the waterfall display remained but new lines had to be manually entered by the participant.

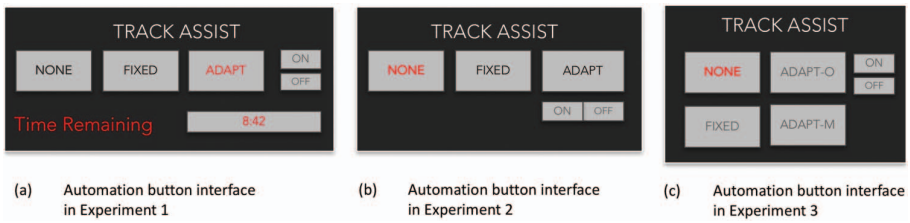

Figure 2. The automation interface buttons for each experiment, which were located at the bottom right of the tactical display. For Experiments 1 and 2 (within-subjects design), participants could see which condition they were operating under by the color of the button label (i.e., a red (darker gray) button meant automation was activated). The adaptable condition in Experiments 1 and 2 (ADAPT), and the adaptable operator-triggered (ADAPT-O) condition in Experiment 3 could be activated and deactivated by the ON and OFF buttons. In Experiment 1, which had limited automation (10 min) a countdown clock was provided. In Experiment 3, the adaptive condition was indicated by the button labeled ADAPT-M (adaptive machine-triggered). See the online article for the color version of this figure.

CPA task. A CPA was defined as the point at which a contact would be at its closest to Ownship. This was essentially the point where a contact heading toward Ownship (e.g., Contact 22 in Figure 1) turned away from the Ownship. The CPA task required participants to (a) track contacts heading toward Ownship, (b) track when these contacts made heading changes, and (c) indicate the time that the CPA occurred by marking the corresponding sound track of that contact on the waterfall display. Each of the 24 contacts presented in each scenario had one CPA each.

The CPA automation supported the information acquisition/analysis stage. When automation was engaged a track history line, resembling an extended ship's wake, was visible on each contact on the tactical display. This reduced the need for the participant to track which contacts made heading changes; however, the automation served no further function. To make accurate CPA decisions, participants needed to mark accurately the timing of the CPA on the waterfall display, and the CPA automation did not interpret the track history lines or alert the participant to when a heading change had occurred. When the automation was turned off in the adaptable automation condition, all track history lines remained on the tactical display but were not updated to reflect further contact movement.

Dive task. The dive task was not automated. Participants were instructed to dive the submarine when (a) all contacts on the tactical display were heading in the same direction and (b) one of the contacts was heading directly toward Ownship. This required participants to be aware of the location and relative headings of contacts, both in relation to other contacts and to Ownship. The time periods that the dive conditions were met were the "dive windows." Dive-window durations varied between 10 s and 30 s. There were 9 or 10 dive windows per scenario. Participants clicked the dive button to signal the Ownship to dive.

Measures.

SA. SA was measured six times during each scenario. SA queries for information related to each of the three tasks and each of the three SA levels (Endsley, 1995a, 1995b) were presented. Although not traditionally used by researchers deploying SPAM, we found Endsley's definitions of SA levels useful in developing our SPAM queries. The six SPAM queries used per scenario were taken from the pool of SA questions presented in Table 1. The SPAM queries were delivered over headphones. SPAM distinguishes workload from SA by warning the operator that a query is in the queue and waiting until the operator accepts the query (Durso & Dattel, 2004). Before a SPAM query was presented, the participant was asked "Are you ready for a question?" An accompanying "yes" and "no" box appeared on screen, and the participant would click to respond to indicate that he or she was ready to accept the question. The time taken between ready prompt audio and when the *yes* box was clicked is referred to as *SPAM accept time* and often correlates with subjective workload (Loft, Bowden, et al., 2015; Vu et al., 2012). SPAM response time (RT) is measured as the time between when the experimenter completes asking the question and the time the participant responds. The logic behind SPAM is that participants with better SA would know where to find the correct answer to a question about the situation on the screen faster and/or more accurately (Durso & Dattel, 2004; Loft, Bowden, et al., 2015).

Workload. There were two subjective measures of workload. The ATWIT (Stein, 1985) was presented on the tactical display every minute, and participants had 10 s to click a workload level between 1 and 10, described as *very low* (1 to 2), *moderate* (3 through 5), *relatively high* (6 through 8), and *very high* (9 to 10). The presentation of the scale did not prevent participants from completing tasks. The NASA TLX (Hart & Staveland, 1987) was completed after each scenario. Participants rated their workload on a 20-point scale on six dimensions of workload: mental demands, physical demands, temporal demands, own performance, effort, and frustration. Participants then indicated the degree to which each of these six dimensions was "the more important contributor to workload" in pairwise comparisons between the dimensions. The overall NASA TLX workload score was calculated by multiplying the resulting weighting of each dimension with the corresponding rating, then dividing the total by the number of pairwise comparisons.

Trust in automation. History-based trust (trust evolved from interaction with the automation) was measured with a six-item questionnaire. The questionnaire was adapted from Merritt (2011). The items were altered to refer to the track assist automation and included items such as "I can depend on track assist," "I have confidence in the information given by track assist," or "I can rely on track assist to behave in consistent ways." This history-based trust was measured after each automated scenario was completed. Participants responded to each item via a 5-point Likert scale (1 = *strongly agree* to 5 = *strongly disagree*).

Procedure. Participants completed two 2.5-hr time slots on consecutive days. On Day 1, they completed a number of individual differences measures (e.g., working memory, personality), as part of a different research project. After this, training started with

Table 1
The SPAM (Situation Present Assessment Method) Queries Used to Measure Participant Situation Awareness in Experiment 1

SA level	SPAM queries	
1	Which vessel is closest to a Y zone?	Has X been visible for more than 4 minutes?
	Is X heading toward a Y zone?	Is X heading towards/away from you?
	Is X in a Y zone?	Which vessel is heading directly towards you?
2	Has X been in a Y zone for more than 1 minute?	Which vessel had the most recent kink in its sound track?
	Which vessel most recently crossed a classification boundary?	Are any vessels on the same course?
3	Could X be within the Y zone in 4 min time?	Which vessel is most likely to show the next CPA?
	Could X cross a boundary within 2 min?	Which vessel requires a course change to open a dive window?
	Which unclassified vessel is most likely to be a trawler/friendly/enemy/merchant?	

Note. SPAM = Situation Present Assessment Method; CPA = Closest Point of Approach.

a 40-min audiovisual PowerPoint presentation that included a number of "learning checks" that had to be answered correctly before the presentation could continue. This was followed by a narrated video of the simulation (prerecorded) where all tasks and the three conditions (none, static, and adaptable) were demonstrated. Finally, on Day 1 participants completed a 27.5-min practice scenario. Participants were told to complete the tasks without engaging automation during the first half of the scenario and then asked to engage automation during the second half.

On Day 2, participants were presented with a 15-min PowerPoint presentation that reminded them of the pertinent points from Day 1. Participants then completed the three 27.5-min scenarios, each of which contained unique contacts presented in different maps. Each participant completed a no automation condition, a static automation condition, and an adaptable automation condition. The order of conditions and the assignment of scenario to condition were counterbalanced. After each scenario, participants were asked to complete the NASA TLX and to complete the history-based trust questionnaire after completing the static automation and adaptable automation scenarios.

Results and Discussion

The CPA hit rate was the number of CPAs correctly marked on the waterfall display per scenario, divided by 24 (i.e., the total number of CPAs presented per scenario). Each time a participant placed a cross on the waterfall display, the coordinates were recorded. The CPA cross could be placed at any time 1.5 s before the actual CPA (to account for the small movement preceding any course change including a CPA, which could be noticed by the participant) or subsequent to when the CPA occurred, as long as the cross was placed on the correct sound track and within a 3-mm radius of the actual CPA point. Otherwise the cross was recorded as a false alarm. The exact potential number of contacts and associated events to make a CPA false alarm response was indeterminable, but we estimated this parameter. It is more likely a CPA false alarm would be made in response to a contact course change. For each scenario, there were 69, 71, or 73 total course changes. Accordingly, the false alarm rate was estimated to be the number of false alarms, divided by the number of course changes for that scenario (minus 24, which was the actual number of CPAs). CPA performance was then calculated by subtracting the CPA false alarm rate from the CPA hit rate. CPA RTs (and RTs for other tasks and SPAM) were based on correct decisions.

The dive hit rate was the number of correct dive responses made during dive windows divided by the total number of dive windows per scenario. The potential number of opportunities to make a dive response was indeterminable. The most likely time a dive false alarm response would be made was during a course change, as course changes were always needed for a dive window to transition from closed to open. As there were fewer dive windows than CPAs, and because all contacts needed to be on the same heading for a dive window to open, it was not likely every course change would have been mistaken for a dive window. Consequently, we calculated the dive false alarm rate as the number of dive false alarms divided by half the number of course changes for the scenario (minus the actual number of dive windows which was 9 or 10). Dive task performance was calculated by subtracting the dive false alarm rate from the dive hit rate.

The means and 95% within-subject confidence intervals for task performance, SA, and subjective workload are presented in Table 2. Within-subject confidence intervals for Experiments 1 and 2 were calculated using the method recommended by Morey (2008). Within-subject design effect sizes for Experiment 1 and 2 were calculated based on recommendations by Morris and Deshon (2002). For each dependent variable, we conducted planned contrasts that directly evaluated our research questions by comparing the static automation condition to the no automation condition and the adaptable automation condition to the no automation condition (Rosenthal & Rosnow, 1985). These inferential statistics are presented in Table 2. Estimates of Cohen's d suggested we had a power of .97 to detect medium-to-large size effects (Cohen, 1988).

Even with appropriate counterbalancing procedures, dependent measures in within-subject designs can be impacted by order effects caused by practice or asymmetric transfer (Poulton, 1982). We therefore assessed for order effects when comparing the static automation condition to the no automation condition, and the adaptable automation condition to the no automation condition. To do this, we originally included the order of condition presentation (static automation presented first vs. no automation presented first or adaptable automation presented first vs. no automation presented first) as an additional variable in each of the analyses reported below for Experiment 1. However, we found no reliable evidence that the order of presentation of condition had a significant impact on the degree to which static or adaptable automation influenced performance, SA, or workload in Experiment 1 (smallest $p = .12$). Thus, for brevity, we collapsed across the condition order variable in the analyses presented in the following text.

Static automation versus no automation. As shown in Table 2, participants were more accurate and faster to classify contacts when using static automation compared with no automation. Participants made more accurate CPA task decisions when using static automation compared with no automation, but were also slower to make these CPA task decisions. The slowed CPA decisions likely reflect the fact that the automated track history allowed participants the option to detect a CPA well after it had occurred. This would have improved accuracy but yielded slower RTs because after making a delayed CPA decision, participants would then have needed to estimate the correct CPA location to mark on the waterfall display by using the contact track history on the tactical display to determine the time passed since the CPA. There was no difference in dive task accuracy or dive task RT when using static automation compared with no automation. As shown in Figure 3, ATWIT scores rose and fell with task load (the number of contacts). Participants reported lower workload on the NASA TLX and the ATWIT when using static automation compared with no automation. There was no significant difference between static and no automation conditions for SPAM accuracy or SPAM RT.

Adaptable automation versus no automation. As shown in Figure 3, automation usage generally coincided with peaks in task load and subjective workload, suggesting participants used automation strategically in response to task demands. However, despite a total of 10 min of automation being available, on average participants only used automation for 6.85 min (95% CI [5.81, 7.89]).

On the classification task, there was no difference in accuracy or RT when using adaptable automation compared with no

Table 2
Descriptive and Inferential Statistics for Performance, Situation Awareness, and Subjective Workload in Experiment 1

Task	Condition	*M*	95% CI	*T*	*p*	Cohens *d*
Classification [Hit]	None	.84	[.80, .87]			
	Static	.91	[.87, .94]	3.08	.004*	.54
	Adaptable	.85	[.83, .88]	.91	.37	.08
Classification [RT]	None	24.91	[21.47, 28.35]			
	Static	18.92	[15.85, 21.99]	−2.96	.005*	−.50
	Adaptable	24.21	[21.55, 26.87]	−.38	.70	−.08
CPA [Hit–FA]	None	.31	[.25, .38]			
	Static	.51	[.44, .57]	4.95	<.001*	.81
	Adaptable	.42	[.35, .47]	2.68	.01*	.45
CPA [RT]	None	18.27	[14.13, 22.42]			
	Static	23.16	[19.93, 26.39]	2.07	.046*	.34
	Adaptable	19.44	[16.72, 22.14]	.55	.58	.09
Dive [Hit–FA]	None	.78	[.74, .82]			
	Static	.76	[.70, .81]	−.77	.45	−.22
	Adaptable	.77	[.71, .82]	−.38	.71	−.10
Dive [RT]	None	8.94	[7.67, 10.21]			
	Static	9.31	[8.15, 10.47]	.55	.59	.09
	Adaptable	9.13	[7.74, 10.52]	.23	.82	.05
NASA TLX	None	63.46	[60.11, 66.80]			
	Static	54.18	[50.78, 57.56]	−4.24	<.001*	−.71
	Adaptable	59.76	[57.74, 61.78]	−2.49	.02*	−.40
ATWIT	None	5.04	[4.83, 5.24]			
	Static	4.32	[4.08, 4.56]	−5.05	<.001*	−.82
	Adaptable	4.85	[4.69, 5.01]	−1.99	.054	−.33
SPAM [Accuracy]	None	.88	[.82, .94]			
	Static	.92	[.88, .97]	1.40	.17	.23
	Adaptable	.92	[.87, .96]	1.24	.22	.20
SPAM [RT]	None	2.78	[2.44, 3.09]			
	Static	2.93	[2.62, 3.23]	.87	.39	.18
	Adaptable	2.87	[2.53, 3.19]	.49	.63	.12

Note. The 95% within-subject confidence intervals are presented in parentheses (Morey, 2008). The inferential statistics present the planned contrasts between the static automation condition and the no automation condition and the adaptable automation condition and the no automation condition for each dependent variable. The degrees of freedom for each planned contrast was 37. CPA = Closest Point of Approach; RT = Response Time in seconds; FA = False Alarm; NASA TLX = National Aeronautics and Space Administration Task Load Index; ATWIT Air Traffic Workload Input Technique; SPAM = Situation Present Assessment Method.
* $p < .05$.

automation. However, participants made more accurate CPA decisions when using adaptable automation compared with no automation. There was no difference in CPA task RT when using adaptable automation compared with no automation. Dive task accuracy and RT did not differ between adaptable and no automation conditions.

NASA TLX scores were lower when using adaptable compared with no automation. However, this benefit only approached significance when measured by ATWIT. There was no difference in SPAM accuracy or SPAM RT when participants used adaptable automation compared with no automation.

Trust and automation preferences. There was no significant difference in the history-based automation trust between the static ($M = 4.44$, 95% CI [4.33, 4.55]) and adaptable ($M = 4.25$, [4.14, 4.37]) automation conditions, $t(37) = 1.66$, $p = .10$, $d = 0.27$. Of the 38 participants, 22 (58%) preferred static automation, 13 (34%) preferred adaptable automation, and 3 (8%) preferred no automation. Postexperiment verbal reports indicated that many participants disliked adaptable automation because of the effort required to budget automation time.

Experiment 2

In Experiment 1 we found that a low degree of static automation reduced workload and improved performance, without any cost to SA or to nonautomated dive task performance. This suggests it may be possible to design static automation to provide significant benefits without associated costs. An alternative explanation for the results, however, is that the SPAM measure lacked the sensitivity to detect underlying changes in SA. To test this possibility, we replicated the paradigm in Experiment 1, while using the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995b) to measure SA.

The SAGAT method involves periodically pausing and blanking the task display in order to query participants about the current and likely future state of the simulated environment. Concurrent evidence from our laboratory found that SA as measured by SAGAT was a stronger predictor of track management performance than SA as measured by SPAM (Loft, Bowden, et al., 2015). SAGAT might be more predictive because the greater number of queries collected with SAGAT provides a more reliable SA estimate,

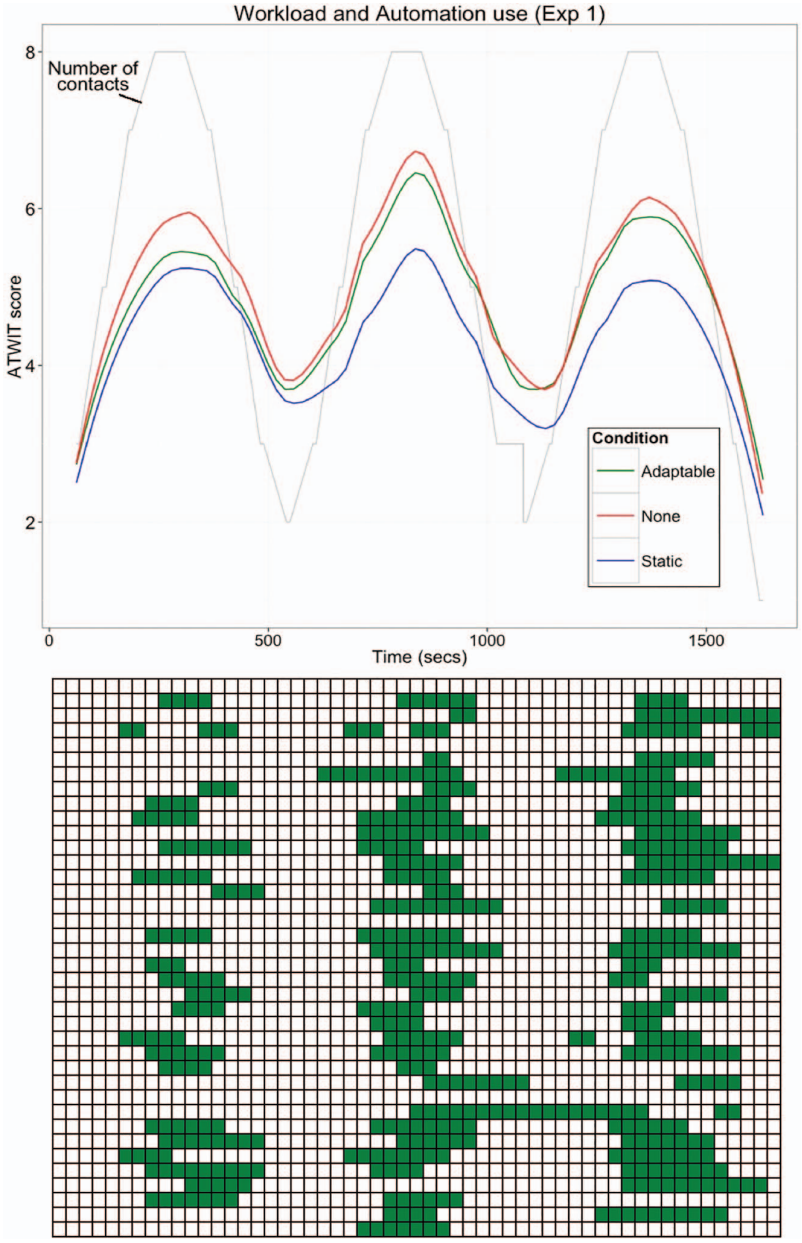

Figure 3. The task load, the workload measured by Air Traffic Workload Input Technique (ATWIT) in all three conditions (top plot) and the times during which adaptable automation was used (bottom plot) during Experiment 1. The number of contacts is shown by the light gray line and rises and falls three times. The workload probe (ATWIT) was presented every minute with a maximum of 27 data points over the 27.5 min (maximum of 1,026 workload points per automation condition when the 38 participants' data was combined). The ATWIT presentation times were identical for each scenario. A locally weighted regression (*loess*) procedure was used to fit the regression line (solid lines) visible in this Figure (for detail see Cleveland & Devlin, 1988). The *loess* method was used in the 'stat_smooth' function in R with the span parameter (degree of smoothing) set at 0.5. The light gray line is the total number of contacts that were visible at the time. The bottom plot indicates the 30-s blocks during which automation was activated in the adaptable condition for each participant (a 30-s block is coded green (dark gray) when automation was used at some point in time during that 30 s, and coded white when automation was not used at any time during that 30-s time period). See the online article for the color version of this figure.

and/or because SAGAT directly taps into participants' memory for the state of the track management display (Loft, Bowden, et al., 2015). The greater number of SA queries posed to participants through the use of SAGAT also provided an opportunity to explore the relationship between SA and workload as a function of variation in task load (Vidulich & Tsang, 2012; Wickens, 2008). It should also be noted that no prior studies have reported any adverse effects of SAGAT administration on primary task performance (e.g., Loft, Bowden, et al., 2015; Strybel, Vu, Kraft, & Minakata, 2008). Thus we have no reason to believe that the change in SA measure in Experiment 2 would adversely affect primary task performance.

In Experiment 1, the use of adaptable automation provided marginal benefits to performance (facilitating only one of two tasks) and to workload (lowering workload on only one of two measures). However, it is possible that the 10-min limit placed on automation usage imposed an unnecessary mental burden on participants, which may have offset some of the benefits the automation provided. The limit was imposed to reduce the potential for overuse of automation, but this control might have been unnecessary because participants were reluctant to use automation (on average they used less than 7 min of the available 10 min of automation in Experiment 1). Therefore, in Experiment 2 we informed participants that there was no time limit for automation usage in the adaptable condition.

Method

Participants. Participants were 43 undergraduate psychology students (21 female; $M = 20.71$ years, $SD = 5.95$) who volunteered to take part in the experiment for course credit. Data from three participants were not included in the analysis. One participant believed the automation was engaged when it was not when using adaptable automation and therefore did not engage automation. One participant did not make a single correct CPA or classification. The third participant had difficulty hearing, understanding the task, and forgot her glasses.

Task, measures, and procedure. The simulation and training were identical to Experiment 1, with the following exceptions. First, there was no time limit for automation use, and participants were instead instructed to use adaptable automation as required. As a result, the countdown clock located in the track assist automation interface in Experiment 1 was removed (see Figure 2b). Second, SAGAT was used to measure SA. During each scenario the simulation was "frozen" six times. During a freeze, the contact symbols would disappear from the tactical display, the sound tracks would disappear from the waterfall display and questions would appear at the top of the tactical display. Seven SAGAT queries were delivered during each freeze. The first question always asked participants to mark on the tactical display where they thought a randomly preselected contact was located. Two questions each then targeted the SA required for classification, CPA, and the dive task. The six SA queries equally represented the three levels of SA (two queries per SA level). The SA queries were taken from the pool of queries presented in Table 3.

SAGAT allowed us to examine how SA covaried with workload and task load, by using a line of best fit on a scatterplot for each SAGAT freeze. This was possible because each SAGAT freeze yielded a correct percentage score. The SAGAT scores therefore provided a spread of SA scores for each SA query point on the same vertical scale as workload (each SAGAT percentage score was multiplied by 10 to match the ATWIT scale). SAGAT scores for the three scenarios were combined for each condition so there were 18 SAGAT scores across each 27.5-min automation condition per participant (the timing of the six SAGAT freezes per scenario was set across the three scenarios so that when combined, they occurred somewhat regularly over the 27.5 min). A locally weighted regression (*loess*) procedure was used to fit a regression line (Cleveland & Devlin, 1988) on the resulting 240 SAGAT scores per automation condition (six SAGAT scores for each of the 40 participants).

Results

As in Experiment 1, for each dependent variable, we conducted planned contrasts that directly evaluated our research questions by comparing the static automation condition to the no automation condition, and the adaptable condition to the no automation condition. The means, 95% within-subject confidence intervals, and inferential statistics for task performance, SA, and subjective workload are presented in Table 4. Estimates of Cohen's *d* suggested we had a power of .98 to detect medium-to-large size

Table 3
The SAGAT Queries Used to Measure Participant Situation Awareness in Experiments 2 and 3

SA level	SAGAT queries		
1	Which vessel is currently in a Y zone?	How many vessels are heading away from you?	How many vessels are on the same course?
	Is vessel X currently in a Y zone?	Is the vessel at bearing X heading towards/away from you?	Are any vessels heading directly towards you?
2	Is vessel X within 2 minutes from an X zone?	Has vessel X had any kinks in its soundtrack?	Which vessel is currently heading directly towards you?
	Which vessel most recently crossed a classification boundary line?	How many times has vessel X changed course?	Are vessels X and Y heading in the same direction?
3	Could vessel X cross a boundary within 4 min time?	Which vessel would make a CPA if it turned to a heading of xxx?	If all vessels turned onto a course of xxx, which vessel would be heading directly towards you?
	Which unclassified vessel is most likely to be a trawler/friendly/enemy/merchant?	Would a CPA be made for vessel X if it turned to a heading of xxx?	Would vessel X head directly towards you if it turned to a heading of xxx?

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

Table 4
Descriptive and Inferential Statistics for Performance, Situation Awareness, and Subjective Workload in Experiment 2

Task	Condition	*M*	95% CI	*t*	*p*	Cohens *d*
Classification [Hit]	None	.79	[.75, .83]			
	Static	.89	[.86, .92]	4.89	<.001*	.81
	Adaptable	.85	[.81, .89]	2.32	.03*	.39
Classification [RT]	None	25.97	[23.20, 29.65]			
	Static	17.92	[15.49, 20.36]	−4.84	<.001*	−.82
	Adaptable	21.02	[17.16, 24.88]	−2.04	.048*	−.33
CPA [Hit–FA]	None	.32	[.26, .38]			
	Static	.51	[.45, .57]	4.99	<.001*	.81
	Adaptable	.41	[.37, .45]	3.06	.004*	.48
CPA [RT]	None	20.79	[15.96, 25.61]			
	Static	21.16	[17.44, 25.08]	.14	.89	.03
	Adaptable	21.39	[17.80, 24.98]	.23	.82	.04
Dive [Hit–FA]	None	.81	[.76, .86]			
	Static	.76	[.69, .82]	−1.48	−.15	−.22
	Adaptable	.74	[.69, .79]	−2.51	.02*	−.40
Dive [RT]	None	7.76	[6.74, 8.78]			
	Static	8.41	[7.48, 9.34]	1.10	.28	.18
	Adaptable	8.66	[7.75, 9.56]	1.55	.13	.25
NASA TLX	None	68.54	[66.10, 70.98]			
	Static	58.69	[55.63, 61.75]	−5.89	<.001*	−.94
	Adaptable	63.95	[61.39, 66.51]	−3.41	.002*	−.55
ATWIT	None	5.06	[4.86, 5.27]			
	Static	4.57	[4.40, 4.74]	−4.35	<.001*	−.68
	Adaptable	5.00	[4.82, 5.19]	−.47	.64	−.05
SAGAT [Accuracy]	None	.60	[.57, .64]			
	Static	.55	[.52, .59]	−2.15	.04*	−.37
	Adaptable	.57	[.54, .61]	−1.27	.21	−.24

Note. The 95% within-subject confidence intervals are presented in parentheses (Morey, 2008). The inferential statistics present the planned contrasts between the static automation condition and the no automation condition, and the adaptable automation condition and the no automation condition, for each of the dependent variables. The degrees of freedom for each planned contrast was 39. CPA = closest point of approach; RT = Response Time in seconds; 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.
* $p < .05$.

effects (Cohen, 1988). We used the same procedure as Experiment 1 to test for order effects (smallest $p = .08$).

Static automation versus no Automation. Participants were more accurate and faster to make classifications when using static automation compared with no automation. Participants made more accurate CPA task decisions when using static automation compared with no automation. There was no difference in CPA task RT when using static automation compared with no automation. There was no difference in dive task accuracy or dive task RT when using static automation compared with no automation. As shown in Figure 4, variation in the ATWIT scores coincided with variation in task load. Participants reported lower workload on the NASA TLX and on the ATWIT when using static automation compared with no automation. SAGAT accuracy was poorer when participants used static automation compared with no automation.

Overall, the findings of improved classification and CPA performance and reduced workload with the use of static automation replicated Experiment 1. However, in Experiment 2, SA as measured by SAGAT was significantly impaired by the use of static automation.

Adaptable automation versus no automation. As shown in Figure 4, automation usage by participants coincided with task-load and workload variation, replicating Experiment 1. As expected, participants also used automation more in Experiment 2 ($M = 11.80$ min; 95% CI [10.05, 13.55]) than in Experiment 1 ($M = 6.85$ min, [5.81, 7.89]), $t(76) = 4.84$, $p < .001$. However, automation was still only used for less than half the scenario time.

Participants made more accurate and faster classification decisions when using adaptable automation compared with no automation. Participants also made more accurate CPA task decisions when using adaptable compared with no automation, but there was no difference in CPA task RT. Participants' accuracy on the dive task was poorer when they used adaptable automation compared with no automation. There was no difference in dive task RT between the adaptable and no automation conditions. Participants reported lower workload, as indicated by the NASA TLX, when using adaptable automation compared with no automation. However, this benefit was not replicated when workload was measured by ATWIT. There was no significant difference in SAGAT response accuracy when participants used adaptable compared with no automation.

Overall, the performance improvements observed for the adaptable condition were more pronounced in Experiment 2 (occurring for the classification and CPA tasks) compared with Experiment 1

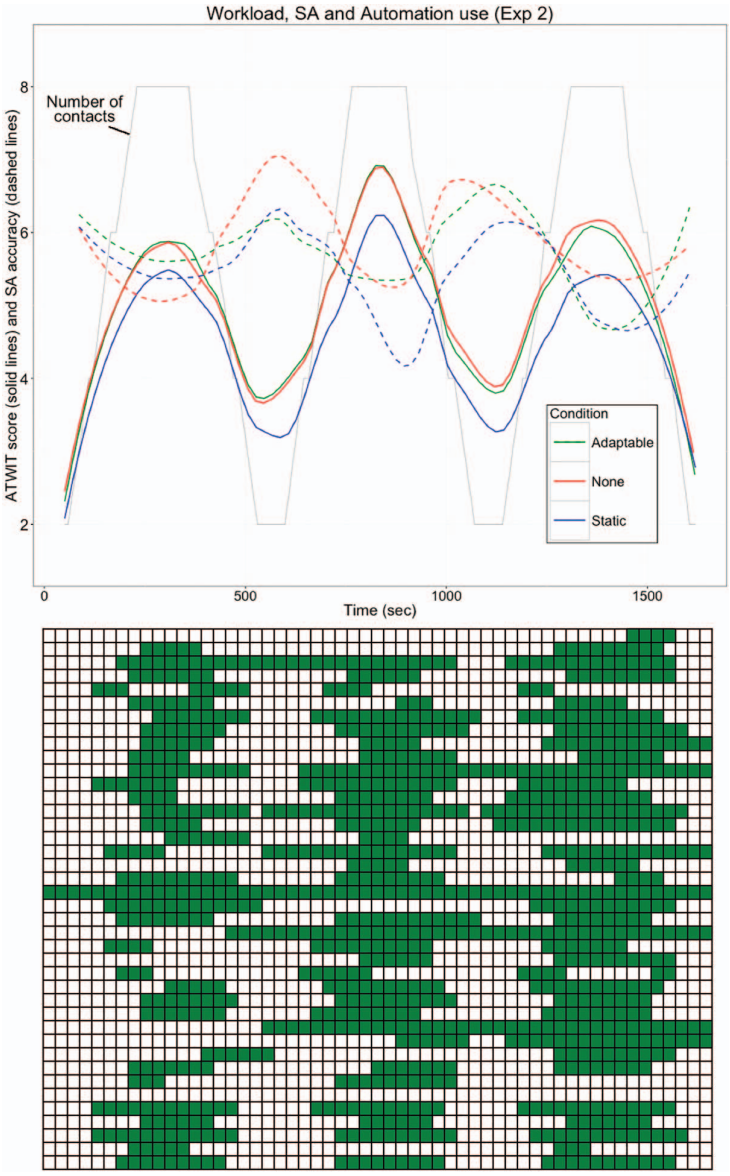

Figure 4. The task load, the workload measured by Air Traffic Workload Input Technique (ATWIT) in all three conditions (top plot) and the times during which adaptable automation was used (bottom plot) during Experiment 2. The number of contacts is shown by the light gray line and rises and falls three times. The workload was plotted using the same method used in Experiment 1, with a maximum of 1,080 workload points per automation condition when the 40 participants' data was combined. The Situation Awareness Global Assessment Technique (SAGAT) scores were also combined and plotted (dashed lines) as a function of time, using the *loess* procedure. The SAGAT freeze times were staggered over the three scenarios and the resulting combined 18 situation awareness (SA) data points were spaced somewhat regularly over the 27.5 min. There were 240 SAGAT freezes (40 participants responding to 6 SAGAT freezes per condition). The relatively fewer data points for SAGAT (compared with ATWIT) meant that the 95% confidence intervals of this function line overlapped between automation conditions at some points and the SA function lines are only used to demonstrate the broad SA changes observed. The bottom plot indicates the 30-s blocks during which automation was activated in the adaptable condition for each participant (a 30-s block is coded green (dark gray) when automation was used at some point in time during that 30 s, and coded white when automation was not used at any time during that 30-s time period). See the online article for the color version of this figure.

(CPA task only). This is likely because participants used automation more in Experiment 2 than in Experiment 1. However, these performance improvements, along with moderate reductions in workload, came at a cost to the dive task.

Trust and automation preferences. There was no difference in the history-based trust scores between the static automation ($M = 4.22$; 95% CI [3.94, 4.49]) and adaptable automation condition ($M = 4.12$, [3.87, 4.37]), $t < 1$). Of the 38 participants, 25 (63%) preferred static automation, 13 (28%) preferred adaptable automation, and 2 (5%) preferred no automation.

SA, workload, and task load. Figure 4 illustrates that SA was negatively related to task load and workload. Increased task load (and consequent increased subjective workload) may have increased the cognitive resources required for maintaining SA (Endsley & Kiris, 1995; Wickens, 2008), and SA may then have become recoverable when task load/workload decreased. Condition did not moderate the variation in SA as a function of task load/workload.

Experiment 3

In Experiments 1 and 2 we found that a low degree of static automation improved performance and reduced workload. However, static automation significantly degraded participant SA in Experiment 2 (using the SAGAT measure). Adaptable automation benefited performance and workload, particularly in Experiment 2 when participants used automation more. There was no evidence of an impact of adaptable automation on SA. However, in Experiment 2, there was a significant 7% decrement in dive task performance in the adaptable automation condition.

In Experiment 3 we replicated our earlier studies using a between-subjects design with each participant completing three scenarios in the same automation condition. The rationale for the between-subjects design was twofold. First, participants may have found it difficult to switch between different conditions and this may have introduced considerable noise in the data. Thus, in Experiment 3 we gave participants greater exposure to an exclusive type of automation. We suspected that participants might use adaptable automation more with extended task and associated adaptable automation exposure, and that this could further facilitate performance and reduce workload compared with Experiments 1 and 2.

Second, the use of a between-subjects design allowed us to introduce a "once off" automation removal state in which participants could no longer use automation. It was not viable to insert an automation removal state into the within-subject designs of Experiment 1 or 2 because research has shown automation complacency induced effects to be high before the first automation removal but to dissipate thereafter (Yeh, Merlo, Wickens, & Brandenburg, 2003). A return-to-manual control impairment would be indicated if, after automation removal, a dependent measure outcome was significantly poorer for participants who had previously used automation compared with those who had never used automation.

The fact that we found only moderate benefits to performance and workload with adaptable automation could reflect the added effort required to decide when to engage and disengage the automation, which might have diverted attentional resources away from task goals (Bailey et al., 2006). Participants may also have experienced metacognitive difficulties in deciding when to use automation. We added an adaptive automation condition to Experiment 3, which removed the need for the participant to decide when to engage/disengage automation while still limiting automation to periods of high workload. We used an adaptive automation trigger based on the number of contacts on the track management displays (task load). We designed the automation to be engaged when there were more than five contacts on the display, and to be disengaged when there were fewer than six contacts on the display. This effectively meant that adaptive automation would be provided for approximately half of each scenario (14.5 min out of 27.5 min), roughly equivalent to the duration and onset/offset times used by participants in the adaptable automation condition in Experiment 2.

To our knowledge, de Visser and Parasuraman (2011) have conducted the only study to have examined task-load adaptive automation triggers, in a supervisory multiple robotic uninhabited vehicle control task. They reported that adaptive automation reduced workload compared with static automation, and improved SA compared with no automation condition; but they found no benefits to performance from adaptive automation. However, the authors suggested that the task was not difficult enough to allow performance differences to manifest. We expected to avoid performance ceilings in Experiment 3 because robust performance differences between conditions were found in our earlier experiments, and thus the adaptive automation could benefit performance compared with no automation.

Method

Participants. Participants were 118 (75 females) undergraduate psychology students ($M = 22.47$ years, $SD = 7.97$) who volunteered to take part in the experiment for course credit. The participants were assigned randomly to one of four automation conditions: No automation ($N = 30$), static automation ($N = 30$), adaptable automation ($N = 29$), and adaptive automation ($N = 29$).

Simulated submarine track management task. The methods in Experiment 3 were identical to Experiment 2, with three exceptions. First, the participants were given training exclusive to their condition. Second, we included an adaptive automation condition. In this condition, participants started each scenario with no automation engaged. The automation was then engaged automatically when task load increased beyond five contacts and was disengaged when task load decreased below six contacts. As illustrated in Figure 5, this meant the three automation durations were 4.67 min, 5.00 min, and 4.83 min (a total of 14.5 min). The automation buttons on the tactical display were redesigned so that all four conditions were present on the interface (see Figure 2c).

Finally, during the third scenario the automation was unexpectedly removed between the first and second task-load peaks, which was at the 10.58-min point in a 27.5-min scenario. At the time of automation removal, a message was overlaid on the tactical display: "Attention. ENEMY SONAR detected. Track Assist turned off. Manual tracking is required." The message was accompanied by an OK button, so that the participant had to acknowledge the message. In the no automation condition, a message was overlaid at the same time. This message read: "Attention, ENEMY SONAR detected. Keep vigilant and continue to track vessels." Automation removal did not occur concurrently with actions required for any task, ATWIT probes or SAGAT queries. At the point of automation removal, the

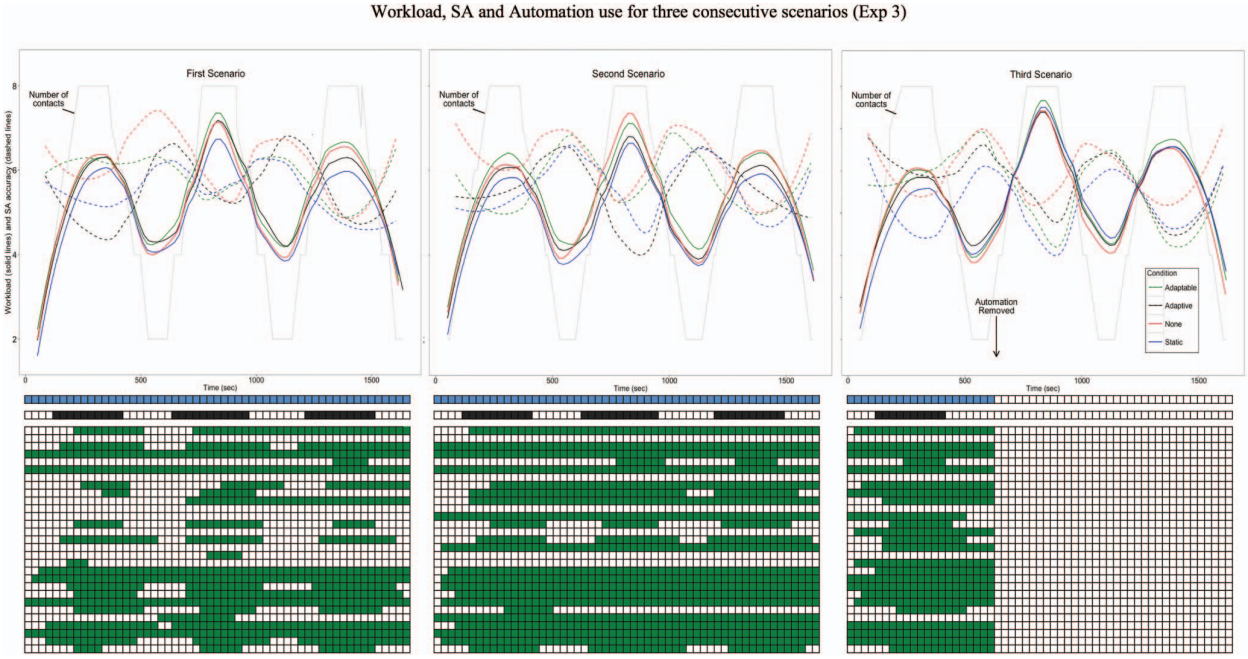

Figure 5. A plot of task load, workload and situation awareness for each of the three scenarios for each of the four automation conditions (top plot), and the times when the automation was used (30s time blocks in the bottom plot) for each condition in Experiment 3. The number of contacts is shown by the light gray line and rises and falls three times. The static automation was activated during the entire first two scenarios and part of the third scenario, and these time periods are coded in blue (light gray) in Figure 5. The adaptive activation times were 2:25 through 7:05, 11:25 through 16:25, and 20:35 through 25:25 min during the 27:30-min scenario, and these time periods are coded in black in Figure 5. The bottom plot indicates the 30-s blocks during which automation was activated in the adaptable condition for each participant (a 30-s block is coded green (dark gray) when automation was used at some point in time during that 30 s, and coded white when automation was not used at any time during that 30-s time period). The 27 Air Traffic Workload Input Technique workload probes per scenario for all participants were combined (max of 810 workload points for the no-automation and the static conditions, and 783 for the adaptable and adaptive conditions), and plotted for each automation condition. The Situation Awareness Global Assessment Technique (SAGAT) scores were similarly combined (180 SAGAT data points for the no-automation and the static conditions and 174 SAGAT data points for the adaptable and adaptive conditions). The workload (solid lines) and SAGAT (dashed lines) were plotted using the same multivariate smoothing method as used in Experiments 1 and 2 (*loess* method using stat_smooth in R, $\alpha = .5$). See the online version of the article for the color version of this figure.

track history for each contact remained visible but ceased updating. The timing lines on the waterfall display also remained but no longer appeared at the point when a contact would enter a new classification area.

Measures.

SA. The SAGAT queries and freeze times were identical to those in Experiment 2. In the adaptive condition, the automation engaged at least 10 s before a SAGAT freeze (maximum time was 3.83 min) and disengaged at least 20 s before the next SAGAT freeze (maximum time was 4.33 min). For each scenario, three of the six SAGAT freezes occurred when automation was disengaged and three when automation was engaged.

Workload. As in Experiments 1 and 2, workload was measured by ATWIT every 60 s. In the adaptive condition, automation was engaged at least 15 s before the next ATWIT probe and disengaged at least 25 s before the next ATWIT probe. Automation removal occurred at 10:35 min in the last scenario, which was 15 s before the next ATWIT probe. There were 10 ATWIT probes before automation removal and 17 ATWIT probes after automation removal in every scenario. The NASA TLX was given after each scenario.

Trust. As in Experiments 1 and 2, the trust questionnaire was presented after each scenario for the automation conditions.

Procedure. The training on Day 1 was identical to that in Experiment 1 and 2, except that training was exclusive to condition. On Day 2, each participant completed three scenarios in their assigned condition. The order of scenarios was counterbalanced.

Results and Discussion

The means and 95% between-subjects confidence intervals for performance, SA, and subjective workload are presented in Table 5. In Experiment 3 we report between-subject effect sizes. The data are divided into time that automation was available (routine

Table 5
Descriptive Statistics for Performance, Situation Awareness, and Subjective Workload by Condition and State in Experiment 3

Automation	Classification		CPA		Dive		SAGAT	Rating	
	Hit	RT	Hit–FA	RT	Hit–FA	RT	Accuracy	ATWIT	NASA TLX
				Routine state					
None	.82 [.76, .88]	26.23 [22.25, 30.21]	.41 [.35, .48]	14.47 [11.28, 17.67]	.87 [.84, .90]	6.82 [5.96, 7.68]	.62 [.59, .65]	5.25 [4.92, 5.59]	61.91 [57.89, 65.92]
Static	.93 [.91, .94]	20.65 [18.75, 22.55]	.47 [.40, .55]	21.43 [17.17, 25.69]	.79 [.73, .84]	9.09 [8.22, 9.97]	.55 [.51, .59]	4.92 [4.49, 5.35]	62.25 [59.43, 65.07]
Adaptable	.87 [.83, .91]	22.98 [19.33, 26.63]	.42 [.33, .50]	20.77 [16.56, 24.99]	.76 [.71, .81]	8.38 [7.41, 9.34]	.57 [.53, .61]	5.36 [4.94, 5.78]	59.16 [54.03, 64.28]
Adaptive	.87 [.83, .92]	24.10 [20.41, 27.88]	.40 [.31, .49]	22.93 [19.30, 26.56]	.82 [.76, .87]	9.03 [7.83, 10.22]	.56 [.52, .60]	5.19 [4.82, 5.56]	64.78 [60.49, 69.07]
				Automation removal state					
None	.85 [.79, .92]	24.40 [20.00, 28.81]	.42 [.33, .51]	11.40 [8.87, 13.94]	.93 [.89, .97]	6.10 [5.20, 7.00]	.60 [.55, .65]	5.75 [5.36, 6.14]	65.56 [61.44, 69.67]
Static	.84 [.79, .90]	24.01 [20.18, 27.84]	.31 [.23, .40]	15.37 [11.34, 19.39]	.92 [.88, .96]	8.28 [6.65, 9.91]	.52 [.49, .56]	5.93 [5.46, 6.41]	68.76 [65.05, 72.47]
Adaptable	.78 [.71, .85]	24.20 [20.23, 28.17]	.37 [.27, .46]	18.30 [10.82, 25.77]	.85 [.78, .91]	7.14 [5.71, 8.56]	.54 [.49, .58]	5.96 [5.54, 6.39]	64.88 [59.89, 69.88]
Adaptive	.85 [.80, .91]	21.10 [16.20, 25.81]	.34 [.23, .44]	15.90 [10.56, 21.25]	.91 [.86, .95]	7.68 [6.20, 9.16]	.56 [.51, .60]	5.86 [5.33, 6.40]	68.24 [63.24, 73.24]

Note. The 95% between-subject confidence intervals are presented in parentheses. CPA = Closest Point of Approach; SAGAT = Situation Awareness Global Assessment Technique; ATWIT Air Traffic Workload Input Technique; NASA TLX = National Aeronautics and Space Administration Task Load Index; RT = Response Time in seconds; FA = False Alarm.

state; i.e., the first two and one third scenarios in each condition) and time that automation was unavailable (removal state; i.e., the last two thirds of the last scenario in each condition).

As in earlier experiments, we conducted statistical tests that directly evaluated our research questions by comparing the performance of the static automation condition to the no automation condition, the adaptable automation condition to the no automation condition, and the adaptive condition to the no automation condition, as a function of automation state (routine, removal). To do this we first ran mixed analyses of variance (ANOVAs), with automation condition as the between—subjects factor and automation state as the within-subjects factor (the resulting inferential statistics are summarized in Table 6). Significant interactions were followed by comparisons of simple effects conducted separately for the routine state and the automation removal state. Estimates of Cohen's *d* suggested we had a power of .70 to detect medium-to-large-size effects (Cohen, 1988).

In the following sections, we report several main effects of automation state. As shown in Tables 5 and 6, CPA performance was poorer, and workload as measured by the ATWIT and NASA TLX increased, during removal states compared with routine states. This may have occurred because for three of the four conditions the removal state represented the time that automation unexpectedly ceased. Furthermore, the automation removal periods always occurred during the last scenario where participants may have been most fatigued. It is also possible that participants may have been temporarily disrupted by the message at the point of automation removal (or the corresponding message in the no automation condition). However, it should be noted that we also found that CPA RT decreased and dive accuracy improved after automation was removed. For brevity, we do not further refer to the effects of automation state and instead focus on the main effects of condition (providing replications of Experiment 1 and 2) and the interactions.

Static automation versus no automation.

Task performance. A 2 (automation condition; no automation, static automation) × 2 (automation state; routine, removal) mixed ANOVA on classification accuracy revealed an interaction. Under routine states, participants made more accurate classification decisions when using static automation compared with no automation, $t(58) = 3.32$, $p = .002$, $d = 0.86$. In contrast, after automation removal, there was no difference in the accuracy of classification between the static automation and the no automation conditions ($t < 1$). For classification RT, there was also an interaction. During routine states, participants made faster contact classifications when using static automation compared no automation, $t(58) = -2.48$, $p = .02$, $d = -0.64$. In contrast, after automation removal, there was no difference in classification RT between the static automation condition and the no automation condition ($t < 1$). Overall, static automation significantly benefited classification accuracy and RT under routine states, replicating Experiments 1 and 2. Because classification accuracy/RT for the static condition did not degrade below that of the no automation condition after automation removal, it can be concluded that there was no return-to-manual deficit.

For CPA task accuracy, there was an interaction. The advantage to CPA task accuracy from using static automation compared with no automation (+6%) during routine states was reversed during automation removal states (−11%). However, there was no statis-

Table 6
Inferential Statistics for Performance, Situation awareness, and Subjective Workload by Condition and Automation State in Experiment 3

	Static automation vs. No automation (df = 1, 58)				Adaptable automation vs. No automation (df = 1, 57)				Adaptive automation vs. No automation (df = 1, 57)			
Dependent variable	Effect	F	p	η_p^2	Effect	F	p	η_p^2	Effect	F	p	η_p^2
Classification [Hit]	Condition	2.02	.16	.03	Condition	.09	.77	.002	Condition	.52	.47	.01
	State	3.48	.07	.06	State	3.87	.054	.06	State	.16	.69	.00
	Interaction	15.31	<.001*	.21	Interaction	14.88	<.001*	.21	Interaction	3.62	.06	.06
Classification [RT]	Condition	1.64	.21	.03	Condition	.45	.51	.008	Condition	.97	.33	.02
	State	.50	.48	.01	State	.07	.79	.001	State	5.63	.02*	.09
	Interaction	5.72	.02*	.09	Interaction	1.85	.18	.03	Interaction	.40	.53	.01
CPA [Hit–FA]	Condition	.20	.66	.003	Condition	.88	.35	.02	Condition	.72	.40	.01
	State	10.91	.002*	.16	State	.17	.68	.003	State	1.63	.21	.03
	Interaction	11.09	.001*	.17	Interaction	1.20	.28	.02	Interaction	2.06	.16	.04
CPA [RT]	Condition	10.77	.002*	.16	Condition	5.74	.02*	.09	Condition	11.00	.002*	.16
	State	6.26	.02*	.10	State	2.60	.11	.04	State	8.37	.005*	.13
	Interaction	.67	.42	.01	Interaction	.03	.86	.001	Interaction	1.29	.26	.02
Dive [Hit–FA]	Condition	3.08	.08	.05	Condition	11.28	<.001*	.17	Condition	2.32	.13	.04
	State	36.16	<.001*	.38	State	16.88	<.001*	.23	State	21.99	<.001*	.28
	Interaction	5.65	.02*	.09	Interaction	.58	.45	.01	Interaction	.88	.35	.02
Dive [RT]	Condition	12.25	.001*	.17	Condition	4.79	.03*	.08	Condition	8.74	.005*	.13
	State	3.06	.09	.05	State	5.24	.03*	.08	State	5.33	.03*	.09
	Interaction	.01	.92	.00	Interaction	.37	.55	.01	Interaction	.49	.49	.01
NASA TLX	Condition	.53	.47	.01	Condition	.38	.54	.01	Condition	.94	.34	.02
	State	38.42	<.001*	.39	State	10.03	.002*	.15	State	14.39	<.001*	.20
	Interaction	3.04	.09	.05	Interaction	.49	.49	.01	Interaction	.01	.92	.00
ATWIT	Condition	.08	.78	.001	Condition	.39	.54	.01	Condition	.01	.92	.00
	State	78.33	<.001*	.58	State	41.50	<.001*	.42	State	34.88	<.001*	.38
	Interaction	9.09	.004*	.14	Interaction	.37	.55	.01	Interaction	.76	.39	.01
SAGAT [Accuracy]	Condition	9.00	.004*	.13	Condition	4.66	.04*	.08	Condition	4.48	.04*	.07
	State	2.98	.09	.05	State	4.76	.03*	.08	State	.90	.35	.02
	Interaction	.02	.89	.00	Interaction	.37	.55	.01	Interaction	.31	.58	.01

Note. RT = Response Time in seconds; CPA = Closest Point of Approach; FA = False Alarms; NASA TLX = National Aeronautics and Space Administration Task Load Index; ATWIT Air Traffic Workload Input Technique; SAGAT = Situation Awareness Global Assessment Technique.
* $p < .05$.

tically significant difference in CPA task accuracy between the static automation condition and the no automation condition for either the routine state, $t(58) = 1.15$, $p = .26$, $d = 0.31$, or the removal state, $t(58) = -1.70$, $p = .09$, $d = -0.46$. For CPA task RT, there was a main effect of condition, with participants making slower CPA decisions when using static automation compared with no automation. Overall, the use of static automation did not significantly improve CPA task accuracy but did significantly impair CPA task RT. Although there was no statistically significant evidence of a return-to-manual deficit for the CPA task, CPA accuracy was 11% poorer for the static automation condition compared with the no automation condition after automation was removed.

For dive task accuracy, there was an interaction. Under the routine state participants had poorer dive task accuracy when using static automation compared with no automation, $t(58) = -2.59$, $p = .01$, $d = -0.68$, but after automation removal there was no difference in dive accuracy, $t < 1$. The dive task RT data indicated a main effect of condition, with participants taking longer to make correct dive decisions when using static automation compared with no automation. In summary, there was a significant cost to accuracy on the nonautomated dive task by the use of static automation during routine states, but this cost was eliminated after automation removal. There was a cost to dive task RT with the use of static automation during both routine and automation removal states.

Workload. For the ATWIT there was an interaction. The decrease in workload when using static automation (−6% workload) during the routine state reversed during the removal state (+3%). However, there was no statistical difference in ATWIT ratings when using static automation compared with no automation during either routine states, $t(58) = -1.25$, $p = .22$, $d = -0.32$, or after automation removal ($t < 1$). The NASA TLX scores of the first and second routine scenarios were averaged for each participant. The NASA TLX score after the third scenario was taken as the participant's subjective workload under the automation removal state. There was no main effect of condition and no interaction for the NASA TLX. In summary, in contrast to Experiments 1 and 2, we found no evidence of a significant reduction in subjective workload with the use of static automation. As discussed later, these results may stem from the change from the within-subject designs used in Experiment 1 and 2 to the between-subjects design used in Experiment 3.

SA. For SAGAT accuracy, there was a main effect of condition, with poorer accuracy of responses to SAGAT queries when using static automation compared with no automation. Thus, replicating Experiment 2, there was a significant cost to SA with the use of static automation under routine states, and this cost to SA did not diminish after automation was removed and participants resumed manual control.

Adaptable automation versus no automation. Consistent with Experiments 1 and 2, and as shown in Figure 5, automation usage coincided with the task-load peaks (see Figure 5). During the first scenario participants used 13 min of automation on average (47% of the available time). In the second scenario participants used 18.35 min (67%). For the third scenario, participants used 7.49 min of the 10.58 min of automation available (71%).

Task performance. A 2 (condition; no automation, adaptable automation) × 2 (state; routine, automation removal) mixed ANOVA on classification accuracy revealed an interaction. The advantage to classification accuracy with static automation (+5%) during routine states was reversed during automation removal states (−7%). However, there was no significant difference in classification accuracy between the adaptable condition and the no automation condition for the routine state, $t(57) = 1.31$, $p = .20$, $d = 0.34$, or after automation removal, $t(57) = -1.48$, $p = .14$, $d = -0.38$. Although there was no statistically significant evidence of a return-to-manual deficit for the classification task, classification accuracy was 7% poorer for the adaptable automation condition compared with the no automation condition after automation removal. For classification RT, there was no main effect or interaction.

For CPA task accuracy, there was no main effect of condition or interaction. CPA task RT showed a main effect of condition, $F(1, 57) = 5.74$, $p = .02$, $\eta_p^2 = .09$, with participants making slower CPA decisions when using adaptable automation compared with no automation.

Overall, in contrast to Experiment 2, adaptable automation did not significantly improve classification accuracy or RT. Additionally, in contrast to Experiments 1 and 2, adaptable automation did not improve CPA task accuracy. In addition, adaptable automation significantly slowed CPA RT, even after participants resumed manual control.

For dive task accuracy, there was a main effect of condition. Dive accuracy was poorer for the adaptable automation condition compared with the no automation condition, replicating Experiment 2. Dive task RT also indicated a main effect of condition. Participants took longer to make correct dive decisions when using adaptable automation compared with no automation. In summary, there was a significant cost to nonautomated dive task accuracy and RT associated with using adaptable automation during routine states, and these costs were not reduced after automation was removed.

Workload. For the ATWIT and the NASA TLX there were no main effects of condition or interactions.

SA. For SAGAT there was a main effect of condition. SAGAT accuracy was poorer when participants used adaptable automation compared with no automation. Thus, in contrast to Experiments 1 and 2, there was a cost to SA arising from adaptable automation that did not decrease after automation was removed.

Adaptive automation versus no automation.

Task performance. A 2 (condition; no automation, adaptive automation) × 2 (state; routine, automation removal) mixed ANOVA on classification accuracy revealed an interaction that approached significance. For classification RT, there was no main effect of condition and no interaction.

For CPA accuracy there was no main effect of condition or interaction. For CPA RT there was a main effect for condition. Participants were slower at marking CPAs in the adaptive condition compared with the no automation condition. In summary, the use of adaptive automation provided no benefit to contact classification, and significantly slowed CPA RT.

For dive task accuracy, there was no main effect of condition and no interaction. For dive RT we found a main effect of condition, with participants taking longer to make dive task decisions in the adaptive condition compared with the no automation condition.

Workload. For ATWIT and the NASA TLX there were no main effects or interactions.

SA. For SAGAT there was a main effect of condition, with SAGAT accuracy poorer for the adaptive automation condition compared with the no automation condition.

Trust. Trust in automation under routine states was calculated from the mean of history-based trust scores of the first two scenarios (static $M = 4.11$; 95% CI [3.99, 4.23]; adaptable $M = 4.07$, 95% CI [3.89, 4.26]; adaptive $M = 4.12$, 95% CI [3.93, 4.31]). Trust in automation under the automation removal state was calculated from history-based trust scores of the last scenario in which automation was removed (static $M = 4.38$, 95% CI [4.17, 4.59]; adaptable $M = 4.07$, 95% CI [3.91, 4.23]; adaptive $M = 4.14$, 95% CI [3.90, 4.39]). A 3 (automation condition; static, adaptable, adaptive) × 2 (automation state; routine, removal) mixed ANOVA revealed no main effects or interactions (smallest $p = .22$). Trust was likely not impacted because automation removal in the third scenario was not interpreted by the participant as a failure of the automation (which might have led participants to doubt the automation reliability), but rather as a deliberate action due to a perceived threat to the submarine.

Meta-analysis. Because there was some inconsistency in the impact of automation on performance, SA, and workload under routine states across the three experiments, and because Experiment 3 was slightly underpowered, to clarify our findings we conducted a meta-analysis across the experiments using the procedures outlined by Cumming (2012). We conducted a meta-analysis on the difference scores between the static automation and no automation conditions, and between the adaptable automation and no automation conditions, for all dependent variables that were common across the three experiments under the routine state. For each meta-analysis, the three experiments were weighted according to the inverse of the variance of their effect sizes (i.e., a study with a shorter confidence interval would have larger meta-analytic weight). We applied the more conservative random effect model (as opposed to a fixed-effects model) to account for the heterogeneity of the three experiments as well as for the variance in the individual effect sizes. The forest plots showing mean effect sizes and 95% CI's (if a CI captures zero, then we cannot say the effect differs significantly from zero) are presented in Figures 6 and 7 and are summarized and discussed where appropriate in the General Discussion section.

General Discussion

In three experiments, we simulated submarine track management to examine the extent to which automation could improve performance while allowing the operator to maintain SA, maintain performance on an interrelated nonautomated task, and regain manual control of automated tasks. To encourage participants to remain engaged when using static automation, we designed the automation only to support information acquisition and analysis,

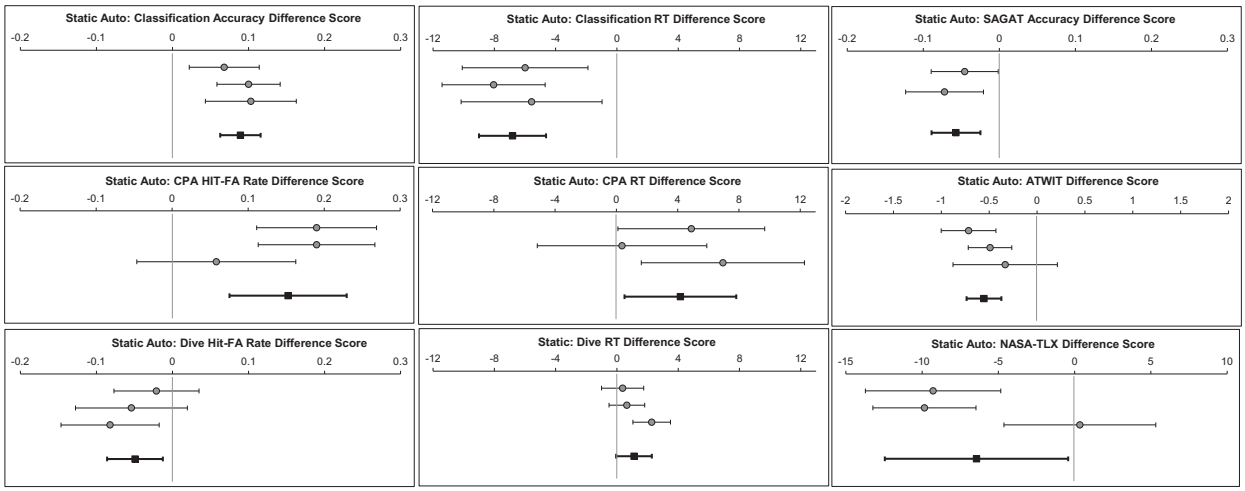

Figure 6. Meta-analysis results for the static automation condition. The gray dots represent the mean differences between the static automation condition and the no automation condition in Experiments 1, 2, and 3 (presented in that order), and the black squares represent the meta-analysis data point. 95% confidence intervals are presented. Note that for Situation Awareness Global Assessment Technique accuracy the gray dots represent Experiment 2 and 3.

the DOA boundary at or below which costs to SA and return-to-manual performance were considered to be less likely (Onnasch et al., 2014). This DOA still required participants to make task decisions and execute associated actions. We also examined the impact of the periodic reallocation of automated tasks to manual control either through operator triggered automation (adaptable automation) or task-load triggers (adaptive automation). Participants performed two tasks that were supported by automation (classification and CPA), and one task that was not supported by automation (dive) but required integration of the same raw information as the tasks supported by automation.

The Benefits and Costs of Using Static Automation

The meta-analyses (see Figure 6) indicate that the low degree of static automation provided consistent benefits to classification task accuracy, classification task RT, and CPA task accuracy. However, as shown in Figure 6, the use of static automation slowed

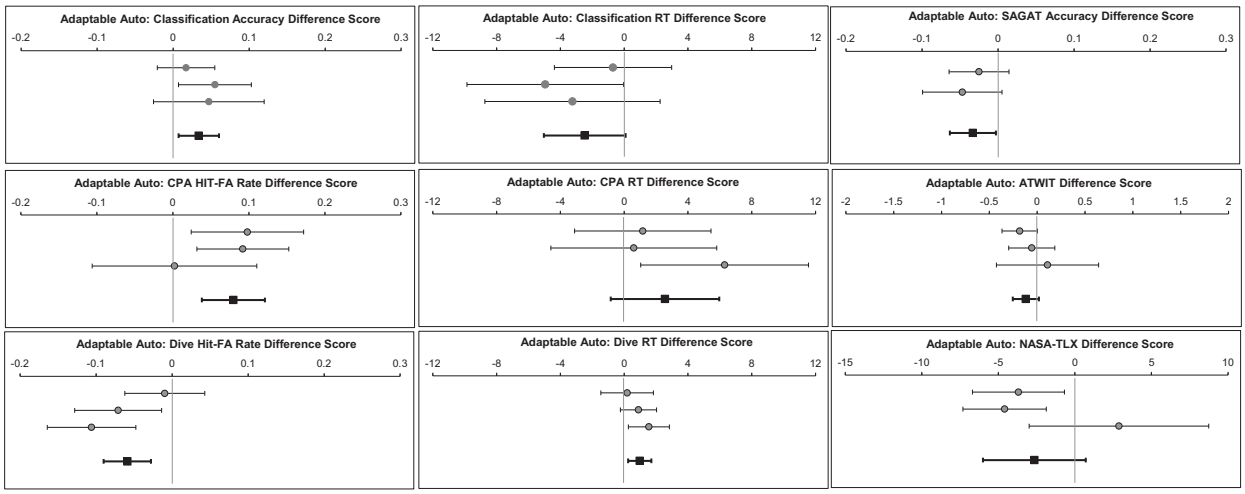

Figure 7. Meta-analysis results for the adaptable automation condition. The gray dots represent the mean differences between the adaptable automation condition and the no automation condition in Experiments 1, 2, and 3 (presented in that order), and the black squares represent the meta-analysis data point. 95% confidence intervals are presented. Note that for Situation Awareness Global Assessment Technique accuracy the gray dots represent Experiment 2 and 3.

CPA task RT under routine conditions. This means that participants took longer to mark a CPA when using static automation compared with no automation. This speed–accuracy trade-off in CPA performance for the static automation condition might reflect the need, or strategic shift, to retrospectively assess the contact track history to make CPA decisions instead of continually monitoring for CPAs. This strategy was viable because if the CPA was missed the participant could always return later to mark it on the waterfall display by looking at the track history to find when the contact changed course away from Ownship. When automation was removed in the last scenario of Experiment 3, benefits to the classification and CPA task were eliminated, but we did not find statistically significant return-to-manual deficits.

The meta-analyses (see Figure 6) indicate that the use of static automation was effective in reducing participant's ratings of workload on the ATWIT and NASA-TLX. However, there was considerable variability in the extent of this reduction across experiments. In Experiments 1 and 2, with the use of the within-subject design, static automation reduced subjective workload, and was the type of automation preferred by participants. However, as shown in Figure 6, static automation provided no benefit to subjective workload in Experiment 3 when the between-subjects design was used (particularly for the NASA-TLX). As indicated in Figure 7, this pattern of subjective workload results across experiments also occurred for the adaptable automation condition when compared with no automation.

The subjective workload data highlight the importance, when interpreting findings, of considering the experimental design that was used. Evidence in the literature suggests that perceptions of mental workload reflect individuals' metacognitive judgments regarding the availability of task relevant information in working memory (Estes, 2015; Yeh & Wickens, 1988). Participants in Experiments 1 and 2 were able to compare the workload changes that they were experiencing across the different within-subject conditions. Of course, such comparisons were not possible for participants in the between-subjects design in Experiment 3, who were exposed exclusively to one type of condition. We believe that a within-subject design is more ecologically valid in the current context (see Greenwald, 1976) because experts working in complex systems, such as submarine track management and air traffic control, often transition between different working conditions, which includes changes in automation availability or suitability for different tasks as a function of the operational context in which tasks are embedded. Within-subject designs also provide statistical efficiency by removing subject variance from error terms used to test treatment effects. A drawback is that within-subject designs can be impacted by practice and asymmetric transfer (Poulton, 1982), and these possibilities need to be evaluated (although note that we found no evidence for order effects in the current studies).

Overall, the results demonstrate that static automation that provides support to information acquisition and analysis (lower level DOA) can deliver workload and performance benefits. However, costs were also clearly apparent. The meta-analyses (see Figure 6) indicate that the use of static automation resulted in costs to SA and accuracy on the interdependent nonautomated dive task. In addition, participants were not able to regain SA after they returned to manual control in Experiment 3. The absence of a SA deficit with the use of static automation in Experiment 1 is likely because of the use of SPAM, which has been shown to be less sensitive than SAGAT for measuring SA in simulated submarine track management (Loft, Bowden, et al., 2015). It is likely that SAGAT is more sensitive because it measures the consciously reportable knowledge being held working memory.

As shown in Figure 6, the meta-analyses indicate that dive task decisions were less accurate, and slower, with the use of static automation compared with no automation. This pattern was replicated in the adaptable condition (see Figure 6), and partly replicated (slowed RT only) for the adaptive condition in Experiment 3. It would have been reasonable to predict that the observed reduction in subjective workload from the provision of static automation would provide participants with the spare cognitive capacity to better manage the nonautomated dive task (e.g., Loft et al., 2013; Rovira et al., 2007). However, in our study, the raw information required for the dive task (assessing the relative location and heading of contacts to each other and to Ownship) overlapped with the information required to complete the classification and CPA tasks. When the classification and CPA tasks were automated, participants presumably did not monitor contact location and heading as closely as they would have without automation (Parasuraman & Manzey, 2010; Wickens et al., 2015). It is likely that this diminished attention to contact location and heading impaired participants' performance on the interdependent nonautomated dive task. To directly test this explanation, future researchers could manipulate whether the nonautomated task shares information processing requirements with automated tasks (as was the case in the current study) or has information processing requirements that are independent of automated tasks. We would expect to find that performance on an independent non automated task would be the same or better for those participants that use automation, compared with those that use no automation. This is because, at least for static automation, we observed consistent reductions in workload, indicating that static automation provided participants with spare cognitive capacity.

In conclusion, we expected that because the static automation functioned below the critical information-processing boundary (Onnasch et al., 2014), it would be less likely to result in costs to SA and nonautomated task performance. This was clearly not the case. Although static automation benefited performance and workload, participant SA was diminished and performance on nonautomated task performance was degraded. These findings support the more general view that static automation may not be optimal for dynamic task environments, regardless of the level or information processing stage at which it is implemented (e.g., Kaber & Riley, 1999; Parasuraman, Cosenzo, & De Visser, 2009).

The Benefits and Costs of Using Adaptable Automation

To our knowledge, this is the first study to have directly compared adaptable to no automation conditions, and is the first to have measured the costs of adaptable automation on SA, return-to-manual performance, and performance on an interdependent nonautomated task. We found that adaptable automation was more likely to be engaged when task load and associated subjective workload increased, and disengaged when demands receded, demonstrating that participants could strategically control their automation usage. Originally, we had some concerns that participants

might overuse automation, but this was clearly not the case. In fact, in Experiments 2 and 3 in which no automation time limits were imposed, participants only used automation for about half the scenario time.

The meta-analyses (see Figure 7) indicate that the use of adaptable automation provided some benefits to performance, improving classification accuracy and CPA accuracy, and marginally improving classification RT. There were also marginal reductions to subjective workload as a result of using adaptable automation. However, as shown in Figure 7, we also found reliable costs associated with adaptable automation. Both accuracy and RT on the dive task were degraded with the use of adaptable automation. Furthermore, unlike the static automation condition, where dive task performance recovered when participants resumed manual control of the classification and CPA task, dive task performance remained impaired after automation removal. In addition, there were costs to SA as a result of using adaptable automation. Thus, our expectation that the process of deciding when to use automation would help participants maintain SA was clearly off the mark. Overall, not only did we find that adaptable automation provided only modest improvements to performance and workload, we also found that it produced substantial costs to SA and to nonautomated task performance.

There are several potential explanations for these poor outcomes of adaptable automation. There may have been a cognitive load associated with the ongoing process of deciding when to engage and disengage automation (Bailey et al., 2006; Harris, Hancock, & Arthur, 1993; Kidwell et al., 2012), which might have been regarded by participants as an "extra task." In line with this, participants in Experiment 1 and 2 reported that they preferred using static automation to adaptable automation. Participants may also have taken some time to recover SA and the associated control of tasks each time they disengaged automation because they needed to switch attention to additional sources of display information (task switching effects; Rogers & Monsell, 1995). Finally, participants may not have been able to monitor their task demands or performance effectively to accurately assess the need for automation (Flavell, 1979; Osman, 2010). In summary, we found little support for our contention that the human operator might be best placed to engage and disengage automation.

The Costs of Using Adaptive Automation

In Experiment 3, there were no benefits to performance or workload from using adaptive automation that was engaged and disengaged based on task load. In fact, CPA decisions were slower with adaptive automation, without any corresponding improvement in CPA task accuracy. Furthermore, there were costs to using adaptive automation. Participants were less accurate and slower to make dive task decisions, and SA was degraded, during both routine states and after automation removal. In sum, the use of adaptive automation was costly, with no concurrent benefits. We did not, however, observe any significant return-to-manual deficits. Overall these results are inconsistent with those reported by de Visser and Parasuraman (2011), who found adaptive automation reduced workload and improved SA. However, note that de Visser and Parasuraman measured subjective SA, rather than using an objective SA measure, and only found a reduction in workload when the adaptive condition was compared with static automation.

Similar to adaptable automation, there may have been a recovery time cost each time automation was disengaged. If so, unlike adaptable automation, this would have been further compounded by the fact that participants were provided with no warning that automation was about to be disengaged. More generally, if adaptive automation was perceived by participants to have engaged or disengaged at inopportune times, this might have created 'automation surprises' (Sarter, Woods, & Billings, 1997), and participants might have used additional cognitive resources to evaluate how the introduction of automation could impact their assessment of the tactical picture (e.g., looking to see if the timing lines and track history matched their own expectations). As a result, the participants' workflow in developing and maintaining a mental model of the situation may have been interrupted, leading to temporary disorientation. A further potential drawback of adaptive automation is that it may reduce the extent of operator engagement in the scenario because they were not deciding when to use automation. It will be important for future research to examine how variations in display design can potentially mitigate potential automation engagement reorientation costs or increase operator engagement (Ballas, Heitmeyer, & Perez, 1992; Kieras, Meyer, & Ballas, 2001). Finally, another potential issue is that the contact count at which we engaged and disengaged automation may not have been optimal.

Practical Implications

Technological innovation and its potential economic benefits mean that humans will continue to deal with ever more highly automated systems. A key challenge concerns how to design automation technology to handle increasingly complex information, while ensuring that key information is translated and communicated to human decision makers in a manner that allows them to maintain SA and take control over automated tasks if required (Hancock et al., 2013). Successfully addressing this challenge is a current priority for defense departments around the world as part of their attempts to design new information handling systems or to evaluate off-the-shelf automation technologies for military command and control systems (Endsley, 2015). Although the current studies used simulations of submarine track management, these issues, and the outcomes of the current work, are also potentially relevant to any work contexts that require humans to monitor demanding perceptual displays, such as air traffic control and unmanned vehicle control.

On the basis of the results of the meta-analysis presented by Onnasch et al. (2014), we used a relatively low DOA that ensured that the operator still made task decisions. The results indicate that a relatively low degree of reliable static automation can consistently improve performance and reduce subjective workload. However, we found that static automation also degraded participant SA. This is a problem because SA represents the quality of an operator's understanding of the task and his or her ability to anticipate the future consequences of task events or their own actions, and SA is a reliable indicator of the capacity of operators to respond to sudden increases in workload/unexpected task events, or to regain manual control of previously automated tasks (Vu & Chiappe, 2015).

Indeed, in industrial settings, automating tasks can improve operator and system performance, but accidents have occurred

because operators have not been adequately prepared to regain manual control. In addition, the current dive task data suggest that because of the loss of SA, operators in complex work systems may find it difficult to maintain adequate performance on nonautomated tasks that share information processing requirements with currently automated tasks. Thus, whether the reduced workload from automation will allow the operator to more effectively manage other nonautomated tasks depends on the nature of that nonautomated task. The implications for automation design are clear. Designers should consider task information overlaps across all subtasks of an information management role, to avoid unintentionally inducing operator complacency for nonautomated tasks.

Although the concept (and promise) of adaptable and adaptive automation has a long history (Rouse, 1988), empirical evidence regarding the effectiveness of adaptable and adaptive automation is limited. Few industries have implemented adaptable or adaptive automation. A large part of the reason for this is that such automation is expensive to design and implement. For industry to seriously consider adaptable and adaptive automation as an alternative to static automation, there needs to be a systematic demonstration that adaptable or adaptive automation can maximize performance (and minimize workload), while ensuring that operators maintain sufficient SA to regain manual control when compared with conditions were no automation is being used. We failed to find evidence for these evaluation criteria in our current simulations of submarine track management using either adaptable or adaptive automation. Of course, there are alternative triggers potentially suited to a submarine control room, such as operator performance (e.g., Calhoun et al., 2011), secondary task workload (Kaber & Riley, 1999), or operator physiology (Wilson & Russell, 2007). We also suspect that the schedule for engaging and disengaging automation is most likely to be optimal if it takes into account how moment-to-moment fluctuations in task load/workload relate to variation in performance within-individuals (Mracek, Arsenault, Day, Hardy, III, & Terry, 2014; Wilson & Russell, 2007). Nevertheless, the current data indicate that allowing the operator to decide when to engage and disengage automation, or triggering automation based on task-load, are not likely to produce sound outcomes.

The design of our submarine track management simulation was informed by observations of real submarine combat systems and by expert submariner opinion. Thus, our experiments have external validity (psychological fidelity) because they represent a prototypical example of a work context that requires operators to monitor demanding perceptual displays. The importance of using representative experimental contexts has long been advocated by the ecological rationality approach to psychology (Brunswick, 1943; Simon, 1956). Nonetheless, we certainly do not dispute the potential problems in generalizing from relatively inexperienced participants to field operations involving experienced operators. There are undoubtedly differences in domain-specific cognitive skill, and in motivation, between experts and students. The expert is also likely to be part of an established team within which communication might provide additional informational cues.

That said, there is also evidence that our results with inexperienced participants can validly inform practical issues in operational contexts. In a recent study assessing SA and performance, we found reasonably consistent results across student participants using the current simulated track management task and follow-up studies with expert submariners using real submarine combat systems (Loft et al., 2016). Further, the meta-analysis conducted by Onnasch et al. (2014) found that the benefits and costs of automation were not moderated by experience; that is, the automation use benefit/costs trade-off was just as statistically likely to occur for experts as it was for more naïve participants. On this basis, it would be reasonable to expect to find similar effects of automation on the performance, workload, and SA of experts as those that have been reported in the current study. These points notwithstanding, it will be crucial for future research to continue to examine potential expert-novice differences in real military and industrial settings so that we can develop a further understanding how experts and novices differ (or not) in the deployment of attention and/or in their use of strategy in approaching tasks that include different types of automation.

In conclusion, the design of automation for supporting performance in complex dynamic work systems is a diverse and challenging problem. It is essential that designs of these automation tools be based on a thorough analysis of human cognition and decision-making processes. Clearly, further basic strategic research is urgently required to discover how best to adapt automation to keep human operators more cognitively engaged with their tasks.

References

Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., & Scott, L. A. (2006). Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors, 48,* 693–709. http://dx.doi.org/10.1518/001872006779166280

Ballas, J. A., Heitmeyer, C. L., & Perez, M. A. (1992). Evaluating two aspects of direct manipulation in advanced cockpits. *CHI '92 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 127–134). New York, NY: ACM.

Bindewald, J. M., Miller, M. E., & Peterson, G. L. (2014). A function-to-task process model for adaptive automation system design. *International Journal of Human-Computer Studies, 72,* 822–834. http://dx.doi.org/10.1016/j.ijhcs.2014.07.004

Brunswick, E. (1943). Organismic achievement and environmental probabilities. *Psychological Review, 50,* 255–272. http://dx.doi.org/10.1037/h0060889

Calhoun, G. L., Ward, V. B. R., & Ruff, H. A. (2011). Performance-based adaptive automation for supervisory control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55,* 2059–2063. http://dx.doi.org/10.1177/1071181311551429

Chen, S., Loft, S., Huf, S., & Visser, T. (2014). Static and adaptable automation in simulated submarine track management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58,* 2280–2284. http://dx.doi.org/10.1177/1541931214581475

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83,* 596–610. http://dx.doi.org/10.1080/01621459.1988.10478639

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, England: Routledge.

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making, 5,* 209–231. http://dx.doi.org/10.1177/1555343411410160

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of

SA. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness* (pp. 137–154). Hampshire, UK: Ashgate.

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. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64. http://dx.doi.org/10.1518/001872095779049543

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65–84. http://dx.doi.org/10.1518/001872095779049499

Endsley, M. R. (2015). *Autonomous Horizons: System Autonomy in the Air Force–A Path to the Future* (Volume I: Human Autonomy Teaming, pp. 1–24). Washington, DC: US Department of the Air Force.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control of automation. *Human Factors, 37,* 381–394. http://dx.doi.org/10.1518/001872095779064555

Estes, S. (2015). The workload curve: Subjective mental workload. *Human Factors, 57,* 1174–1187. http://dx.doi.org/10.1177/0018720815592752

Farrell, S., & Lewandowsky, S. (2000). A connectionist model of complacency and adaptive recovery under automation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 395–410. http://dx.doi.org/10.1037/0278-7393.26.2.395

Flavell, J. H. (1979). Metacognition and cognitive monitoring. A new area of cognitive-development inquiry. *American Psychologist, 34,* 906–911. http://dx.doi.org/10.1037/0003-066X.34.10.906

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83,* 314–320. http://dx.doi.org/10.1037/0033-2909.83.2.314

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. *Ergonomics in Design, 21,* 9–14. http://dx.doi.org/10.1177/1064804613477099

Harris, C. W., Hancock, P. A., & Arthur, E. J. (1993). The effect of taskload projection on automation use, performance, and workload. *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 25–30). Columbus, OH: Naval Air Warfare Center-Aircraft Division.

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.

Kaber, D. B., & Riley, J. M. (1999). Adaptive automation of a dynamic control task based on secondary task workload measurement. *International Journal of Cognitive Ergonomics, 3,* 169–187. http://dx.doi.org/10.1207/s15327566ijce0303_1

Kidwell, B., Calhoun, G. L., Ruff, H. A., & Parasuraman, R. (2012). Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56,* 428–432. http://dx.doi.org/10.1177/1071181312561096

Kieras, D., Meyer, D., & Ballas, J. A. (2001). Towards demystification of direct manipulation: Cognitive modeling charts the gulf of execution. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 128–135). New York, NY: ACM.

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.

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. http://dx.doi.org/10.1177/0018720814522830

Loft, S., Bowden, V., Braithwaite, J., Morrell, D. B., Huf, S., & Durso, F. T. (2015). Situation awareness measures for simulated submarine track management. *Human Factors, 57,* 298–310. http://dx.doi.org/10.1177/0018720814545515

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. http://dx.doi.org/10.1177/0018720816652754

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

Loft, S., Smith, R. E., & Bhaskara, A. (2011). Prospective memory in an air traffic control simulation: External aids that signal when to act. *Journal of Experimental Psychology: Applied, 17,* 60–70. http://dx.doi.org/10.1037/a0022845

Loft, S., Smith, R. E., & Remington, R. W. (2013). Minimizing the disruptive effects of prospective memory in simulated air traffic control. *Journal of Experimental Psychology: Applied, 19,* 254–265. http://dx.doi.org/10.1037/a0034141

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. http://dx.doi.org/10.1177/1555343411433844

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

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

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105–125. http://dx.doi.org/10.1037/1082-989X.7.1.105

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy, J. H., III, & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human Factors, 56,* 1401–1413. http://dx.doi.org/10.1177/0018720814533964

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. http://dx.doi.org/10.1177/0018720813501549

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin, 136,* 65–86. http://dx.doi.org/10.1037/a0017815

Parasuraman, R., Cosenzo, K. A., & De Visser, E. (2009). Adaptive automation for human supervision of multiple uninhabited vehicles: Effects on change detection, situation awareness, and mental workload. *Military Psychology, 21,* 270–297. http://dx.doi.org/10.1080/08995600902768800

Parasuraman, R., Galster, S., Squire, P., Furukawa, H., & Miller, C. A. (2005). A flexible delegation-type interface enhances system performance in human supervision of multiple robots: Empirical studies with RoboFlag. *IEEE Systems. Man and Cybernetics—Part A: Systems and Humans, 35,* 481–493. http://dx.doi.org/10.1109/TSMCA.2005.850598

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52,* 381–410. http://dx.doi.org/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. http://dx.doi.org/10.1207/s15327108ijap0301_1