INL/CON-12-26093
PREPRINT

Alarm Strategy and Complexity: Predictions of Operator Response

Nuclear Power Instrumentation and Control and Human Machine Interface Technology

Austin Ragsdale
Roger Lew
Brian Dyre
Ronald Boring
David Gertman

July 2012

ALARM STRATEGY AND COMPLEXITY: PREDICTIONS OF OPERATOR RESPONSE

Austin Ragsdale, Roger Lew, Brian Dyre
University of Idaho
Moscow, Idaho, USA
rags7638@vandals.uidaho.edu; rogerlew@vandals.uidaho.edu; bdyre@uidaho.edu

Ronald Boring, David Gertman
Idaho National Laboratory
Idaho Falls, Idaho, USA
ronald.boring@inl.gov; david.gertman@inl.gov

ABSTRACT

Decision support for operators is not new, and much has been written regarding the potential usefulness of digital support systems and alarm filtering strategies. However, determining the appropriate characteristics of decision support tools is difficult, especially when alarms can vary in the manner which diagnostic information is formulated and displayed and when event scenario types are complex and numerous. When first reviewed, the advantages or disadvantages of a particular alarm approach may not be apparent to the designer or analyst. The present research focuses on the review of two particular alarm strategies, binary alarm type (BAT) and likelihood alarm type (LAT), and reviews their influence upon accuracy, bias, and trust for tasks performed at a computer workstation capable of replicating a series of control-room-like alarms. The findings are discussed in terms of the of the performance advantages of likelihood alarm technology and related research as an aid to the alarm design process.

Key Words: likelihood alarm theory, human factors, operator performance, automation, cognition, complexity

1 INTRODUCTION

The implementation of digital support systems in control rooms across industries is becoming a commonplace occurrence. While the temptation to present as much information as possible is great, care needs to be taken in terms of the structuring, i.e., filtering, and presentation of the numerous alarms available to the operator. As a result of the complexity of the underlying plant processes and the availability of sophisticated information technologies, different strategies for alarm presentation exist. At present, there is need to include empirically determined findings and insights from human factors in the design of these systems. The nuclear industry has sought to guide operators to changes in key parameters and processes through the design of procedures that support directing operators when and where to check for key indications including alarms. This process is aided by the alarms themselves being highly salient. The introduction of automation including digital support tools is not without its concerns. For a review of some of the challenges associated with introducing automation see, for example, Parasuramman et al. [1], O'hara and Gunther [2], and Boring and Gertman [3]. Finding the appropriate level of diagnosis and automation has followed approaches outlined by [1], who determined principles by partitioning components of automation into levels and stages. Alarms work to cue operator response and help with information extraction from the immediate environment. The necessity for these alarms to be salient has been established by Woods [4] and others.

The issues of alarm presentation and filtering have challenges in a dynamic control room. Events may not always occur in isolation, and enhanced diagnostic information and display strategies may need to consider the potential for one event to mask another. Further, we note that alarms may be valid or not. That is, there is some probability that an alarm is actually true. Getty et al. [5] note that the accuracy of the alarm, its threshold, and the probability of the event itself are all factors. Total alarm response accuracy can be reduced by having fewer true alarms. This can happen due to such factors as having a low threshold that results in alarming during normal conditions such that it is viewed as a nuisance. In these situations, it is not unusual for operators to lose trust in alarms, the result of which is a delay in operator response. As Bustamante [6] has noted, adjusting the alarm threshold upward can result in decreasing critical event detection. The effects of false alarms as well as missed alarms can be expensive in the consideration of the implementation of likelihood alarm technology.

1.1 Binary vs. Likelihood Alarm Technology

Binary alarm types (BAT), once an alarm threshold is crossed, orient operator attention by turning an alarm on. If a threshold is not crossed, then the alarm is off. This class of alarms is somewhat error prone due to threshold and reliability issues. Findings in Bustamante and Clark [7] suggest that accompanying alarms with diagnostic information gathered from multiple alarm advisories may be a more fruitful approach. This approach is called the likelihood alarm type (LAT). Aside from providing an ok, warning, or alarm state, they can have multiple thresholds that are mapped to probability space. Mapping fundamentals include two principles, positive predictive value (PPV), and urgency mapping. Urgency mapped alarms are associated with increased operator tendency to respond [8]. Earlier research by Bustamante [6] has shown improved operator attention allocation and appropriate decision making in the presence of these conditions [9]. In this paper, we address findings that contribute to our understanding of the probability matching process.

1.2 Trust

By understanding the connection between compliance, reliance, and trust, in relation to the reliability of alarms, a better understanding of how alarm reliability affects the operator-technology interaction may be achieved. Trust in automation has generally focused on single automated aids; however, Keller and Rice [10] have recently evaluated trust when operators are interacting with multiple automated aids. The significance of this research lies in determining if operator trust is based within each component of the automated system or if it is based on a more general system-wide view. Keller and Rice found that when operators interact with multiple automated aids with differing PPVs, they appear to develop system-wide trust rather than component-specific trust in the automated system. For a definitive paper on *trust* and a number of human performance findings see the work of Lee and See [11].

1.3 Matching Behavior

Operators tend to probability match to a low PPV and generalize when the PPV is high. Thus, operators tend to inefficiently interact with multiple automated aids when the PPV of those aids differ.

1.4 Additional Factors

Proximity has a role in trust as well. Subject response to alarms and displays may be influenced by the PPV of adjacent displays. For example, Keller and Rice [10] found that this influenced operator sensitivity to alarms but, not operator bias.

2 METHOD

2.1 Overview

Participants joined in a study that evaluated measures of trust, bias, accuracy, and for a design that evaluated two approaches to alarm presentation (binary case, likelihood alarm case) for two events. EVENT A was more apparent, and EVENT B was capable of masking some conditions of EVENT A. Trust was measured through a 3-indicator self report that reviewed the participants' perception of alarm reliability, dependability, and trustworthiness, which has been found to possess high internal consistency [12-13]. Accuracy was defined in terms of alarm acknowledgement in response to the alarm. This represented a combined "acknowledge" and "action" score described subsequently in this paper. It was predicted that accuracy and trust would be higher and bias lower with the LAT presentation style of alarm informing. An interaction effect was also hypothesized, whereby trust would be highest in the EVENT B–LAT condition and lowest in the EVENT B-BAT condition. Sensitivity was defined in terms of signal detection theory. A number of additional minor hypotheses were evaluated and summarized but are to be reported in-depth in subsequent publications.

2.2 Loading Task

A loading task was used to relate the alarms to a process and to raise workload to a realistic level required participants to operate DURESS [14], which engaged participants in process control tasks constrained by physical laws and functions that govern operational control possibilities.

2.3 Participants

Sixty university students (26 females and 34 males) participated in this experiment. Participants ranged from 18 to 41 years of age (M = 21.86, SD = 4.01). All participants reported having required normal or corrected-to-normal vision and hearing. As an incentive, participants were compensated with two research credits. Throughout the experiment, researchers treated participants according to the American Psychological Association Ethical Guidelines.

2.4 Materials and Apparatus

There were a total of four testing stations; each station was equipped with one computer, mouse, keyboard, and two monitors. The monitors were connected to the computer so that participants could move the mouse pointer between both monitors. For all trials, the simulated environments were rendered using DURESS (Vicente & Pawlak, 1994) and NPC software (Lew, 2011) running on Microsoft Windows XP. The display for the process control task was viewed on a 19-inch diagonal Samsung LCD monitor at a spatial resolution of 1280 x 1024 pixels (H x V) and a refresh rate of 105 Hz, and was placed directly in front of participants. The display for the alarm monitoring task was viewed on a 19-inch diagonal Hanns-G LCD monitor at a spatial resolution of 1440 x 900 pixels (H x V) and a refresh rate of 105 Hz, and was placed just left of the center monitor at a 45-degree angle. The rational for this configuration was to simulate real world conditions where there is often separation among a primary process display and various alarm presentation screens.

2.5 Tasking

The participants' job was to interact with a simulated simplified nuclear process control task—the loading task—to monitor and maintain an optimal water level, flow, and temperature by manipulating various valves and heaters. The initial water flow started at the left side of the display and had a predetermined temperature and flow rate. This inflow of water moved from left to right and split into two main valves labeled 'VA' and 'VB'. Additionally, 'VA' split into two valves labeled 'VA1' and 'VA2', whereas 'VB' split into two valves labeled 'VB1' and 'VB2'. All six valves could be opened or closed on a continuous scale ranging from 0 to 20. Eventually, water flow from these valves entered into one of two reservoirs. Both of these reservoirs hold anywhere between 0 and 100 units of water, and are labeled Reservoir 1 and Reservoir 2. Beneath both reservoirs were corresponding heaters, one for each reservoir. Reservoir 1's heater was labeled 'H1' whereas Reservoir 2's heater was labeled 'H2'. These heaters controlled water temperature within the corresponding reservoir and could be adjusted on a continuous scale ranging from 0 through 1. On the far right of both reservoirs was a temperature gauge. Each gauge corresponded to the reservoir to the right of it, and displays reservoir water temperature. These gauges were labeled 'T1' or 'T2', ranged from 0 to 50, and displayed the current reservoir temperature in red within the temperature scale (like a thermometer). Just right of both reservoirs were water release valves labeled 'VO1' and 'VO2'. These valves controlled the amount of water flowing out of the corresponding reservoir and could be opened or closed on a continuous scale of 0 through 20.

The secondary task simulated a crucial function that nuclear power plant (NPP) operators perform when critical events arise; that is, monitoring alarms, diagnosing problems, and taking corrective action when necessary. Participants performed this task through a simulated NPP alarm system called 'NPC' (Lew, 2011) that creates an environment where participants monitor and interact with a variety of alarm systems that may be present during steam breaks and recirculating events.

The simulated alarm systems were equipped with either BAT or LAT. A system equipped with BAT emitted one of two types of advisories (i.e., *OK* or *ALARM*) while a system equipped with LAT emitted one of three types of advisories (i.e., *OK*, *WARNING*, or *ALARM*). An auditory stimulus was presented for each alarm at 65 dB(A) to ensure the alarm was salient enough for participants to easily detect. Participants monitoring alarm advisories represent the first processing stage (attention). After receiving an alarm advisory, participants either ignored the advisory and continued performing the primary task or they acknowledged the advisory by clicking on the corresponding EVENT A button for the alarm in question. The EVENT A buttons were labeled *X EVENT A 1*, *X EVENT A 2*, and *X EVENT A 3*, and were located below the main control panel. Participants had up to five seconds to acknowledge alarm advisories and an additional 5 seconds, once alarm advisories were acknowledged, to monitor critical task information and decide if corrective action was necessary. Monitoring task critical information was accomplished by going to an additional window.

Monitoring critical task information and deciding if corrective action was necessary represent the second processing stage (decision making). When participants monitored Event A, they evaluated three main steamline pressure gauges to determine if a simulated steamline break had occurred (attention stage) and if main steamline isolation valves (MSIVs) should be closed (decision stage). When steamlines were in a non-critical state, gauges displayed mid level pressure and were marked in green. The occurrence of EVENT A was indicated by a low pressure gauge displayed in red. Once participants determined if corrective action was needed, they clicked on the *EVENT A Critical* or *EVENT A OK* button for the major steamline for which the alarm sounded. Consequently, once participants finished interacting with the EVENT A event, they shortly, 10 seconds later, received an alarm advisory for the EVENT B event. Thus, a fairly rapid pace was presented to subjects.

Participants acknowledged EVENT B alarm advisories the same way they did for the EVENT A event. If participants decided to acknowledge EVENT B alarm advisories and perform monitoring, they evaluated secondary radiation indicators for three steam generators. When the EVENT B was in a non-critical state, gauges displayed mid levels of leakage and were marked in green. When an EVENT B occurred, the leakage indicator for the steam generator in question was elevated above that of the other steam generators and was marked red, the caveat being that leakage indicators were misleading at times. That is, when EVENT A and EVENT B occur concurrently, the leakage indicator in question showed the leakage gauge as normal, giving the appearance that the EVENT B did not occur when it actually did. The occurrence of EVENT A was expected to contribute to lower performance in the EVENT B–BAT condition. Participants then had 10 seconds to interact with the EVENT A system (5 seconds to acknowledge the alarm and 5 seconds to monitor critical task information/take corrective action). Once the 10 seconds for interacting with the EVENT A system were up, there was a space of 10 seconds before participants received an EVENT B alarm advisory. Participants then had 10 seconds to interact with the EVENT B system (5 seconds to acknowledge the alarm and 5 seconds to monitor/take corrective action). Once the 10 seconds for interacting with the EVENT B system were up, there was a 10 second pause before participants received the next EVENT A alarm advisory. This cycle continued 40 times so that participants received 40 alarm advisories for EVENT A and 40 alarm advisories for EVENT B.

Note that EVENT A approximated a main steamline break (MSB), and EVENT B approximated a steam generator tube rupture (SGTR). It must be noted that the occurrence of these events greatly exceeded the frequency of such events that would be expected in actual operations. The participants performing the task were not skilled operators, and the simulated environment was designed to elicit responses at a much higher rate than would be possible in a simulation of higher fidelity.

2.6 Positive Predictive Value

Both alarm systems, regardless of technology type, had the same parameters; both EVENT A and EVENT B alarm systems had a positive predictive value (PPV) of .5 (i.e., a total of 20 hits and 20 false alarms during the test trial). However, the alarm advisories had different PPVs when comparing them across the different alarm technology types. Regarding the PPV of BAT alarm advisories, *OK* advisories had a PPV of 1 and *ALARM* advisories had a PPV of .5. Regarding the PPV of LAT alarm advisories, *OK* advisories had a PPV of 1, *WARNING* advisories had a PPV of .1, and *ALARM* advisories had a PPV of .9, suggesting that the presence of the advisory system had somehow helped to raise confidence in or perhaps increase the overall situation awareness of the participants.

Training to the system was established, and proficiency in alarm acknowledgement and screen navigation was achieved within one hour. A combination of PowerPoint slides and narrative on an audio loop were used to reduce the potential for experimenter bias. The DURESS model was configured so that EVENT A and EVENT B were easily conceptualized as two types of breaks, the first a break in a large feed line, the second in a recirculation system.

3 RESULTS

3.1 Overview

Although a more complex design was employed and is presented in a larger report, in this paper we report on a 2x2 factorial design where participants responded to Events (A, B) and alarm type (BAT, LAT). Additionally, there was a manipulation where the underlying alarm was either true or false. Participants were randomly assigned to one of these four conditions:

1. EVENT A-BAT + EVENT B-BAT,
2. EVENT A-LAT + EVENT B-LAT,
3. EVENT A-LAT + EVENT B-BAT, or
4. EVENT A-BAT + EVENT B-LAT.

Additionally, the state of EVENT A or B upon accuracy of the other was also calculated. The overall design was mixed, whereby participants were randomly assigned to conditions but participated in a number of measures. Additionally for ease of understanding, we analyze these data for cognitive processing Stage 1 and 2—attention and decision—based upon Wickens' cognitivie processing model [15]. Another paper is planned that discusses findings in terms of cognitive functions.

3.2 Trust and Technology Type

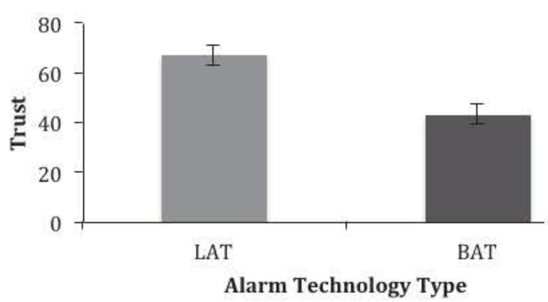
Results for trust in EVENT A and EVENT B are summarized in Table I. For EVENT A, there was a significant main effect for technology type: $F(1, 56) = 20.19$, $p < .05$, partial $\eta^2 = .27$, observed power = .99. On average, trust was greater in the LAT condition ($M = 66.83$, $SE = 3.78$) than the BAT condition ($M = 42.88$, $SE = 3.78$). Note that the standard error (SE) is reported, because it is the standard deviation of a sampling distribution.

Results for trust in the EVENT B alarm system showed a significant main effect for technology type: $F(1,56) = 22.36$, $p < .05$, partial $\eta^2 = .29$, observed power = .99. On average, trust was greater in the LAT condition ($M = 62.66$, $SE = 4.11$) than the BAT condition ($M = 35.19$, $SE = 4.11$).

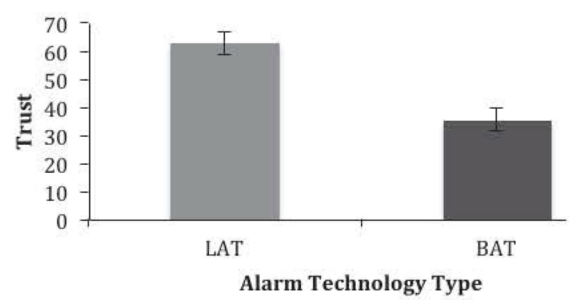
There was a trend for EVENT B-BAT trust scores to be low, which we feel reflects those trials where EVENT A and EVENT B occurred concurrently, and as a result, EVENT B was more difficult to detect. This was most pronounced for the BAT condition, where additional diagnostic information was not present. Figure 1 presents trust scores for event types and LAT and BAT alarm presentation conditions.

Table I. Mean trust scores for event type by alarm technology.

Event Type	Likelihood Alarm Type	Binary Alarm Type
Event A	66.8 (SE = 3.78)	42.9 (SE = 3.78)
Event B	62.7 (SE = 4.11)	35.2 (SE = 4.11)



EVENT A



EVENT B

Figure 1. Trust in alarm display type for EVENT A and B.

3.3 Relationship Between Accuracy, Bias, and Trust

Regarding overall accuracy (aggregated at both EVENT A and EVENT B) and overall trust, results showed a statistically significant positive correlation, $r(60) = .50$, $p < .05$. That is, as overall-accuracy increased, overall trust increased. Regarding overall bias and overall trust, results showed no statistically significant correlation between these two constructs.

3.4 Notable Interactions

Accuracy scores reflected the influence of adjacent displays. EVENT B-LAT condition accuracy scores were significantly higher when the EVENT A had LAT compared to BAT alarm display presented. This explanation is somewhat consistent with the Keller and Rice (2009) findings that operators tend to probability match to an aid with a low PPV while generalizing it to an aid with a higher PPV. This can be seen in Figure 2 below.

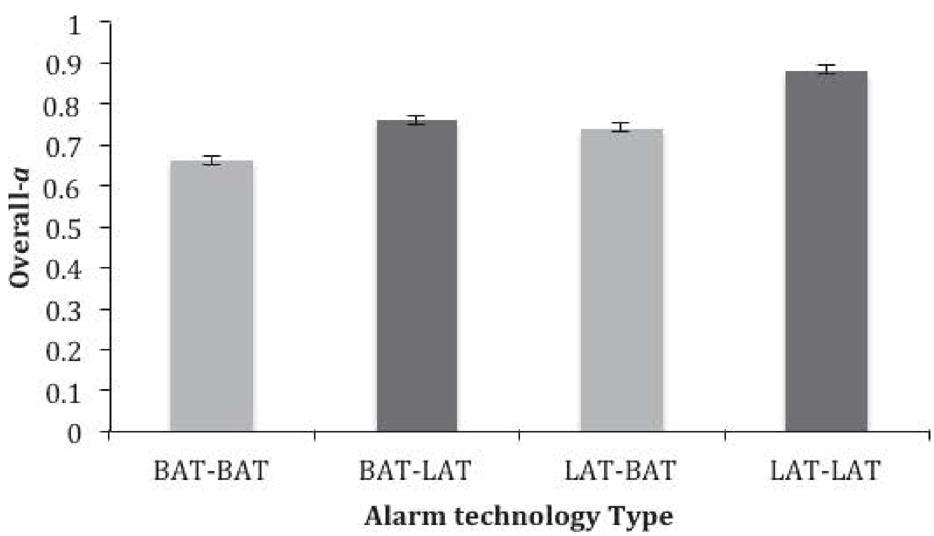


Figure 2. Overall accuracy across all four alarm technology type configurations. (Error Bars represent ± 1 standard error of the mean)

3.5 Summary

Support was found for the assumption that different performance would be observed with the use of binary alarms when compared to likelihood alarms, i.e., likelihood alarms were associated with greater participant trust and accuracy. There was also support for arguments that the reliability and trust in adjacent displays can have an overall influence on participant response to a particular display of interest. Further, concurrent events where the one event could mask some of the indication of the other event were associated with lower scores. Bias was highest for binary type alarm displays.

Overall, trust was lower for situations where participants did not have additional diagnostic information available that could cue them that indicators were misleading. Performance accuracy was significantly lowest for conditions where response to both events was presented in binary form. Also, trust was lowest for these conditions.

4 DISCUSSION

Overall, for a process control task, the hypothesis that alarm type, whether binary or likelihood, is related to operator accuracy and trust was supported. Results for bias were less clear. In the present study, the overall reliability of adjacent alarms influenced subsequent performance. The finding that participants had greater trust in likelihood displays suggests that further research in this area should be conducted. The LAT also proved to be effective for conditions where indication may have been misleading. Variations on this research could inspect the differences found when the number of misses is increased, that is, changes to the PPV.

This study is not without limitations. First, future research could be done within high fidelity simulators and with licensed plant operators. Thus, the true extrapolation of these findings to control room environs could be assessed. Next, a scenario could be selected that shares greater fidelity with plant operations and where the alarms selected are the same alarms associated with a real plant process. The DURESS model used in the present study is not nearly as complex as many of the nuclear plant processes present in today's control rooms. As part of this research, it would be valuable to see how these influences may contribute to the quality and timeliness of crew decision making, for example do LAT presentations take longer to process but offer increased operator response, or is processing and response times largely the same?

These results are also important for human reliability analysis (HRA). Human error probabilities in HRA methods assume a base failure rate that encompasses a wide variety of failures. These errors are then adjusted by manipulation of performance shaping factors [16]. The influence of the reliability of adjacent displays on response accuracy and differences between BAT and LAT is not well characterized in HRA, and data of this sort can be useful when included in overall characterization and quantification. The nuance of misleading indicators in combination of alarm technology type is largely uncharacterized but could provide useful findings and directions for control room design including upgrades and guidelines for implementation.

The findings from this research carry significant theoretical implications and practical applications as they apply to improving efficiency and effectiveness in operational settings and aid in our understanding and characterization of operator response.

5 ACKNOWLEGEMENTS

This research was originally conceived and carried out while the first author was a graduate student under the supervision of Professor Ernesto Bustamante, now deceased. We are grateful for Professor Bustamante's assistance on the design of the study as well as his deep insights on likelihood alarm types. Portions of this paper were completed under a grant from the Center for Advanced Energy Studies at Idaho National Laboratory to the Department of Psychology at the University of Idaho. The Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517.

6 REFERENCES

1. Parasuraman, R., Sheridan, T., & Wickens, C. (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(3), 286-297.

2. O'Hara, J., & Gunther, W. (2010) The Effects of Degraded Digital Istrumentaion and Control Systems on Human System Interfaces and Operator Performance: HFE Review Gudiance, BNL Report 91047, Brookhaven National Laboratory, Brookhaven NY.

3. Boring, R. L., & Gertman, D. I. (2012). Human Reliability Analysis for Small Modular Reactors, Probabilistic Safety Analysis and Management 11 and ESREL 12, Helsinki Finn, in Press.

4. Woods, D. D. (1995). The alarm problem and direct attention in dynamic fault management. Ergonomics, 38(11), 2371-2393.

5. Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. Journal of Experimental Psychology: Applied, 1, 19-33.

6. Bustamante, E. A. (2007). Using Likelihood alert technology in cockpit displays of traffic information to support free flight. In Proceedings of the 14th International Symposium on Aviation Psychology (pp. 100-102). Dayton, OH.

7. Bustamante, E. A., & Clark, R. M. (in Press). Differential effects of likelihood alarm technology, type of automation, and type of task on decision making as applied to aviation and UAV operations.

8. Edworthy, J., & Loxely, S. (1991). Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. Human Factors, 33(2), 205-231.

9. Clark, R. M., Ingebritsen, A., & Bustamante, E. A. (in press). Effects of likelihood alarm technology and false-alarm vs. miss-prone automation on decision-making accuracy and bias. Human Factors.

10. Keller, D., & Rice, S. (2009). System-wide versus component-specific trust using multiple aids. In Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting. San Antonio, TX: Human Factors and Ergonomics Society.

11. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46 (1), 50-80.

12. Bustamante, E. A. (2009). A reexamination of the mediating effect of trust among alarm systems' characteristics and human compliance and reliance. In Proceedings of the 53th International Symposium on Aviation Psychology. San Antonio, TX.

13. Stanton, N., Ragsdale, S. A., Bustamante, E. A. (2009). The effects of system technology and probability type on trust, compliance, and reliance. In Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting (pp.1368-1372). San Antonio, TX: Human Factors and Ergonomics Society.

14. Vicente, K. J., & Pawlak, W. S. (1994). Cognitive work analysis of the DURESS system. Unpublished manuscript, University of Toronto, Toronto, ON.

15. Wickens, C. D. (1987). Information processing, decision-making, and cognition. In G. Salvendy (Ed.), Handbook of human factors.

16. Gertman et al., (1995). SPAR-H Human Reliability Analysis, NUREG/CR-6863, US Nuclear Regulatory Commission, Washington DC.