

Heuristic Automation for Decluttering Tactical Displays

Mark St. John, Harvey S. Smallman, and Daniel I. Manes, Pacific Science & Engineering Group, San Diego, California, and Bela A. Feher and Jeffrey G. Morrison, Space and Naval Warfare System Center, San Diego, California

Tactical displays can quickly become cluttered with large numbers of symbols that can compromise effective monitoring. Here, we studied how heuristic automation can aid users by intelligently “decluttering” the display. In a realistic simulated naval air defense task, 27 experienced U.S. Navy users monitored a cluttered airspace and executed defensive responses against significant threats. An algorithm continuously evaluated aircraft for their levels of threat and decluttered the less threatening ones by dimming their symbols. Users appropriately distrusted and spot-checked the automation’s assessments, and decluttering had very little effect on which aircraft were judged as significantly threatening. Nonetheless, decluttering improved the timeliness of responses to threatening aircraft by 25% as compared with a baseline display with no decluttering; it was especially beneficial for threats in more peripheral locations, and 25 of 27 participants preferred decluttering. Heuristic automation, when properly designed to guide users’ attention by decluttering less important objects, may prove valuable in many cluttered monitoring situations, including air traffic management, crisis team management, and tactical situation awareness in general.

INTRODUCTION

Clutter can become a serious problem for users monitoring situation displays. For example, in naval air defense, users must monitor airspaces to find threatening aircraft. These airspaces are frequently in busy environments near land and contain multiple commercial air lanes and other air traffic. Clutter increases search times by increasing the number of objects that must be sifted through or searched to find objects of interest (e.g., Treisman & Gelade, 1980). Clutter also increases the chance for “change blindness,” the chronic human inability to detect changes occurring in a scene when attention is focused elsewhere (Rensink, 2002). These problems can result in reduced situation awareness and delayed response times to critical events.

A common method for reducing clutter and promoting situation awareness is to identify important objects and then mark or highlight them in some manner. Highlighting, when the

identification process is reliable, allows users to focus on a subset of objects and thereby effectively reduces the number of objects that must be sifted through or monitored. For example, in a search through a matrix of words, Fisher, Coury, Tengs, and Duffy (1989) found that highlighting a subset of words improved response time, even when the highlighting was less than completely reliable. In a visual search task for symbols on a tactical map display, Van Orden, DiVita, and Shim (1993) found that highlighting a category of symbols improved response time. In an augmented reality search task, Yeh and Wickens (2001b) found that highlighting targets improved response time. However, one downside of highlighting is that because it is such an effective form of cuing, it can impede the detection of important objects that are mistakenly left unhighlighted (and hence uncued) when the automation is imperfect or the situation is uncertain (e.g., Baddeley, 1972; Posner, 1980; Yeh & Wickens, 2001b).

A related method for reducing clutter is to identify less important objects and then declutter them from the display by making them less visually salient in some manner. This method also reduces the effective search space by eliminating some objects from the search set. In several studies of visual search for targets in tactical map displays, researchers have shown that users appreciate and benefit from the decluttering of irrelevant categories of symbols (Johnson, Liao, & Granada, 2002; Nugent, 1996; Osga & Keating, 1994; Schultz, Nichols, & Curran, 1985; Yeh & Wickens, 2001a).

A number of methods have been used to declutter objects by reducing their visual salience, including size reduction, dimming, turning symbols into dots, and even complete removal. Ideally, a good declutter method should visually segregate important from less important objects but with minimal disruption to the information content of the symbols. For example, in a visual search task for target symbols on a cluttered display, St. John, Feher, and Morrison (2002) found that simply dimming irrelevant symbols to one third of their initial luminance (thereby reducing their contrast against a dark background) supported easy segregation but without removing any identifying information.

An often overlooked issue, which we address here, is how the highlighted or decluttered objects are identified in the first place. In most experimental studies, the identification function is simply assumed to exist, but it is left unspecified. In applied tactical domains such as air defense, the identification functions are typically simple classification rules, such as all friendly aircraft or all aircraft with altitudes over 25,000 feet (standard U.S. Navy practice). Although attractive because of their simplicity, these rules often fail to meet the needs of sophisticated users because they do not align with the categories of most interest to these users.

A more sophisticated approach is to define meaningful categories of objects and then use these categories as the basis for decluttering. For example, in air defense, rules can be defined to identify commercial versus military aircraft, and then the commercial aircraft can be decluttered. Of course, such rules are necessarily heuristic and are bound to miscategorize aircraft on occasion. Moreover, the identification

function of most interest to tactical users is not the type of aircraft, *per se*, but its level of threat to own ship or other assets. Navy users monitor tactical situations in order to assess threats and then execute responses in order to minimize them. Threat, however, is an ill-defined and complex function of many aircraft attributes and requires years of experience to train (Kaempf, Wolf, & Miller, 1993; Liebhaber, Kobus, & Feher, 2002; Marshall, Christensen, & McAllister, 1996; Morrison, Kelly, & Hutchins, 1996).

Development of reliable automated threat assessment algorithms has long been a goal for aiding situation awareness generally, and air defense in particular. Unfortunately there are several challenges to producing reliable threat evaluation automation. First, the problem can grow extremely complex in attempting to account for all possible variables, including aircraft kinematics, coordinated aircraft behaviors (the big picture), intelligence information, and situational factors such as the geopolitical context. Second, the problem can suffer from ambiguity because important data may be unknown or unknowable. For example, aircraft identity is often based on electronic emissions that may not be detectable or that may have multiple interpretations; ultimately, the intent of an aircraft can never be established with certainty.

Third, expert decision makers frequently disagree about the threat of individual aircraft. For example, Marshall et al. (1996) found that all six of the teams they studied agreed on the interest level of only 41% of the aircraft. Consequently, an automated algorithm can never perfectly match the threat ratings of every user. Fourth, well-known problems of automation trust, complacency, and confirmation bias (e.g., Parasuraman & Riley, 1997) can undermine the effective use of automation and lead to disastrous consequences. On one hand, for example, a user might monitor only those aircraft indicated as threats by the automation, or if the automation missed a threat, the user might be significantly delayed in noticing it. If the automation mistakenly overrated the threat of an aircraft, a user might treat it more aggressively than necessary. On the other hand, distrust of automation might actually increase workload by driving users to increase their monitoring of lower threat aircraft.

Our approach is to treat the automation and

the user as a “mixed initiative” system that combines “heuristic automation” that is known to be imperfect with engaged, knowledgeable users who use the automation as a guide but ultimately rely on their own best judgment. According to this design strategy (e.g., Parasuraman & Riley, 1997, pp. 244, 249; St. John & Manes, 2002; St. John, Oonk, & Osga, 2000), users are taught how and where the automation is likely to be trustworthy or make errors, and they verify the automation accordingly. This design strategy fits well with what are termed “low levels of automation” (e.g., Kaber & Endsley, 2004; Parasuraman, Sheridan, & Wickens, 2000), which might involve merely identifying alternative solutions rather than recommending a single best solution or executing a solution unless countermanded by the user. For example, in a visual search task, St. John and Manes (2002) used heuristic automation in the form of an imperfect target detection tool to make a rough first cut at identifying the likely locations of hidden targets. Users then exploited this information to guide their own searches. This approach led to a 23% improvement in search times, even when the automation was only 70% reliable. In a dual-task paradigm, Sorkin, Kantowitz, and Kantowitz (1988) used a “likelihood alarm display” to indicate the likelihood of a signal occurring in the secondary task. Users exploited the likelihood information to decide how carefully to attend to the secondary task. In both studies, knowledgeable users exploited the information provided by imperfect, heuristic automation to guide their attention.

We applied this heuristic automation design strategy to air defense. First, a heuristic threat assessment algorithm evaluated all aircraft every second as they moved about the display by weighing several aircraft attributes and computing a “threat score.” Then, lower scoring, less threatening aircraft were decluttered by reducing the salience of their symbols on the display. In this way, the decluttered aircraft would not distract from the higher threat aircraft, yet they would remain available for inspection. We predicted that users would be able to exploit the information provided by the automation to focus the majority of their attention on the fully visible threatening aircraft while periodically scanning the entire display to verify the automation’s

assessments of the decluttered aircraft. Situation awareness would be enhanced and responses speeded because significant threats would be clearly visible. Decluttering might be especially useful for facilitating the early detection of significant threats at longer ranges from own ship. Time freed up from searching the cluttered display could be used to verify decluttered aircraft opportunistically on the chance that the heuristic algorithm decluttered an aircraft in error. Thus the potential costs of automation-induced misses would be minimized.

The current experiment tests these predictions in a scenario-based, quasi-realistic air defense task with experienced naval users. Our goal was to assess whether heuristic automation in combination with decluttering could facilitate performance and garner user acceptance within the naturalistic constraints of a real task with experienced users in realistic scenarios. Accordingly, participants performed the normal tasks involved in air defense – namely, monitoring an airspace, evaluating aircraft, and responding to the “significantly threatening” ones by issuing queries and warnings. In the real world, the air defense task involves a team of naval personnel. The experiment, however, was designed to be performed by a single individual by removing many of the subsidiary, technical tasks such as correlating raw radar data and operating radio circuits. The scenarios were designed to be cluttered and reasonably challenging by providing a variety of aircraft types and levels of threat.

Figure 1 shows a screenshot of the display used in the experiment (the actual display was in color). The tactical display showed a 170 × 120 nautical mile (315 × 222 km) area reminiscent of the Persian Gulf. Three relatively friendly countries, labeled F1, F2, and F3, appeared on the left, and a relatively hostile country, labeled H1, appeared on the right. Commercial air lanes appeared as faded (violet) lines that crisscrossed the display. Own ship was represented by the (blue) circle near the center of the display. Friendly aircraft appeared as (blue) bullet shapes. All unknown, potentially threatening aircraft, including commercial airliners, oil platform helicopters, maritime patrols, and tactical fighter aircraft appeared as (yellow) clover shapes (MIL-STD-2525B, Department of Defense, 1999). Less threatening aircraft appeared dim,

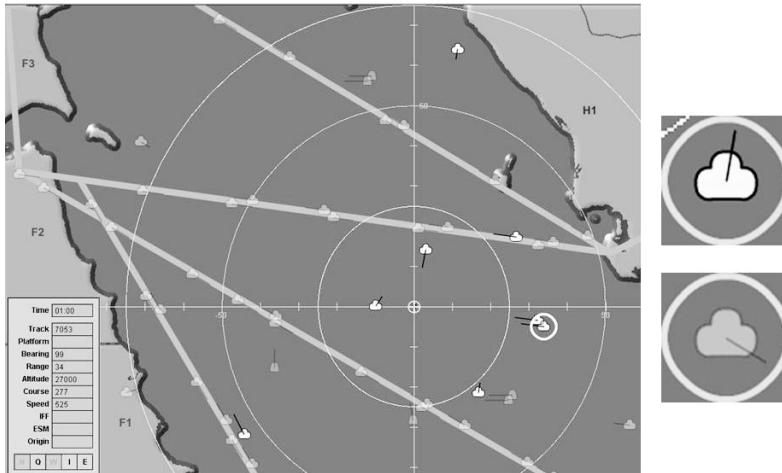


Figure 1. Screenshot of the task display (left). Close-up view of a fully visible aircraft (top right) and a decluttered aircraft (bottom right).

and the significantly threatening aircraft stood out as bright (yellow), amid the clutter.

Because participants were required to respond only to significantly threatening aircraft, a natural place to set the declutter threshold was to declutter all but the significantly threatening aircraft (defined as aircraft scoring an 8 or higher on a 10-point scale of threat). However, given the heuristic nature of the automated threat algorithm and variation among expert assessments, it was likely that the algorithm would occasionally declutter an aircraft that one or more participants might determine to constitute a significant threat. Lowering the threshold to keep more “borderline” threatening aircraft fully visible might reduce this problem, but at the cost of leaving more aircraft fully visible and increasing clutter on the display. More clutter means that users must spend more time searching among and evaluating a larger set of fully visible aircraft, only some of which are actually significantly threatening, in their view.

To investigate this trade-off empirically, we manipulated the declutter threshold as an independent variable in the study. In the high-threshold declutter condition, only aircraft that the threat assessment algorithm evaluated to be significantly threatening remained fully visible. In the medium-threshold declutter condition, all aircraft that the algorithm evaluated to be either significantly threatening or borderline threatening remained fully visible (6 or higher

on a 10-point scale). The declutter conditions were compared against a no-declutter condition, in which all aircraft symbols were equally salient.

METHOD

Participants

The participants were 27 U.S. Navy personnel (26 men and 1 woman). Ages ranged from 24 to 54 years, with a mean of 35 years. Eight of the participants were chiefs or senior chiefs (E-7 to E-8) from the Aegis Training and Readiness Center Detachment, San Diego; 3 were senior officers (O-5 to O-6) from the Tactical Training Group, Pacific; and 16 were junior officers (O-2 to O-4) from the Airborne Early Warning Wing, Pacific. The participants had from 3 to 30 years of service in the U.S. Navy, with an average of 13 years. Air defense expertise and experience was rated on a 3-point scale for each participant by an independent subject matter expert. Fourteen of the participants were given a very high rating, 2 were given a high rating, and 11 were given a moderate rating.

Task, Apparatus, and Stimuli

The experiment was run on a laptop with a 15-inch (38-cm) screen running at 1024 × 768 pixel screen resolution and viewed by the participant from a comfortable viewing distance.

In all conditions, users could access a variety of information about an aircraft (hereafter called

a track) by selecting a track with the mouse and then viewing a set of track data that appeared in a window in the lower left corner of the display. The track data included a track number for identification; the platform or type of aircraft; the bearing and range of the track from own ship; the altitude, course, and speed of the track; its country of origin; and two types of electronic/radar information: identification friend or foe (IFF) and electronic signal measures (ESM). For the purpose of realism, not all information was available for every track. For example, Track 7053 in Figure 1 is emitting no identifying electronic or navigational radar information; therefore its IFF and ESM are unknown and, consequently, the platform is also unknown. Additionally, the track flew in from the east over water, so its country of origin is unknown.

There were three equivalent scenarios, each lasting 15 min. During each scenario, tracks moved slowly about the display at realistic physical rates: from 95 to 560 nautical miles/hr (176–1037 km/hr), which is equivalent to 10 to 55 pixels/min (0.006° to 0.035° of visual angle/s). There were approximately 50 tracks on the display at all times, with tracks occasionally entering or exiting the displayed area. Most tracks appeared benign and nonthreatening, behaving like normal commercial airliners, oil platform helicopters, or other light commercial aircraft. At each moment, however, approximately seven tracks appeared significantly threatening (8 or higher on a 10-point scale) – for example, behaving like tactical fighter aircraft, moving at high speed, from hostile origins, toward own ship. Approximately 12 additional tracks appeared potentially threatening or “borderline” (6 or 7 on a 10-point scale of threat). These tracks presented a mix of benign and threatening attributes.

As tracks moved about the display, their threat levels changed. For example, as tracks approached own ship, their threat levels rose, and then as they passed, their threat levels dropped again. Occasionally, an aircraft would start out behaving like a commercial airliner following an air lane and would then abruptly change course and head inbound at high speed. Such actions would raise its threat score abruptly. Other tracks appeared suddenly from islands or oil platforms. In general, the scenario was designed to present

a range of aircraft behaviors and keep the participants engaged.

There were three conditions: no declutter, medium-threshold declutter, and high-threshold declutter. Assignment of scenarios to conditions was counterbalanced across participants. In the no-declutter condition, all track symbols appeared equally bright, and the user received no aid in evaluating the tracks for their levels of threat to own ship. In the two declutter conditions, less threatening tracks were decluttered.

The threat assessments were accomplished using an algorithm based on research into how navy experts evaluate threat (Liebhaber, 2001; Liebhaber et al., 2002; Marshall et al., 1996). Namely, the algorithm took as input 12 attributes (e.g., range, speed, origin, and whether a track was on an air lane) that are known to impact threat assessments. These attributes were weighed according to their mean impact on threat, as rated by a group of experts (Liebhaber, 2001), and then summed to produce a raw threat score. For example, a speed greater than 450 nautical miles/hr (833 km/hr) raised the raw threat score 1.8, whereas a speed of less than 150 nautical miles/hr (278 km/hr) raised the raw threat score 0.2. This algorithm treated each attribute independently, meaning that the algorithm did not take into account the implications of any high-order conjunctions of attributes. Hence the algorithm was relatively simple and heuristic in nature. More detail on the algorithm is available in St. John, Manes, Smallman, Feher, and Morrison (2004). Finally, the raw scores were transformed using the logistic function and rescaled between 1 and 10 to accentuate the midrange of the threat scale, given that few tracks ever received extreme scores.

Decluttering of the lower threat tracks was then accomplished by making their aircraft symbols semitransparent (65% transparent) so that the much darker background color showed through. In effect, the semitransparency reduced the luminance of the symbols to about one third of their initial values, similar to the approach used by St. John et al. (2002).

During the task, participants monitored the tracks and responded to the significantly threatening ones. Participants were instructed that the evaluation part of the task was their own judgment. They were also told that the threat

algorithm and declutter operation was only an imperfect aid: “The algorithm is not designed to be perfect – you are the final judge of threat and which tracks require actions. Instead, the algorithm is meant to provide a reasonable ‘first cut’ at evaluating threat. You should act on each track that you evaluate to be a significant threat. The algorithm is only there to help you focus on high threats.” These instructions both allowed and encouraged users to judge for themselves which tracks were significantly threatening. From postexperiment interviews, it was clear that these experienced participants were quite willing to believe the algorithm was fallible and to check its choices using their own judgment.

Once a track was judged to be a significant threat, however, the “rules of engagement” (ROE) determined how participants were *required* to respond. The ROE defined three concentric range rings around own ship and two types of “significant events” that required a response from participants: ring crossings and threat level increases. For ring crossings, participants were required to “notify alpha bravo” (i.e., click a button to notify a superior command element about a track) if a significantly threatening track crossed the ring at 75 nautical miles (139 km) from own ship; to “query” the track (i.e., click a button to initiate a radio message to the track) if it crossed the ring at 50 nautical miles (93 km) from own ship; and to “warn” the track (i.e., click a button to initiate a radio warning to the track) if it crossed the ring at 25 nautical miles (46 km) from own ship. Participants were required to perform these responses as quickly as possible. Only inbound ring crossings (toward own ship) counted as significant events. For threat level increases, if a previously less threatening track became a significant threat by performing some threatening action, such as turning inbound and increasing speed, then participants were asked to respond immediately with the response appropriate for that distance from own ship. Responses were always attributed to the most recent significant event.

These rules provided a good method for handling a common difficulty found in experiments on tasks that involve substantial expert user judgment, such as air defense. This difficulty arises from the variability among experts in their assessment of threat and in the timing of their responses (e.g., Morrison et al., 1996). This vari-

ability can make it difficult to measure performance benefits. In the current experiment, the assessment variability problem was addressed by allowing participants to exercise their own judgment in identifying significantly threatening aircraft and then, in the analyses of response timeliness, including only those aircraft that individual participants determined to be significantly threatening. The strict ROE for responding to significant threats, however, meant that any delay in responding could then be attributed to a loss of situation awareness rather than to user judgment about the appropriate timing of a response.

The complete lack of a response to a significant event, however, can still be attributed to a loss of situation awareness – the event was not observed – or to a participant’s judgment that the event was not significant. Any difference in response rates attributable to decluttering, therefore, can be interpreted either as a change in situation awareness or a change in participants’ threat assessments.

Participants made responses by first selecting the track, then clicking on the appropriate button underneath the track data display (N for notify, Q for query, or W for warn). Two additional responses, “illuminate with fire-control radar” (I) and “request to engage” (E), were also available to participants if they felt tracks represented an especially elevated level of threat. Unlike notify, query, and warn, however, no specific ROE were provided for when these two actions should be taken. These extra response options were included to provide added realism and to keep users occupied and engaged with the most threatening tracks, as they would be in the real task. They were not analyzed further because they were optional and subject to variable interpretation, unlike the concrete ROE.

The threat assessment algorithm identified 24 significant events during each scenario. It also identified 29 “borderline events,” when a borderline track crossed a ring or a track increased its threat level to become a borderline track, and 40 “low-threat events.” Of course, participants were required to respond only to those events that they personally judged to be significant. Additionally, at the beginning of each scenario, participants were required to “come up to speed” on the situation by immediately responding to

each significantly threatening track currently on the display with the response appropriate for that distance from own ship.

Procedure

Participants were given a basic description of the task and were then asked to sign informed consent forms. Participants were then given a detailed orientation to the display, the task, the ROE, and the tactical situation using a static screenshot of the basic, no-declutter condition. They were then briefly exposed to all three conditions and told that the purpose of the experiment was to see how the different displays might influence their performance. Participants then ran through a practice scenario with assistance from the experimenter. The practice scenario used the no-declutter condition and lasted 5 min.

Each participant performed in all three declutter conditions, one with each scenario, in a counterbalanced order. Twenty-four of the participants were administered the NASA Task Load Index (TLX; Hart & Staveland, 1988; National Aeronautics and Space Administration, n.d.) following each scenario in order to assess their subjective levels of workload.

RESULTS

Behavioral Measures

We first evaluated the benefits of decluttering for speeding responses to threats; then we evaluated the potential costs of decluttering for biasing users' threat assessments or causing potential threats to be missed.

Response times were computed by taking the difference between the time a response occurred (i.e., when the N, Q, or W button was clicked) and the time of the most recent ring crossing or threat level change event. Mean response times, both overall and for each level of threat, were then computed for each condition for each participant.

First, it is interesting that the response times were as long as they were: The mean response time was 31 s. Monitoring for significant threats and critical events must have required careful evaluation and close observation of individual tracks, which sometimes delayed the detection of other critical events. These long response

times underscore the need for any tool that can reduce this delay, albeit without incurring other large costs.

Our hypothesis was that decluttering the low-threat tracks would facilitate timely noticing and responding to the ring crossings and threat changes of significantly threatening tracks. To test this hypothesis, overall response times for each declutter condition were submitted to a one-way repeated measures analysis of variance (ANOVA). Decluttering significantly reduced response times, $F(2, 52) = 3.5, p = .037$ (see Figure 2). Response times were 25% faster in the high-threshold declutter condition than in the no-declutter condition (significant by Tukey-Kramer post hoc test). In a separate one-way repeated measures ANOVA of response times to only the significantly threatening tracks, response times were 28% faster in the high-threshold declutter condition than in the no-declutter condition, $F(2, 52) = 3.6, p = .035$.

Response times to only the borderline threat tracks were not significantly different between declutter conditions, $F(2, 40) = 1.2, p = .31$. Note that the reduced degrees of freedom in this analysis was attributable to the fact that 6 participants responded to no borderline threatening tracks in one or more declutter conditions. Response times to low-threat tracks could not be analyzed because so few participants ever responded to these tracks. The infrequency of responses to borderline and low-threat tracks limited their impact on the overall results. Overall, decluttering substantially improved the timeliness of most responses. It is important to note that the order of presentation of conditions did not affect the results: An analysis of the first condition presented to each participant showed the same pattern of results.

To investigate the effect of decluttering more closely, we split the response times based on the type of significant event that prompted them: ring crossings or threat level increases. The overall response times for each declutter condition and significant event type were submitted to a two-way repeated measures ANOVA. As expected, there was a significant main effect of declutter condition, $F(2, 52) = 4.6, p = .015$. The high-threshold declutter condition was significantly faster than the no-declutter condition (by Tukey-Kramer post hoc test). There was also

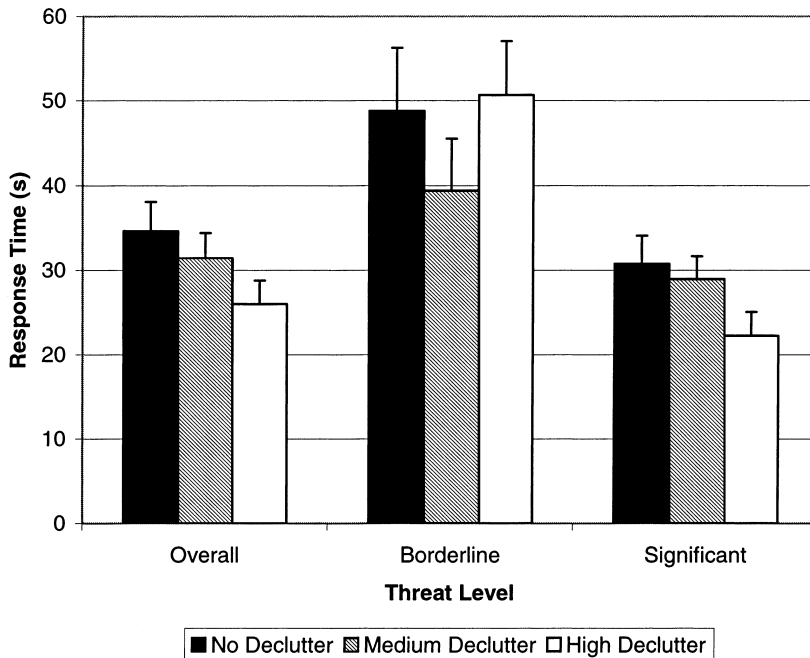


Figure 2. Effect of decluttering on response times, overall and broken down by level of threat.

a main effect of event type, $F(1, 26) = 57.6, p < .0001$. Response times to ring-crossing events (27 s) were faster on average than response times to threat level increase events (40 s). The interaction between declutter and event type was not significant, $F(2, 52) < 1$. These results indicate that decluttering facilitated the detection of both the relatively salient and predictable ring-crossing events and the relatively less salient and unpredictable threat change events.

Next, we split the response times based on the type of response: notify, query, or warn. Because these responses were designated to occur at different ranges from own ship, the three responses provided a convenient way to examine the effects of decluttering at different ranges from own ship and the center of the display. As Figure 3 shows, response times were fast and similar across declutter conditions for warnings, which occurred within 25 nautical miles (46 km) of own ship. However, for the queries at 50 nautical miles (93 km) and the notifications at 75 nautical miles (139 km), response times were slower and strongly influenced by decluttering. The response times were submitted to a two-way repeated measures ANOVA of response type and declutter condition. The main effect of response

type was significant, $F(2, 50) = 21.1, p < .0001$, and the main effect of declutter condition was significant, $F(2, 50) = 3.9, p = .028$. The interaction of response type and declutter condition was also significant, $F(4, 100) = 2.9, p = .025$. These results indicate that even the baseline display was sufficient for monitoring tracks close to own ship and that the real benefits of decluttering lie in facilitating the rapid detection and response to threats farther away from own ship. For the peripherally located notify responses, high-threshold decluttering improved response times by an impressive 44%. In effect, decluttering “buys time” for the user by helping the user to notice threats sooner and while they are farther away from own ship.

Finally, we split the response times by the participants’ levels of experience at air defense. To perform this analysis, the 2 highly experienced participants were dropped because of the small sample size. This reduction left 14 very highly experienced participants and 11 moderately experienced participants. Experience level did not influence response times ($F < 1$), indicating that decluttering was beneficial for both groups.

Given these response time benefits, did decluttering change the overall response rate or

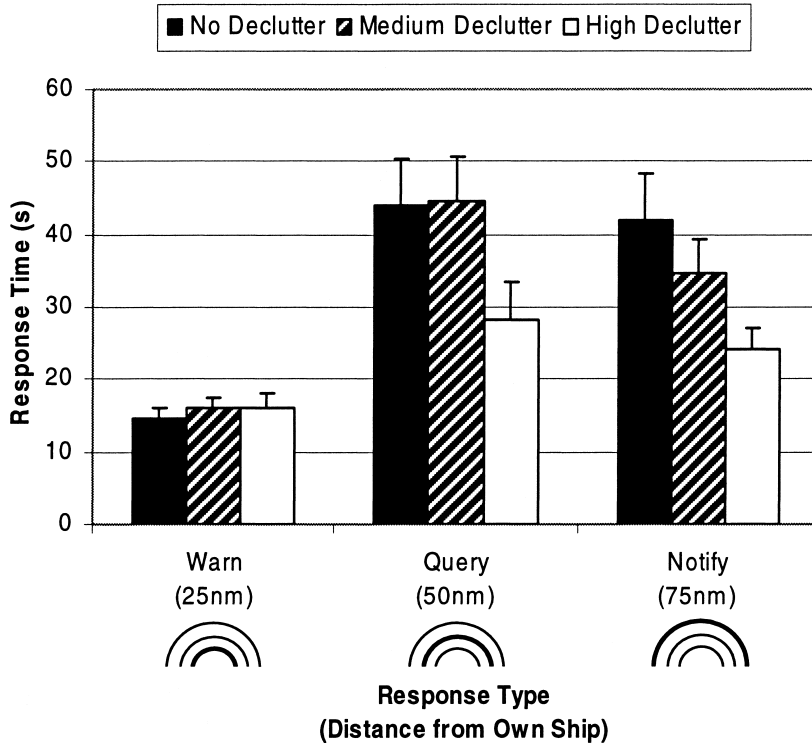


Figure 5. Effects of decluttering on response times, broken down by type of response (and distance from own ship). The semicircles indicate the response rings (warn, query, and notify).

which tracks elicited notify, query, and warn responses from participants? To answer these questions, we first asked how well participants agreed with the heuristic threat assessment algorithm. Participants responded an average of 21.4 times during each scenario (recall, for comparison, that the threat assessment algorithm identified 24 significant events during each scenario). On average, 80% (17.2/21.4) of participants' responses were made to tracks that the threat assessment algorithm identified as significant threats, and 16% (3.5/21.4) of their responses were made to tracks that the threat assessment algorithm identified as borderline threats. Only 3% (0.7/21.4) of participants' responses were made to low-threat tracks, and 81% of the participants responded to none of the low-threat tracks. These results indicate that the threat assessment algorithm and the participants closely aligned with one another in evaluating threat at this basic yet critical level of categorization. The heuristic automation was by no means perfect, however, and no participant responded to every automation-identified significant event.

An inspection of a sample of participants' responses revealed that different participants omitted responses to different tracks. One salient pattern, however, was that participants tended to omit responses to tracks that were following known commercial air lanes, even though the algorithm identified them as significant threats. The algorithm apparently underweighted the reduction in threat that participants attributed to this attribute. This finding was later confirmed in user interviews.

Second, did decluttering change the response rate? The numbers of responses for each declutter condition and level of threat were submitted to a two-way repeated measures ANOVA. Significant threats were responded to more frequently than were borderline or low threats, $F(2, 52) = 268, p < .0001$, but there was no overall effect of declutter, $F(2, 52) < 1$ (see Table 1). There was, however, a significant interaction between threat level and decluttering, $F(4, 104) = 4.4, p = .003$. To examine this interaction, we looked at each level of threat in separate one-way repeated measures ANOVAs. There

TABLE 1: Number of Responses

Condition	Threat Level			
	Overall	Low	Borderline	Significant
No declutter	21.3	1.0	3.5	16.8
Medium declutter	22.1	0.6	4.4	17.1
High declutter	21.0	0.7	2.7	17.6

were no differences in the number of responses to low-threat tracks, $F(2, 52) = 1.1, p = .33$, or to significantly threatening tracks, $F(2, 52) = 1.2, p = .30$, but the number of responses to borderline threats was affected by decluttering, $F(2, 52) = 5.3, p = .008$. Namely, there were more responses to borderline events when the medium-threshold declutter condition made borderline tracks fully visible than when the high-threshold declutter condition made these tracks decluttered ($p < .05$ by Tukey-Kramer post hoc test). The difference in responding, however, was very small in absolute terms: 4.4 responses in the medium-threshold declutter condition versus 2.7 responses in the high-threshold declutter condition.

One explanation for this difference is that decluttering led to a subtle bias in threat assessments. Namely, making borderline tracks fully visible (medium-threshold declutter) led users to judge these tracks as slightly more threatening, and therefore slightly more of these tracks received responses. Conversely, making borderline tracks decluttered (high-threshold declutter) led users to judge these tracks as slightly less threatening, and therefore slightly fewer of these tracks received responses. According to this explanation, the declutter manipulation led to a slight cost in terms of biased threat assessments, the very occasional underestimation of a threatening track, and a missed response to a significant event made by that track.

Fortunately, because of the nature of the task, these biases are less likely to affect decision making close to own ship, given that threat becomes more clear cut as tracks move closer. For the closest (warn) range ring, the number of responses to borderline threats dropped from 3.5 overall to 0.5, and the difference between conditions was not significant ($F < 1$).

A second explanation, which is not mutually exclusive from the first, is that decluttering led

to a small number of missed observations of significant events. If the threat assessment algorithm occasionally misevaluated a significant threat and decluttered it inappropriately, then this mistakenly decluttered significant threat might go unobserved, or missed, as it crossed a range ring, thereby lowering the response rate to borderline tracks. According to this explanation, the declutter manipulation led to a slight cost in terms of mistaken decluttering of significant threats that then went unobserved. As with the first explanation, this chain of events is more likely to occur in the periphery of the display. Tracks close to own ship are closely observed, as indicated by the fast response times to warn ring crossings (Figure 3). Ultimately, these small costs must be weighed against the performance benefits described earlier.

Did the declutter operation change the process of monitoring the display and maintaining situation awareness? Situation awareness was estimated by tabulating which tracks participants hooked (selected) in order to view and evaluate their detailed attribute values. The assumption was that participants would tend to repeatedly hook tracks that were threatening or otherwise worth a close examination. Therefore, elevated levels of hooking high-threat tracks should correspond with better situation awareness. The numbers of hooks for each declutter condition and level of threat were submitted to a two-way repeated measures ANOVA. Confirming the assumption, across all three conditions, participants primarily hooked the significantly threatening tracks, $F(2, 52) = 145.5, p < .0001$ (see Figure 4).

The overall amount of hooking, however, was not affected by decluttering, $F(2, 52) = 0.5$. This finding is important because it indicates that decluttering did not reduce participants' attention to and close monitoring of the situation, nor did it create extra work for participants by

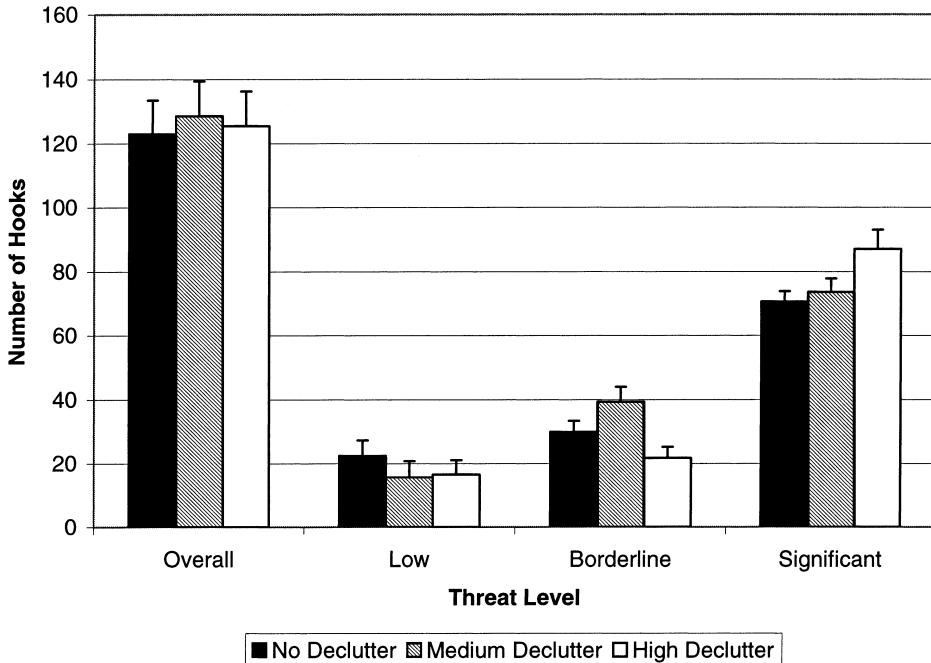


Figure 4. Effect of decluttering on the number of hooks, overall and broken down by level of threat.

influencing them to increase their hooking. Rather, when participants were in the declutter conditions, they continued to hook and evaluate tracks at the same rate as they did when they were in the no-declutter baseline condition.

Decluttering did influence *which* tracks were hooked, as indicated by a significant interaction between declutter condition and threat level, $F(4, 104) = 13.9, p < .0001$. To examine this interaction, we looked separately at each level of threat in one-way repeated measures ANOVAs. For significantly threatening tracks, high-threshold declutter increased the amount of hooking, $F(2, 52) = 9.4, p = .0003$. This finding indicates that participants watched and evaluated the significantly threatening tracks more closely when the declutter operation kept these tracks fully visible and decluttered the rest. Interestingly, this increase did not occur in the medium-threshold declutter condition, even though the medium condition also kept these tracks fully visible. Instead, the medium-threshold declutter condition increased the number of borderline threats that were hooked, $F(2, 52) = 19.7, p < .0001$. In other words, making only the significantly threatening tracks fully

visible (high-threshold declutter) increased participants' situation awareness of the high-threat tracks. Making both the significant and borderline threat tracks fully visible (medium-threshold declutter) increased participants' situation awareness of only the borderline threat tracks.

Perhaps participants hooked these borderline tracks more frequently than otherwise in order to understand why they had been made fully visible. In terms of costs, increased situation awareness of borderline threats might facilitate finding the occasional mistakenly decluttered threatening track but at the price of reducing surveillance of tracks that clearly are threatening.

Finally, even though the response time benefits of decluttering were similar for both experience levels, experience level did lead to several general differences in response rates and hooking rates. The overall numbers of responses in each declutter condition and experience level were submitted to a three-way mixed effects ANOVA of experience level, threat level, and declutter condition. Moderately experienced participants responded more (24 times) than did very highly experienced participants (19 times), $F(1, 25) = 8.5, p = .008$. Moderately experienced

participants responded to 1.4 more significantly threatening tracks, 2.1 more borderline threatening tracks, and 1.7 more low-threat tracks. In separate two-way ANOVAs at each level of threat, only the difference for borderline threats was significant, $F(1, 23) = 4.4, p = .048$. In a similar analysis of hooking rates, moderately experienced participants also hooked more borderline threat tracks, $F(1, 23) = 4.2, p = .05$, and more low-threat tracks, $F(1, 23) = 5.8, p = .024$, than did the very highly experienced participants. These increases were similar for all three declutter conditions.

The most likely explanation for these results is that the moderately experienced participants played the task more conservatively by judging more tracks to warrant responses. In the high-threshold declutter condition, this higher rate of responding meant that moderately experienced participants were actually more likely than the very highly experienced participants to disregard the automation's threat assessments, given that they responded to several decluttered borderline threat tracks. Contrary to conventional wisdom (including the conventional wisdom of the participants themselves), the less experienced participants did not doggedly follow the automation. If one assumes that experience leads to greater self-confidence at the task, then the very highly experienced participants should have been more confident and therefore more skeptical and less trusting of the automation (Lee & Moray, 1994). Instead, the moderately experienced participants appeared to be more skeptical of the automation than were the very highly experienced participants.

However, it seems likely that this conservatism is more a reflection of these participants' stance toward the task than of their stance toward the automation per se. It is also possible that the very highly experienced participants felt so confident of their abilities that they were more willing to follow the automation's recommendations, knowing they could change their minds if they chose. In general, the effects of trust in automation are extremely complex and multivariate (see Lee & See, 2004; Parasuraman & Riley, 1997). Moreover, it seems likely that attitudes toward mixed-initiative systems and lower level automation may be quite different from attitudes toward higher levels of automa-

tion that "take over" a task. The most important point, however, is that decluttering led to similar response time benefits for both groups.

Subjective Measures

Immediately following each scenario, 24 of the participants rated their subjective workload using the NASA-TLX (Hart & Staveland, 1988; NASA, n.d.). The overall indices for each declutter condition were submitted to a one-way repeated measures ANOVA. The effect of declutter was not significant, $F(2, 46) = 1.1, p = .35$. We then examined only the workload subscale that participants judged to be most relevant to the task: mental demand. In a similar analysis of mental demand only, the effect of declutter was significant, $F(2, 46) = 6.1, p = .004$. The subjective mental demand in the no-declutter condition was given an average rating of 49 out of 100, whereas both the medium- and high-declutter conditions were given average ratings of 40 out of 100. In terms of mental demand, then, decluttering reduced subjective workload ratings by an average of 18%. In a two-way mixed effect ANOVA of experience level by declutter condition, there was no effect of experience level on mental demand ($F < 1$), although mental demand was numerically lower for very highly experienced participants.

In interviews following the experiment, participants reiterated that decluttering reduced their workload, relieved the pressure to act and decide quickly, allowed time to concentrate on suspects, and aided situation awareness. Comments included "I actually had more time to spend scanning the display because I could see where the high threats were" and "With decluttering I had more time to loiter on a track of interest and put the puzzle pieces together."

When asked which condition they preferred, highly and very highly experienced participants split their preferences between the high-threshold declutter and the medium-threshold declutter interfaces. Moderately experienced participants overwhelmingly preferred the medium-threshold declutter interface. Only 2 of the 27 participants preferred the no-declutter condition. A common opinion was that "medium-threshold declutter helped narrow down the tracks that were better candidates to recheck" whereas the "high threshold left me more suspicious of the decluttered

tracks, [causing] greater workload.” This more conservative stance matches the behavioral data on the number of responses and the number of hooks, but it contrasts with the data on response times. Participants at all experience levels benefited similarly and solely from the high-threshold declutter interface. The medium-declutter interface may have felt “safer,” but it was the high-declutter interface that improved response times.

The similar effects of decluttering on response time and mental workload for both the highly and moderately experienced participants might appear to run counter to the classic findings of the expertise literature. One might have expected the moderately experienced participants to benefit more than the highly experienced participants. However, the effect of decluttering most likely influences fairly low-level visual search processes for quickly finding and refinding tracks of interest, and visual search processes are not likely to be strongly influenced by air defense experience. Similarly, Hollands and Merikle (1987) found that psychology experts were no faster than novices in searching an alphabetically organized menu system of psychology terms, although experts were faster to search a semantically organized system. To the extent that our scenarios contained ad hoc clutter, there was little structure for the experts to utilize.

DISCUSSION

Decluttering a naval air defense display using a heuristic threat assessment algorithm was successful in a number of ways. First, 25 out of the 27 experienced U.S. Navy users preferred one or the other of the two declutter interfaces over the baseline no-declutter interface. Second, participants rated the mental demands of the task as lower when using the declutter interfaces.

Third, the high-threshold declutter interface significantly improved response times to threatening tracks by 25%. Fourth, decluttering increased situation awareness of significant threats. Participants spent significantly more time monitoring and evaluating the tracks that the threat assessment algorithm identified as significantly threatening, as measured by which tracks were hooked during the scenarios.

The benefit of high-threshold decluttering was 9 s overall and more than 16 s for the middle and outer range rings. These are very substantial

differences, in terms of both absolute time and percentage increase. Although it is true that even a fast-flying aircraft will not travel more than a few miles within that time, it gives users a substantial period during which they can weigh decisions or evaluate additional aircraft. In one participant’s words, “Decluttering allowed me to get out in front of my [rules of engagement], rather than behind, where mistakes are made.”

These benefits must be weighed against the evident cost of decluttering, given that high-threshold decluttering slightly, but significantly, reduced the number of responses to borderline threat tracks as compared with medium-threshold decluttering. This difference may have been attributable to decluttering either slightly biasing users’ threat assessments of these tracks, leading to the occasional missed response, or to the occasional inappropriate decluttering of a significant threat, leading to an increased chance of a significant event going unnoticed. Because of the nature of the task, these costs are much more likely to occur in the periphery of the display. In our view, the large benefits in responding to sure threats outweigh the small costs in missed peripheral responses to unclear threats.

In important respects, the threat assessment algorithm performed quite well, even though it used relatively simple heuristics to assess threat. Rather than attempting to strictly rank order tracks from most threatening to least threatening, it merely attempted to categorize tracks as high threat (fully visible) or low threat (decluttered). At this less ambitious task, the algorithm was reasonably successful in that it reasonably closely matched the judgments of participants. In the no-declutter condition, in which the algorithm rated tracks but did not influence the display, 5% of participants’ responses, on average, were to low-threat tracks, 17% were to borderline threat tracks, and fully 79% were to significantly threatening tracks. Most important, this good, but imperfect, categorization performance by the threat assessment algorithm enabled the task performance benefits we have described. These benefits, we believe, derive from the way in which the automation was designed into the interface and used by the participants – namely, it suggested where users should focus their attention but still allowed them to scan the entire situation and respond as they saw fit.

The response time benefits for the high-threshold declutter interface are easy to understand. For the tracks that the algorithm assessed to be significant threats, ring-crossing events were clearly visible because these were the only fully visible tracks on the display. Threat level increase events were also easy to observe because these events typically caused a decluttered track to turn fully visible. Even if a participant did not see the actual change in status, once a track became fully visible, it was easy to notice quickly. On the rare occasion when participants determined that a decluttered track was in fact a significant threat, response times were substantially longer. However, these longer times were in fact about the same length as those in the baseline condition. Therefore, the high-threshold declutter interface led to substantial response time benefits when the participants and automation agreed and led to no delays when they disagreed.

For the medium-threshold declutter interface, in contrast, detecting ring crossings was more difficult because there were substantially more fully visible tracks to monitor, only some of which were actually significantly threatening. Similarly, threat level increases that turned a borderline track into a significant threat would have been difficult to detect because the borderline tracks were already fully visible. Consequently, this interface required close monitoring of the borderline tracks. These extra burdens placed on participants in the medium-threshold declutter condition may explain the relative lack of response time benefits.

Participants were split, however, in their preference for the medium- and high-threshold declutter interfaces. The medium-declutter interface was viewed as safer, and it fit with a more conservative stance toward decluttering. Similarly, our hypothesis going into the experiment had been that the medium-threshold declutter interface represented a sensible compromise between the “aggressive” decluttering of the high-threshold declutter interface and the baseline no-declutter interface. By leaving borderline threats fully visible, we reasoned that participants would never miss a threat but would still realize benefits from monitoring a reduced set of fully visible tracks. However, the response times do not support this conservative stance. The response times indi-

cate that high-threshold decluttering allowed participants to focus easily on the unambiguous threats of the fully visible tracks and still maintain a broad awareness of additional potential threats.

Note also that all participants were conservative and appropriately skeptical of the automation in the sense that all participants continued to hook and evaluate decluttered tracks and even occasionally ordered responses to decluttered tracks. No participant mistook the threat assessment algorithm for a perfect indicator of threat. This level of continuing verification may be surprising to some, but it may make more sense in light of two facts. First, U.S. Navy users have substantial experience with new technology during naval exercises and tend to be sensibly wary. Second, the declutter system made verification very easy, so that users could remain engaged and continue to evaluate tracks for themselves simply by selecting tracks and viewing their data. This continued verification limited the chance of a misevaluated track failing to be attended, as indicated by the occasional response to a borderline threat track.

An interesting compromise between high- and medium-threshold declutter might be a two-level decluttering that codes significant and borderline threats differently from each other and from low threats. This design is reminiscent of multilevel alerts (Sorkin, Kantowitz, & Kantowitz, 1988; St. John & Manes, 2002) and fuzzy signal detection (Parasuraman, Masalonis, & Hancock, 2000). Such a display would still clearly identify significant threats but also provide support for more conservative performance by identifying borderline threats as well. The danger is that the different codings must remain easily discriminable, or users will be unable to efficiently focus their attention.

It is interesting and important to consider how these results might change, and where the optimal threshold might lie, as the numbers of significant, borderline, and low-threat tracks and the reliability of the algorithm change – for example, by placing the algorithm in a different scenario context. Here, we found that setting the declutter threshold to match the definition of significant threat was better than setting the threshold lower. If the number of significant threats increased, then the burden of closely

monitoring them would increase, but the benefit of identifying them against the background of lower threat tracks ought to remain. If the number of low threats increased, then the occasional burden of scanning the decluttered tracks for hidden threats would increase somewhat, but once again, the benefit of identifying the significant threats ought to remain. If, however, the number of borderline threats increased or if the reliability of the algorithm decreased, then the burden of scanning the decluttered tracks for hidden threats would rise rapidly. The clear identification of the obvious threats, however, should continue to provide important response time benefits in comparison with the baseline no-declutter display. We are currently planning to systematically vary these parameters in a controlled laboratory version of this monitoring task and to observe how the costs and benefits play out.

Another limit on the generalization of these findings is the short-term nature of the scenarios. In practice, users stand watch for hours at a time, over periods of weeks and months, and significant threats are typically few and far between. Whether these differences in task duration and threat frequency would change the results of the study are unknown. Users who guard against automation bias in the short term might be lulled into complacency over the long term. Future research and development of the declutter concept will need to take this possibility into account. For instance, it may be possible to implement design features to help guard against this potential hazard. One such possibility, which resonated with participants during their interviews, would be to adapt the declutter threshold to suit the situation. This adaptation could be controlled by the system based on performance or task variables (e.g., Parasuraman, Mouloua, & Molloy, 1996), user physiology variables (e.g., Mikulka, Scerbo, & Freeman, 2002), or by the users themselves (see Kaber & Endsley, 2004, for a recent review). For instance, users could set the threshold low during relatively benign situations in order to see any potential threats, and they could set the threshold high during more tense situations in order to focus on the more significant threats.

Finally, although the current experiment was designed to demonstrate the basic benefits of

decluttering, there are, in fact, numerous ways in which the declutter interface may be improved. The most important suggestion from participants was, in our view, to better indicate changes in threat and declutter status. During the experiment, threat level increases that changed a track from nonthreatening to significantly threatening produced a relatively salient change in visibility: from a dimmed, decluttered symbol to a fully visible symbol. However, these relatively large visibility changes still led to fairly long response times. It seems likely that in many cases, participants did not actually observe the status and symbol changes but found the already-changed tracks during their normal scanning around the display. Research in change blindness (see Rensink, 2002) shows that even salient changes in a display may be difficult to observe unless they happen to be directly attended at the moment of change. Causing the track symbols to flash following a status change might effectively draw users' attention, although perhaps at the price of distracting users from other critical tasks. An alternative concept, in the vein of "negotiated interruptions" (McFarlane, 2002), is a "change history" tool (Smallman & St. John, 2003) that preserves a record of important changes on a tactical display and improves users' ability to remain apprised of important changes without unduly distracting them.

CONCLUSIONS

The current experiment shows, within in the context of a realistic monitoring task, that decluttering less important tracks, by dimming their symbols, can produce important performance benefits. It further shows that even relatively simple heuristic automation for identifying potential threats and assisting display search can prove quite effective when it is designed to support and guide users, rather than to replace them, and when the verification process is easy and built into the normal tasking of the user.

ACKNOWLEDGMENTS

The authors would like to thank subject matter experts Gene Averett, especially for his assistance in creating the scenarios used in the experiment, and Ronald Moore. This research

was sponsored by the Office of Naval Research and the Space and Naval Warfare System Center, San Diego.

REFERENCES

- Baddeley, A. D. (1972). Selective attention and performance in dangerous environments. *British Journal of Psychology*, *63*, 537–546.
- Department of Defense. (1999). *Department of defense, interface standard, common warfighting symbology* (MIL-STD-2525B). Washington, DC: Department of Defense, Defense Information Systems Agency. Available at <http://symbology.disa.mil/>
- Fisher, D. L., Coury, B. G., Tengs, T. O., & Duffy, S. A. (1989). Minimizing the time to search visual displays: The role of highlighting. *Human Factors*, *31*, 167–182.
- Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier.
- Hollands, J. G., & Merikle, P. M. (1987). Menu organization and user expertise in information search tasks. *Human Factors*, *29*, 577–586.
- Johnson, W. W., Liao, M., & Granada, S. (2002). Effects of symbol brightness cueing on attention during a visual search of a cockpit display of traffic information. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 1599–1603). Santa Monica, CA: Human Factors and Ergonomics Society.
- Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics*, *5*, 113–153.
- Kaempf, G. L., Wolf, S., & Miller, T. E. (1993). Decision making in the AEGIS combat information center. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 1107–1111). Santa Monica, CA: Human Factors and Ergonomics Society.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*, 153–184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80.
- Liebhauer, M. J. (2001). *Description and evaluation of an air defense threat assessment algorithm* (Tech. Rep.). San Diego, CA: Pacific Science & Engineering Group.
- Liebhauer, M. J., Kobus, D. A., & Feher, B. A. (2002). *Studies of U.S. Navy air defense threat assessment: Cues, information order, and impact of conflicting data* (Tech. Rep. SSC-1888). San Diego, CA: Space and Naval Warfare Systems Center.
- Marshall, S. P., Christensen, S. E., & McAllister, J. A. (1996). Cognitive differences in tactical decision making. In *Proceedings of the 1996 Command and Control Research and Technology Symposium* (pp. 122–132). Washington, DC: Department of Defense, Command and Control Research Program.
- McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, *17*, 63–139.
- Mikulka, P. J., Scerbo, M. W., & Freeman, F. G. (2002). Effects of a biocybernetic system on vigilance performance. *Human Factors*, *44*, 654–664.
- Morrison, J. G., Kelly, R. T., & Hutchins, S. G. (1996). Impact of naturalistic decision support on tactical situation awareness. In *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting* (pp. 199–203). Santa Monica, CA: Human Factors and Ergonomics Society.
- National Aeronautics and Space Administration. (n.d.). *Task Load Index [TLX] Version 1.0, user's manual*. Available at <http://iac.dtic.mil/hsiac/Products.htm#TLX>
- Nugent, W. A. (1996). Comparison of variable coded symbology to a conventional tactical situation display method. In *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting* (pp. 1174–1178). Santa Monica, CA: Human Factors and Ergonomics Society.
- Osga, G., & Keating, R. (1994). *Usability study of variable coding methods for tactical information display visual filtering* (Tech. Rep. NOSC-2628). San Diego, CA: Naval Command, Control and Ocean Surveillance Center, Research, Development, Test, and Evaluation Division.
- Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors*, *42*, 636–659.
- Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, *38*, 665–679.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, *30*, 286–297.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.
- Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, *53*, 245–277.
- Schultz, E. E., Nichols, D. A., & Curran, P. S. (1985). Decluttering methods for high density computer-generated graphic displays. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 300–303). Santa Monica, CA: Human Factors and Ergonomics Society.
- Smallman, H. S., & St. John, M. (2003). CHEX (Change History EXplicit): New HCI concepts for change awareness. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 528–532). Santa Monica, CA: Human Factors and Ergonomics Society.
- Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors*, *30*, 445–459.
- St. John, M., Feher, B. A., & Morrison, J. G. (2002). *Evaluating alternative symbologies for decluttering geographical displays* (Tech. Rep. SSC-1890). San Diego, CA: Space and Naval Warfare System Center.
- St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332–336). Santa Monica, CA: Human Factors and Ergonomics Society.
- St. John, M., Manes, D. I., Smallman, H. S., Feher, B. A., & Morrison, J. G. (2004). *An intelligent threat assessment tool for decluttering naval air defense displays* (Tech. Rep. SSC-1915). San Diego, CA: Space and Naval Warfare System Center.
- St. John, M., Oonk, H. M., & Osga, G. A. (2000). Designing displays for command and control supervision: Contextualizing alerts and “trust but verify” automation. In *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 6.646–6.649). Santa Monica, CA: Human Factors and Ergonomics Society.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.
- Van Orden, K. F., DiVita, J., & Shim, M. J. (1995). Redundant use of luminance and flashing with shape and color as highlighting codes in symbolic displays. *Human Factors*, *35*, 195–204.
- Yeh, M., & Wickens, C. D. (2001a). Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques. *Human Factors*, *43*, 543–562.
- Yeh, M., & Wickens, C. D. (2001b). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, *43*, 355–365.

Mark St. John is director of the Cognitive Systems Division at Pacific Science & Engineering Group, Inc. He received his Ph.D. in cognitive psychology in 1990 at Carnegie-Mellon University.

Harvey S. Smallman is a senior scientist at Pacific Science & Engineering Group, Inc. He received his Ph.D. in experimental psychology in 1993 at the University of California, San Diego.

Daniel I. Manes is a senior human factors engineer at Pacific Science & Engineering Group, Inc. He received his M.S.E. in industrial and operations engineering in 1997 at the University of Michigan, Ann Arbor.

Bela A. Feher is a senior scientist at the Space and Naval Warfare System Center, San Diego. He received

his Ph.D. in social psychology in 1970 at Wayne State University.

Jeffrey G. Morrison is a senior scientist at the Space and Naval Warfare System Center, San Diego. He received his Ph.D. in psychology in 1992 at the Georgia Institute of Technology.

Date received: December 31, 2003

Date accepted: November 5, 2004