Reliability requirements for augmented reality in visual search tasks

Samuel S. Monfort,
John J. Graybeal
KINEX, Inc.
Manassas, VA

Ewart de Visser
Warfighter Effectiveness Research Center, U.S. Air Force Academy
Colorado Springs, CO

Todd Du Bosq
U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate
Fort Belvoir, VA

ABSTRACT

In military operations, quick and accurate target detection and identification is critical for mission success. Augmented reality (AR) technologies can aid target detection and identification by layering digital imagery atop a Soldier's field of view to increase situational awareness. These systems are rarely perfect, however, and in some cases unreliable AR may actually interfere with performance. This investigation focused on the capacity for unreliable AR to impair performance. We showed participants a series of 2D simulations where highly-visible AR cues were superimposed over tanks placed randomly in a grassland environment. The reliability of these cues varied (from 25% to 100%) throughout the experimental session, as some valid targets were erroneously un-marked (false negatives) while some invalid targets were erroneously marked (false positives). Participants were asked to search for the vehicles while being assisted by the AR; search accuracy and response time were analyzed, and participants provided feedback regarding their mental workload and trust in the AR. We found the expected negative relationship between unreliability and performance, but also found that AR false positives were more damaging to performance than AR false negatives. Unreliable AR also hurt performance more when marking vehicles at greater distances. Further, although error type and target distance had powerful effects on participant performance, they had less of an impact on subjective trust and workload, suggesting that Soldiers using AR might not be consciously aware of how their own performance changes as a function of AR properties. In summary, unreliable AR hurt performance differently depending on the type of errors produced by the system, and impaired some aspects of performance but not others. These results carry important implications for how AR is designed to improve performance on the battlefield.

Keywords: augmented reality, trust in automation, visual detection and identification, mental workload, human performance

ABOUT THE AUTHORS

Samuel S. Monfort is a former Research Psychologist with KINEX, INC., who worked with the U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate at Fort Belvoir. He received his Ph.D. in 2017 from George Mason University in Human Factors and Applied Cognition. Dr. Monfort's research interests include trust in automation, human-computer interaction, user interface design, measurement, and statistics. He has conducted research on augmented reality in a number of military contexts, including minesweeping and basic navigation. His work has focused on matching machine characteristics to Soldier capabilities, and has directly informed the design of augmented reality devices. He recently accepted a position with the Insurance Institute for Highway Safety.

John J. Graybeal is a Research Psychologist with KINEX, INC., working with the U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate at Fort Belvoir (NVESD). He received his Ph.D. in 2017 from George Mason University in Psychology, with a concentration in Cognitive and Behavioral Neuroscience. Since 2017, he has worked in NVESD's Perception Lab, where he evaluates human performance with a wide range of sensor imagery and emerging technologies. He regularly conducts research with Soldiers and other human subjects using visual tasks (e.g. object detection, object recognition) with both real and simulated imagery. He is responsible for leading the NVESD Perception Lab's efforts to modernize and refine human testing methodologies, including training practices for vehicle identification skills that support NVESD perception tests. His research also focuses on human performance with augmented reality technologies. He is responsible for conducting augmented reality simulation experiments in addition to providing design recommendations for and evaluations of augmented reality displays.

Dr. Ewart de Visser received his Ph.D. in applied cognitive psychology from George Mason University and his B.A. in Film Studies from the University of North Carolina at Wilmington. He also received his propedeuse in Cognitive Artificial Intelligence (CKI) from Utrecht University in the Netherlands. Dr. de Visser is currently an affiliated faculty of George Mason University and Clemson University. He is also the founder and president of de Visser Research, LLC. Previously, Dr. de Visser worked for 11 years at Perceptronics Solutions, Inc (2005-2016), where he served as the Director of Human Factors and User Experience Research and created robotic interfaces and displays, produced mobile experiences and visualizations in the areas of emergency medicine, stress training, and cyber security, and applied the use of multi-agent systems to manage multiple mixed initiative human-robot teams. In the area of HRI, Dr. de Visser published peer-reviewed papers on the measurement of trust in robotics, adaptive human-robot teaming, human-automation etiquette, and virtual agent anthropomorphism. Dr. de Visser has explored human-automation trust from different theoretical perspectives—cognitive, social, and neural and has published numerous papers on the measurement of trust in robotics, adaptive human-robot teaming, human-automation etiquette, and virtual agent anthropomorphism.

Todd Du Bosq is the Field Performance Branch Chief in the Modeling and Simulation Division at the U.S. Army RDECOM CERDEC Night Vision and Electronics Sensors Directorate. He received his B.S. degree in physics from Stetson University in 2001 and his M.S. and Ph.D. degrees in physics from the University of Central Florida in 2003 and 2007, respectively. Since 2007, Dr. Du Bosq has provided innovative research and guidance to the U.S. Army on the human acquisition performance of infrared and reflective sensors regarding military vehicles, human targets, laser markers, and IEDs for source selections, war games, simulation, training, and system performance trade studies. He also oversees a diverse set of visual perception experiments, training, and demonstrations being conducted in the NVESD Perception Laboratory.

Reliability requirements for augmented reality in visual search tasks

Samuel S. Monfort,
John J. Graybeal
KINEX, Inc.
Manassas, VA

Ewart de Visser
Warfighter Effectiveness Research Center, U.S. Air Force Academy
Colorado Springs, CO

Todd Du Bosq
U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate
Fort Belvoir, VA

INTRODUCTION

Soldiers rely on advanced image processing to assist with target identification (Biros, Daly, & Gunsch, 2004; Yeh & Wickens, 2001), and the application of this processing has increasingly involved augmented reality (AR). In head-mounted displays and other systems, AR layers digital imagery atop a Soldier's field of view to increase situation awareness and combat effectiveness. These systems are rarely perfect, however, and the extent to which unreliable AR adversely affects performance is the focus of this investigation. Specifically, we sought to estimate the reliability level necessary for AR cues to be useful in a target search and identification context. We were also interested in understanding when unreliable AR might actively interfere with performance, resulting in AR-aided performance that is worse than performance during "manual" un-augmented search.

Automation and Supervisory Control

Manual tasks are automated out of a desire to improve performance and efficiency while reducing human workload and fatigue. There are a great number of automation success stories, from commercial aviation (e.g., Wiener & Curry, 1980) and medical procedures (e.g., Armato et al., 2002), to military operations (e.g., Deng, Han, & Mishra, 2003). However, far from removing human responsibility altogether, automation has largely increased the responsibilities of human operators who now must assume managerial positions over automated systems. Even for tasks that are entirely automated (so-called "unmanned" tasks), a human operator is typically placed in a supervisory role. The assumption of supervisory duties introduces the potential for errors related to the human's failure to attend or respond to the automation's mistakes (Bagheri & Jamieson, 2004; Parasuraman & Riley, 1997). Although AR on the battlefield does not necessarily equate to automation per se (e.g., a person instead of a machine could be determining displayed AR information), the net effect on Soldiers is the same: they must incorporate information provided to them from an outside source with their own expertise to make decisions. AR that fails to provide correct information may have a net negative effect on overall performance if those failures go unnoticed, if the failures hurt operator trust to an extent that non-failures are viewed skeptically, or if they simply distract the operator. Thus, at a certain level, having poorly-performing AR may be worse than having no AR at all.

AR Error Type

The extent to which AR unreliability affects performance, trust, and workload in a visual search task may depend on the nature of that unreliability: whether errors are misses (i.e., false negatives) or false alarms (i.e., false positives). In many instances, AR provides feedback to a human operator through alerts—prompting subsequent action through an alert's presence or suggesting inaction through its absence (cf. errors of omission/commission; Parasuraman & Manzey, 2010). System designers can affect whether AR errors will tend to be misses or false alarms by adjusting alert sensitivity. In this context, AR misses and false alarms should have qualitatively different effects on operator behavior: the former encourages action and the latter inaction (Meyer, 2001; Meyer, 2004).

Dixon, Wickens, and McCarley (2007) suggest that because false alarms are salient, intrusive, and annoying, the same level of overall unreliability should be more impactful if errors are false alarms than if they are misses. However, other research has found false alarms and misses to have a similar negative effect on performance (Madhavan, Wiegmann, & Lacson, 2006; Rovira & Parasuraman, 2010), and a few studies have even found that participants may trust a false alarm-prone system more than a miss-prone system (Davenport & Bustamante, 2010). Whether misses or false alarms induce lower trust and/or are more detrimental to performance likely depends on the particular task at hand. For visual search tasks, the intrusiveness of false alarms may interfere with a Soldier's ability to reflexively orient visual attention towards important visual stimuli (i.e., bottom-up attention), as attention may be captured by false but salient AR cues

(Treisman, 1985; Nothdurft, 1992). That is, a great number of false alarms will likely spread visual attention too diffusely, mitigating the benefits of AR visual cues that help a Soldier reflexively orient on valid threats (i.e., "visual clutter": Yeh et al., 2003). Furthermore, the salience of false alarms may also interfere with conscious efforts to direct visual attention (i.e., top-down attention), as Soldiers may begin to distrust the system and subsequently change the way they consciously direct attention to AR visual cues. If false alarms are too prevalent, Soldiers may begin ignoring valid visual cues from the AR system (i.e., automation "disuse": Parasuraman & Riley, 1997). Thus, in a target identification context, false alarms may interfere with both bottom-up and top-down attentional processing.

Although false alarm-prone AR in a visual search task may be closely related to performance, trust, and workload, miss-prone AR may not. One study on visual search found that a system that missed as many as 70% of targets nonetheless improved performance over a manual control baseline (de Visser & Parasuraman, 2011). The authors suggest that in a field of multiple targets, any that a system identifies will alleviate the workload of a human observer. Put simply, even a miss-prone AR system may help Soldiers locate targets because the errors are not distracting. Indeed, the visual information presented to an operator with perfectly unreliable miss-prone AR is equivalent to a no-AR condition of the same visual search scenario (in both cases, all targets are unmarked). Thus, in contrast to false alarms, misses may not interfere with bottom-up or top-down attentional processes, and may therefore be less damaging to performance, trust, and workload.

To explore the moderating effect of AR error type on the relationship between AR reliability and overall performance, the current study assigned participants to either miss-prone or false alarm-prone AR error type conditions while keeping the overall number of errors consistent between groups.

Perceptual Difficulty

Research consistently finds that greater task difficulty increases the potential risks of unreliable systems (Dixon & Wickens, 2006; Meyer, 2001; Parasuraman et al., 1993). As targets become more difficult to perceive, unreliable AR may therefore be more likely to adversely affect performance. In fact, past research has found that system unreliability only affected performance for difficult tasks; easy tasks were associated with relatively high, invariant performance (Dzindolet, Pierce, Pomranky, Peterson, & Beck, 2001). Thus, the second focus of this investigation was to validate the effect of perceptual difficulty on the relationship between AR reliability and performance in a military context. We manipulated perceptual difficulty by varying the target's distance from the observer. Soldiers must be capable of identifying targets at extreme ranges, often when few visual cues are present (McDowell, 1992; Sterling & Jacobson, 2006). Therefore, understanding the effect of AR reliability when targets are distant and difficult to perceive is critically important.

Summary and Hypotheses

Soldiers stand to benefit from AR that assists them with target detection and identification, but the reliability requirements for such a system remain unclear. Unreliable AR may not only fail to improve performance, but may also interfere with combat operations if errors are intrusive or distracting.

This investigation will focus on both objective performance (i.e., response accuracy and response time) and subjective perceptions (i.e., trust in AR, cognitive workload, and mental resource drain). Our two hypotheses pertain to the effect of AR reliability on these outcomes. First, we hypothesized that errors in false alarm-prone AR will be more damaging to both objective performance and participant subjective state than errors in miss-prone AR. Second, we hypothesized that errors (either false alarms or misses) above distant targets will be more damaging to objective performance than errors above close targets. This study did not examine the relationship between range and subjective state, however, as subjective surveys were administered after scenes that included a mix of close and distant targets.

METHOD

Participants were randomly assigned to one of two conditions corresponding to the type of error issued by the AR: misses or false alarms. In either case, reliability of the AR varied throughout the experiment, with probability of a correct AR response set to six levels: 25%, 40%, 55%, 70%, 85%, and 100%.

Participants were instructed to mark all tanks in a series of static scenes composed of grassland, sparse trees, and several buildings as quickly and as accurately as possible by clicking on them (no time limit). They were further instructed that six different AR systems would take turns assisting them with their search, and that some systems would provide more accurate guidance than others. Guidance took the form of AR icons superimposed over targets. After working with each system, participants were asked to evaluate its reliability and trustworthiness.

Participants

A total of 184 participants were recruited, both from the George Mason University undergraduate research pool (n = 83) and from Amazon's Mechanical Turk (MTurk; n = 101). MTurk participants were excluded for poor monitor resolution (<900 vertical pixels; n=32) or if they did not finish the task (n=12), reducing the MTurk sample size to 57 and the total sample size to 140. Undergraduate participants were compensated with research credit hours and MTurk participants were paid $3. Undergraduate participants were typically younger (M=21.3 years, SD=2.62) and took less time completing the experiment (M=31.6 minutes) than MTurk participants (M=35.9 years, SD= 10.6; M=38.6 minutes). Military personnel were not required for this experiment because we used a simple visual perception task that did not require military experience or training. To this end, we used readily interpretable AR symbology (i.e., a picture of a tank) rather than standard military symbology.

Scene Generation

A total of 54 grassland scenes were generated for participants to search. These 54 scenes contained between zero and eight "targets" each, placed randomly in the environment. These targets were either tanks with tank icons above them (accurate), tanks without tank icons above them (inaccurate: miss), or houses with tank icons above them (inaccurate: false alarm). Only one of the two inaccuracies were shown to any single participant, as each was assigned to either the miss condition or the false alarm condition (a between-subjects factor; Figure 1).

Whether or not a given target was accurate depended on the reliability block to which the scene belonged. The 54 grassland scenes were divided into six blocks corresponding to the various reliabilities of the AR system that would be assisting participants (25%, 40%, 55%, 70%, 85%, 100%). Each reliability block contained nine scenes (54÷6), with a single instance of each 0 to 8 targets. In other words, all reliability blocks contained one scene that had 0 targets, one scene that had 1 target, one scene that had 2 targets, and so on, up to a maximum of 8 targets per scene, for a total of 36 targets.

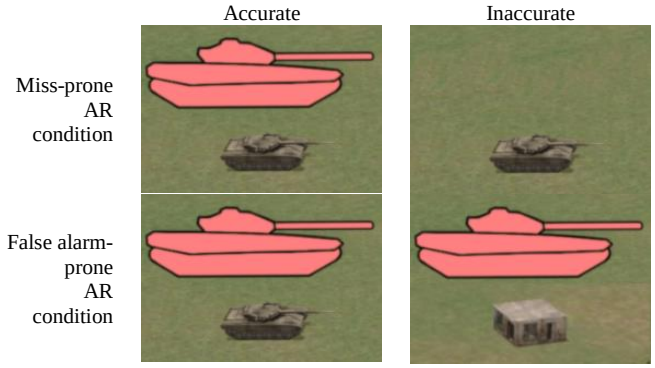

Figure 1. Targets for each AR error type.

The proportion of the 36 targets that were accurate was either 25% (9 accurate, 27 inaccurate), 40% (14 accurate, 27 inaccurate), 55% (20 accurate, 16 inaccurate), 70% (25 accurate, 11 inaccurate), 85% (30 accurate, 6 inaccurate), or 100% (36 accurate, 0 inaccurate). In addition to the targets described above, each scene contained 30 trees; three building clusters that appeared at short, medium, and long ranges; and three small buildings. With the exception of the three building clusters that appeared at fixed ranges (300, 500, and 700 meters), the placement of the various background elements was determined randomly for each scene. The horizon was populated with a dense forest (see Figure 2 for an example).

Self-report Surveys

A number of self-report surveys were also used. Subjective assessments provide insight into participants' mental state as they complete the various search tasks. Discrepancies or similarities between participants' subjective experience and their performance can have critical practical consequences (e.g., participants failing to recognize an objective safety hazard), and subjective responses often provide critical insight into why certain objective performance results were observed. The first survey pertained to trust in AR, modified from Lee and Moray's (1992) trust scale; the second was an overall workload scale taken from the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988); and

the last was the Gas Tank Questionnaire (GTQ), assessing remaining mental resources (Monfort et al., 2017). Participants also completed nine scenes without any AR; the trust scale items that followed these scenes were modified to refer to self-confidence instead.

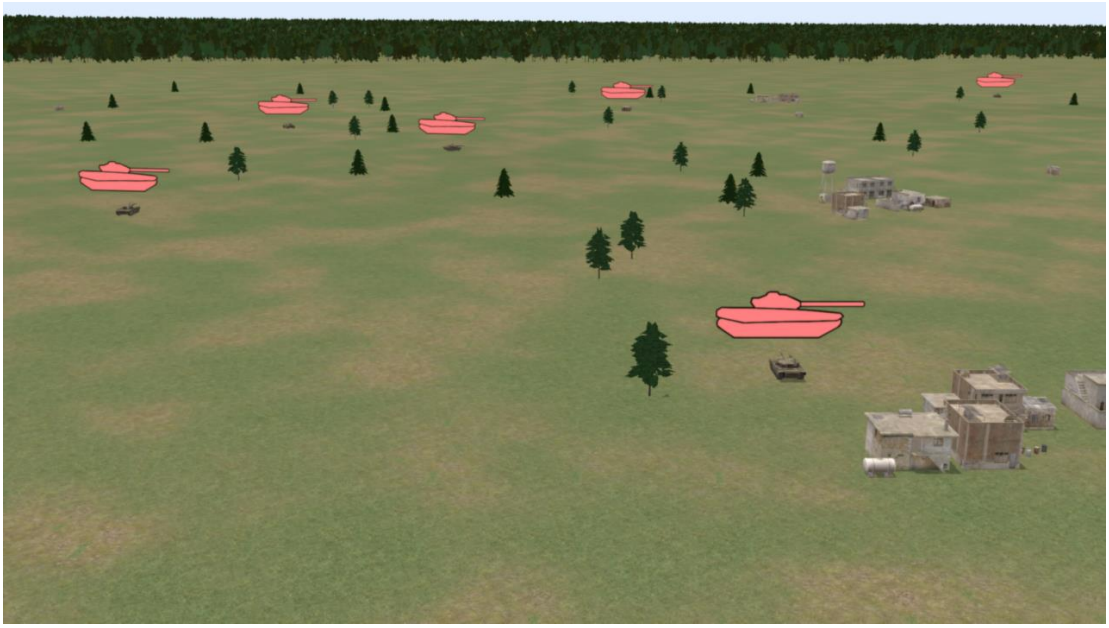

Figure 2. Example scene from the experiment, selected from the 85% reliable false alarm-prone AR block. The incorrectly-marked target (a house) is fourth from the left.

Procedure

After arriving at the lab, participants completed informed consent and were assigned to an experimental condition (i.e., AR error type). They then underwent a tutorial and training session to familiarize themselves with the task. This session included descriptive text as well as two sets of training scenes: one with very reliable AR (100%), and one with very unreliable AR (25%). These trials were presented in sequence to familiarize the participants with the concept of variable AR reliability, and to highlight the fact that some systems were more reliable than others. Very unreliable AR was selected for training both because we wanted participants to experience the full range of AR performance and because research suggests the first time operators experience a major system failure, trust decreases substantially (Rovira, McGarry, & Parasuraman, 2007); we wanted this initial drop in trust to be uniform for all participants. Depending on the condition to which they were assigned, participants either completed a tutorial for a miss-prone or for a false alarm-prone tank detection system.

After completing the training, participants were asked to complete a trial without any AR. In this trial, none of the nine scenes contained any AR marks. At its conclusion, participants were asked to rate their self-confidence (i.e., "Slide the bars below to express your feelings about your ability to spot tanks without help from the tank detection system"). They were also asked how demanding they found the task and to rate their remaining mental resources. Subsequent trials with AR marks also used these scales, but inquired about trust rather than self-confidence.

Participants were then advised that the tank detection system was "coming online," and that they would be completing a number of trials with different systems. They were told that they would be rating the trustworthiness of six systems, and were instructed to rate each system independently, and to "reset their trust" after each trial. This instruction was given to reduce the likelihood of any trust carry-over effects. See Figure 3 for a graphical overview of the procedure.

Data Analysis

All analyses were conducted using generalized linear mixed-effects regression models (Bates, Mächler, Bolker, & Walker, 2015), and Sattherwaite (1946) approximations were used to determine denominator degrees of freedom for *t* and *p* values. *P* values are associated with statistical tests of our hypotheses, and for our purposes, *p* values that are

less than .05 represent effects that are sufficiently unlikely to have occurred by chance. Performance outcomes were mostly captured at the response-level as binary data: whether or not a given target was clicked (miss) and whether or not a given click was on a target (false alarm). These outcomes were analyzed using logistic regression models. How long participants spent searching each scene was recorded in seconds and analyzed using Poisson regression models. The three self-report assessments measuring trust, workload (NASA-TLX), and remaining mental resources (GTQ) were collected at the end of each nine-scene reliability block, and were analyzed using linear regression models. The change in mental resources during each block (i.e., mental resource drain) is the difference between resources reported at the end of the block and resources reported at the end of the prior block.

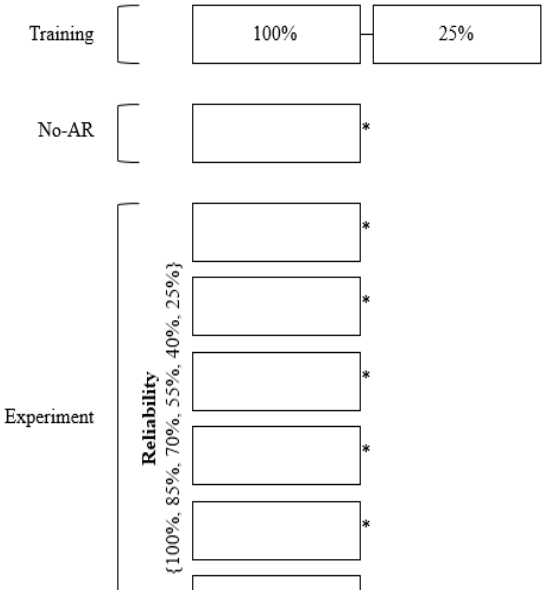

Figure 3. Graphical summary of the study procedure. Rectangles represent trials, curly brackets enclose randomized elements, and asterisks represent survey administrations.

Because we are interested in determining the reliability required to improve performance above a no-AR baseline, all regressions were adjusted using data from the no-AR baseline. To make this adjustment, we conducted an additional mixed-effects regression predicting each of the six outcomes during the no-AR block. These regressions were all empty models, with a nested error structure but no predictors. We extracted the empty model intercepts for each of these regressions by participant, which were then subtracted from the corresponding intercepts in subsequent fully-specified models, producing an adjusted intercept term that represented deviation from the no-AR baseline. These adjusted intercepts were used to calculate an estimated value for each outcome and participant across reliability levels relative to each participant's own no-AR baseline. For a complete description of this procedure, see Monfort et al. (2017).

RESULTS

Our first hypothesis predicted the effect of AR reliability on the various outcomes of interest will depend on the type of error (miss or false alarm) produced by the AR system. Specifically, we hypothesized participants searching for targets with false alarm-prone AR will require a more reliable system compared to participants paired with miss-prone AR. System unreliability should result in more human error when it fails to provide useful information and distracts from the primary task (i.e., intrusive false alarms) compared to when it merely fails to provide useful information (i.e., unhelpful misses). To test this hypothesis, our regression models predicted six outcomes (probability of a miss, probability of a false alarm, search time, trust, mental workload, and change in mental resources) with AR reliability, error type, and a reliability by error type interaction term.

Response Accuracy

As expected, unreliable AR caused participants to miss more targets when the unreliability stemmed from AR false alarms compared to AR misses. That is, participants missed more tanks when they were distracted by erroneously-marked houses compared to when they were faced with erroneously-unmarked tanks. This effect can be observed in Figure 4. Even a relatively small amount of unreliability in false alarm-prone AR made participants miss more targets than they did with no AR at all. In contrast, miss-prone AR never hurt performance: participants were able to compensate for poor AR performance, spotting targets even at very low levels of reliability.

Soldiers often have difficulty identifying targets when those targets are distant, when relevant perceptual cues are small or even imperceptible. We had therefore hypothesized that distant targets would magnify any undesirable effects of unreliable AR, as Soldiers would be less able to leverage their own perceptual abilities under these circumstances to compensate for poor AR performance. Consistent with this hypothesis, participants in our study missed more targets when those targets were distant compared to when they were close. The effect of range also magnified the undesirable effects of false alarm-prone AR. For participants paired with false alarm-prone AR, distant targets were much more

likely to be missed compared to close ones. As depicted in Figure 5, participants were most likely to miss tanks when those tanks were distant and the scene was populated with AR false alarms. Thus, although all of the tanks in the false alarm-prone AR groups were properly marked, the presence of so many false alarms may have interfered with participants' ability to properly scrutinize valid targets, particularly when those targets were distant from the observer.

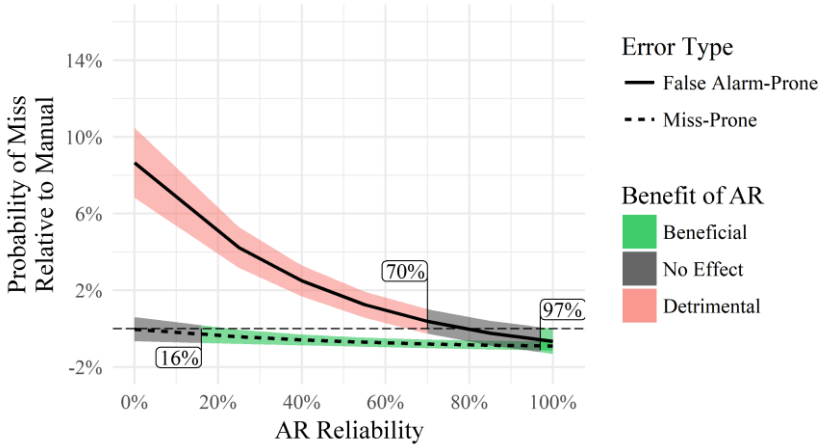

Figure 4. Probability of missing a target by AR reliability and error type. The shaded regions represent the calculated 95% confidence intervals (CI) surrounding the reliability range when joint human-AR performance was worse (red; $p < .05$), neither better/worse (gray; $p > .05$), and/or better (green; $p < .05$) than human-only performance.

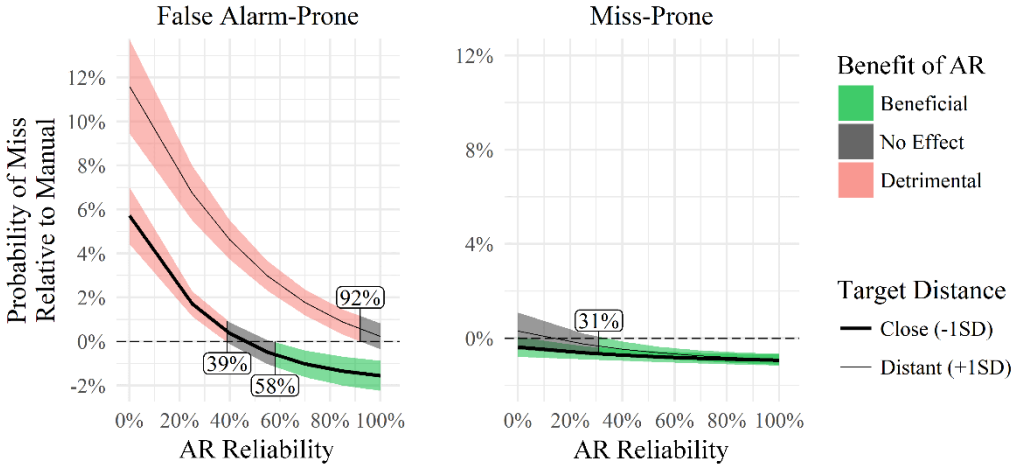

Figure 5. Probability of missing a target by AR reliability, error type, and target range. The shaded regions (95% CI) represent the reliability range when joint human-AR performance was worse (red; $p < .05$), neither better/worse (gray; $p > .05$), and/or better (green; $p < .05$) than human-only performance.

Unreliable, false alarm-prone AR also had a greater effect on participant false alarm rate compared to miss-prone AR. That is, in addition to missing more correct targets, these participants also "hit" more invalid targets (Figure 6). In addition to causing participants to miss more valid targets, increased range also caused them to hit more invalid targets, especially when the AR system incorrectly placed AR tank icons above these distant, invalid targets (Figure 7).

In sum, AR that issued false alarms caused observers to both miss correct targets and to "hit" incorrect targets. In contrast, AR that missed targets did not have a substantial effect on target accuracy. Both of these effects were magnified when targets were distant compared to when they were close.

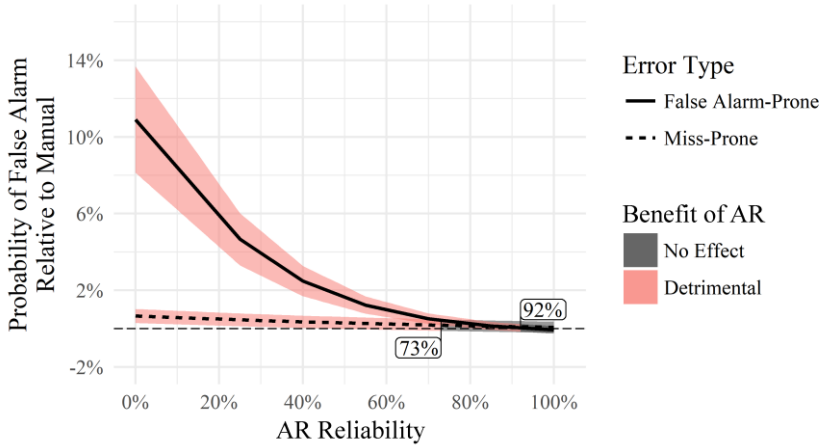

Figure 6. Probability of issuing a false alarm by AR reliability and error type. The shaded regions (95% CI) represent the reliability range when joint human-AR performance was worse (red; $p < .05$) or neither better/worse (gray; $p > .05$) than human-only performance.

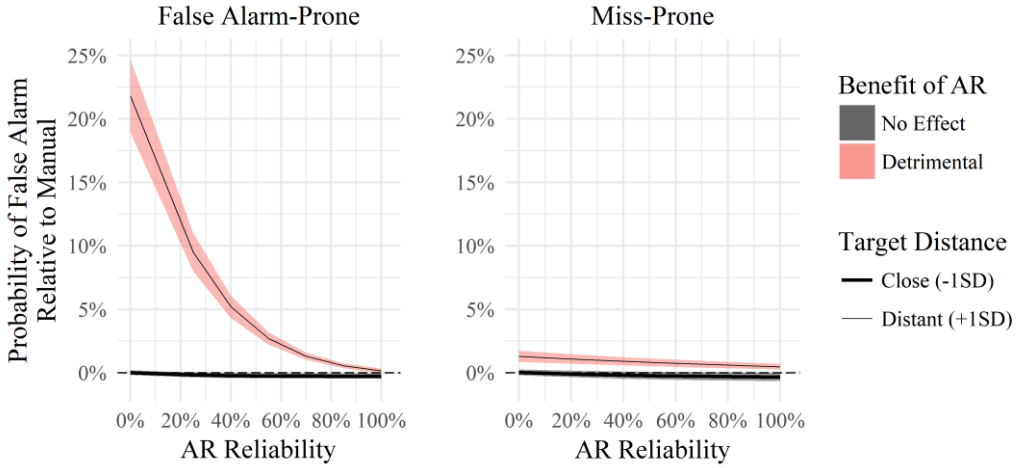

Figure 7. Probability of issuing a false alarm by AR reliability, error type, and range. The shaded regions (95% CI) represent the reliability range when joint human-AR performance was worse (red; $p < .05$) or neither better/worse (gray; $p > .05$) than human-only performance.

Search Time

Overall, participants spent an average of 11.8 seconds (SD = 17.9) searching each scene for targets. As expected, the amount of time spent searching varied by error type (Figure 8). Participants interacting with false alarm-prone AR tended to reduce their time searching each scene as the number of AR false alarms increased. In contrast, participants interacting with miss-prone AR tended to increase their time searching as the number of AR misses increased. The former represents a maladaptive response to worsening AR reliability, and represents a potential mechanism for the previously discussed worsening observer accuracy. That is, participants interacting with AR that injected many false alarms into their visual field responded poorly by less diligently searching the scene. It is possible that the visual clutter from AR false alarms overwhelmed observers, causing them to "give up" rather than persevere (although "giving up" should not be interpreted as a total lack of responses, as participants in the false alarm-prone AR group were more likely to mark additional invalid targets).

Subjective Outcomes

Although participants demonstrated clear (objective) performance differences between error types, fewer such differences arose for the subjective measures of trust, workload, and mental resources. Both false alarm-prone and

miss-prone AR affected workload and trust to a similar extent across reliability levels, and participants in both conditions never trusted the AR more than themselves. However, participants did report greater resource drain after interacting with miss-prone AR compared to false alarm-prone AR (Figure 9). The greater expenditure of cognitive resources reported by participants with miss-prone AR is consistent with the greater amount of time participants searched in the face of AR failure. That is, participants faced with AR that did not mark targets properly increased the amount of time they spent searching and leveraged more cognitive effort, so their performance did not suffer. Participants faced with false alarm-prone AR did *not* increase their search time or cognitive effort, and performed worse as a result.

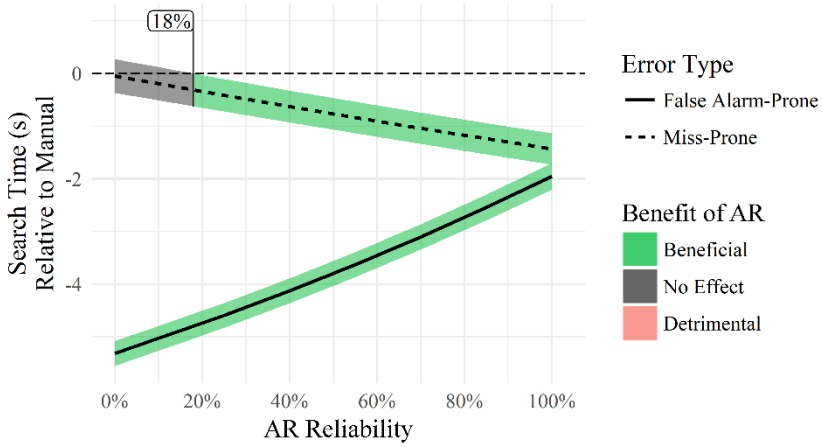

Figure 8. Search time by AR reliability and error type. The shaded regions (95% CI) represent the reliability range when joint human-AR performance was neither better/worse (gray; $p > .05$) or better (green; $p < .05$) than human-only performance.

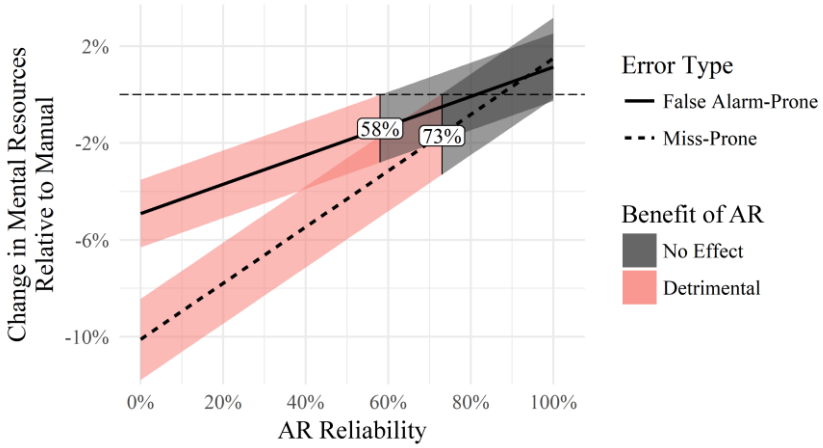

Figure 9. Change in mental resources by AR reliability and error type. The shaded regions (95% CI) represent the reliability range when joint human-AR performance was worse (red; $p < .05$) or neither better/worse (gray; $p > .05$) than human-only performance.

DISCUSSION

The purpose of the current research was to explore variables that might affect observer performance in a visual search task while using an augmented reality (AR) system. Specifically, we sought to test several variables' effects on the reliability required for AR to improve performance above a no-AR baseline. The data show that the minimum reliability required for automation to improve human performance on a visual search task—far from being a single fixed threshold—varies according to context-specific factors (i.e., range to target, error type), and further depends on which outcome of interest is being considered.

AR Error Type

Our data suggest that observer performance is the most sensitive to AR errors when those errors are false alarms. That is, when the AR system mistakenly designates non-targets as targets, observers tend to perform worse than when the AR system fails to designate targets altogether, even at the same overall level of unreliability. This finding is consistent with Dixon et al. (2007), who suggest that because false alarms are salient, intrusive, and annoying, the same level of overall unreliability may be more impactful if AR errors are false alarms than if they are misses.

Interestingly, as reliability decreased in false alarm-prone AR systems, observers became more likely to accept invalid cues (i.e., increased compliance) and less likely to accept valid cues (i.e., decreased reliance). The fact that false alarm-prone AR increased compliance runs contrary to research that suggests frequent false alarms should make participants less likely to heed automated warnings (i.e., the "cry wolf" effect; Wickens et al., 2009). It is possible that the large number of false alarms overwhelmed participants and caused them to disengage from the task, clicking on targets with less deliberate effort. Past research suggests that increasing AR false alarms might aggravate observers (Yeh et al., 2003). Indeed, participants in the present study spent less time with each scene as AR reliability decreased and reported less resource drain after these trials. The response to worsening AR from participants in the false alarm condition suggests that the mechanism for poor performance is related to a premature withdrawal of effort (i.e., the participants "gave up."). Although participants experienced greater resource drain in the miss-prone AR, likely due to their increased search times and active compensation for AR failures, ultimately the finding that participants randomly assigned to false alarm-prone AR either could not compensate for AR mistakes or were unwilling to pay the cognitive cost of compensating, suggests that the false-alarm condition was more difficult.

It is important to question whether this type of task disengagement (caused by false alarm-prone AR) would occur in combat situations, where Soldiers are highly motivated, inevitably more so than the participants of the present study. Many combat operations take place over a long period of time where contact with the enemy is relatively limited, requiring effortful sustained attention. The prolonged application of Soldier effort, even if highly motivated, will invariably result in attentional failure, which will occur sooner with more difficult tasks (i.e., vigilance is hard; Warm, Parasuraman, & Matthews, 2008). Thus, although Soldiers may perform adequately with false alarm-prone AR in the short-term, our research suggests their conscientious persistence will be costlier in terms of the mental effort required to sustain it. As a result, longer-term exposure to false alarm-prone AR should be limited when possible.

System Sensitivity

Whether a sensor is miss-prone or false alarm-prone is a consequence of sensor sensitivity. The sensitivity, or the critical threshold required to trigger a signal to the human operator, is typically chosen by the system designer. However, some sensors currently under development will allow operators to adjust the sensitivity of target detection algorithms (Gans et al., 2015), which will shift the relative prevalence of AR misses and false alarms. The knowledge that AR false alarms are more damaging than misses (even at the same level of overall unreliability) is relevant to determining the amount of freedom to give Soldiers for sensitivity adjustments. Compared to participants paired with miss-prone AR, participants with false alarm-prone AR were more than four times as likely to miss targets and two times as likely to issue false alarms. Nonetheless, trust levels for both groups were essentially similar throughout the experiment, suggesting the consequences of false alarm-prone AR, at least in part, escape conscious awareness. This is not a surprising finding as human operators often poorly calibrate their trust (Dzindolet et al., 2003), and many catastrophic, news-worthy stories about human-automation collaboration involve unreliable automation escaping notice from a human overseer (e.g., Sparaco, 1995). Soldiers may therefore unknowingly increase their risk of error if they increase sensitivity to the point of generating too many false alarms. While operators find false alarms annoying (Yeh et al., 2003), the consequences of missing a target are generally more life-threatening than having to further scrutinize additional false alarms. This may incentivize Soldiers to increase device sensitivity, even though too many false alarms may result in the worst performance for correctly detecting real threats. Although allowing Soldiers to customize sensor sensitivity may improve outcomes by improving the fit between sensor capabilities and user needs (Horvitz, 1999), and adjustable settings may be necessary to use the same equipment for different tasks or in different environments, the data from the current study suggest user-controlled sensitivity adjustments should be constrained by the system, preventative user training, or both to prevent alert over-saturation and user disengagement. Further research exposing participants to both types of AR errors (instead of a single error type) and research investigating operator tendencies with user-controlled sensitivity adjustments should be conducted to confirm these findings.

Target Distance

The capacity for AR to interfere with Soldier performance was further magnified when targets were distant. When AR systems label distant, difficult-to-see objects, the human observer is less able to verify these claims. The "lumberjack effect" (Onnasch et al., 2014) suggests the more support an AR system provides, the greater risk a human operator assumes in the event of AR failure. The AR in the current study provided more support for distant targets, and therefore AR failures resulted in more damage to performance. Sheridan and Parasuraman (2000) suggest that the true risk of automating a task can be calculated by the cost of failure multiplied by the probability of failure incurred by the human-machine pairing. In our study, the probability of human-machine failure increased with range, meaning the reliability required for AR to have a net-positive effect on distant targets was higher (particularly with false alarm-prone AR). Thus, AR systems aiding target search will require higher reliability for distant, more difficult targets in order to improve operator performance. If sufficient reliability cannot be generated for distant targets, designers may wish to consider only automating closer targets while instructing operators to primarily search for distant targets.

Future Research Directions

One limitation of our research is it only accounts for visual performance at one level of environmental difficulty and visual clutter. In addition to range, target size, and other target characteristics, the difficulty of a visual detection task is also heavily influenced by the environment in which the target is placed (e.g., clutter: Rosenholtz, Li & Nakano, 2007). Thus, investigating the extent to which AR reliability requirements change when targets are placed in increasingly difficult search environments would be useful. Although participants in the current study were generally more able to compensate for misses than false alarms, this relationship might invert with more difficult search tasks. Likewise, all targets in this study were completely unobstructed by trees and buildings, but real detection tasks are rarely so simplistic, so future research could examine AR reliability requirements with partially obstructed targets. Future research should also account for factors directly related to specific military technologies not addressed by this simulation, such as head-mounted displays or thermal and night vision sensors. Some research has found meaningful differences in how humans experience information presented on a screen versus a head-mounted display (Thomas & Wickens, 2004; McKendrick et al., 2016). Likewise, target detection at night/using visual information outside the visible spectrum can induce greater stress on human observers (Sterling & Jacobson, 2006). To the extent that such technologies alter demands placed on the visual system, AR reliability requirements may change.

CONCLUSION

The current study was designed to evaluate the level of reliability required for an AR system to assist with a common Soldier visual search task: vehicle detection and identification. Further, we sought to test the degree to which contextual factors (i.e., error type and range to target) related to visual search tasks might alter reliability requirements. Our results suggest the reliability required for AR to improve performance varies considerably depending on these contextual variables. First, AR reliability requirements were highest when AR erroneously marked invalid targets compared to when it missed valid targets. Second, we found that unreliable AR impaired human performance the most when targets were distant and difficult to see. Lastly, we observed a discrepancy between the various performance outcomes; although unreliable AR tended to hurt performance and search speed, it had comparatively small effects on trust and workload. The differential effect of AR reliability on objective and subjective outcomes suggests Soldiers may not be able to gauge AR performance accurately. Soldiers may benefit from training or AR design changes that increase operator awareness of system reliability. While practical constraints will almost always prevent AR from becoming perfectly reliable, designers should ensure AR reliability is high enough that joint human-AR performance exceeds the performance of the human alone. The current study demonstrated the AR reliability level required to elevate joint performance above human-only levels can vary dramatically by context. Rather than relying on a fixed estimate, designers should carefully consider contextual factors when developing AR to pair with a human operator.

ACKNOWLEDGEMENTS

This research was partially funded under contract number W909MY-12-D-0004 from the U.S. Army RDECOM CERDEC Night Vision and Electronic Sensors Directorate (NVESD). NVESD supports the testing and development of next-generation augmented reality displays to increase Soldier effectiveness.

REFERENCES

Armato, S.G., Li, F., Giger, M.L., MacMahon, H., Sone, S., & Doi, K. (2002). Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology*, *225*(3), 685-692.

Bagheri, N. & Jamieson, G.A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency." In D.A. Vicenzi, M. Mouloua, & O.A. Hancock (Eds.), *Human performance, situation awareness, and automation: Current research and trends* (pp. 54-59). Mahwah, NJ: Erlbaum.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). lme4: Linear mixed-effects models using Eigen and S4, 2014. *CRAN*, 1(4).

Biros, D.P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173-189.

Davenport, R.B. & Bustamante, E.A. (2010, September). Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1513-1517.

Deng, J., Han, R., & Mishra, S. (2003). *Enhancing base station security in wireless sensor networks*. Technical Report CU-CS-951-03, Department of Computer Science, University of Colorado.

de Visser, E.J. & 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*(2), 209-231.

Dixon, S.R. & Wickens, C.D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, 48(3), 474-486.

Dixon, S.R., Wickens, C.D., & McCarley, J.S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*(4), 564-572.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697-718.

Gans, E., Roberts, D., Bennett, M., Towles, H., Menozzi, A., Cook, J., & Sherrill, T. (2015, May). Augmented reality technology for day/night situational awareness for the dismounted soldier. In *SPIE Defense + Security*, 1-11.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139-183.

Horvitz, E. (1999, May). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159-166.

Madhavan, P., Wiegmann, D.A., & Lacson, F.C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241-256.

McDowell, E. D. (1992). *Automatic Target Recognition Display Format Study* (No. NAWC-WPNS-TP-8072). Naval Air Warfare Center Weapons Division, China Lake, CA.

McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., & Ayaz, H. (2016). Into the wild: Neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy. *Frontiers in Human Neuroscience*, *10*(216), 1-15.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors*, *43*, 563-572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*, 196-204.

Monfort, S.S., Graybeal, J.J., Harwood, A., McKnight, P.E., & Shaw, T. (2017). A single-item assessment for remaining mental resources: Development and validation of the Gas Tank Questionnaire (GTQ). *Theoretical Issues in Ergonomics Science*, 1-23.

Nothdurft, H.C. (1992). Feature analysis and the role of similarity in preattentive vision. *Attention, Perception & Psychophysics*, *52*(4), 355-375.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Onnasch, L., Wickens, C.D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476-488.

Parasuraman, R. & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *Human Factors*, 52, 381-410.

Parasuraman, R., Molloy, R., & Singh, I.L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, *3*, 1-23.

Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of vision*, *7*(2), 17-17.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76-87.

Rovira, E. & Parasuraman, R. (2010). Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance. *Human Factors*, *52*(3), 411-425.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110-114.

Sheridan, T.B. & Parasuraman, R. (2000). Human versus automation in responding to failures: An expected-value analysis. *Human Factors*, *42*(3), 403-407.

Sterling, B.S. & Jacobson, C. N. (2006). *A Human Factors Analysis of Aided Target Recognition Technology* (No. ARL-TR-3959). Army Research Laboratory Aberdeen Proving Ground, MD.

Sparaco, P. (1995). Airbus seeks to keep pilot, new technology in harmony. *Aviation Week and Space Technology*, 30, 62-63.

Thomas L.C. & Wickens C.D. (2001, October). Visual displays and cognitive tunneling: Frames of reference effects on spatial judgments and change detection. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(4), 336–340.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, & Image Processing*, *31*, 156-177.

Warm, J.S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433-441.

Wickens, C.D., Rice, S., Keller, D., Hutchins, S., Hughes, J., & Clayton, K. (2009). False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect? *Human Factors*, *51*(4), 446-462.

Wiener, E.L. & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*(10), 995-1011.

Yeh, M. & Wickens, C.D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, *43*(3), 355-365.

Yeh, M., Merlo, J.L., Wickens, C.D., & Brandenburg, D.L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors*, *45*(3), 390-407.