Collective Mistrust of Alarms

James P. Bliss, Susan Sidone, & Holly Mason

Old Dominion University
Norfolk, VA 23529

Abstract

Recently, alarm systems have become more sensitive and ubiquitous. Unfortunately, sensitive alarm systems may produce greater numbers of false alarms, lowering an operator's level of trust and degrading task performance. In the past, researchers have considered only situations where individuals react to alarms. Because of the frequency and variability of teamed alarm reaction scenarios, we investigated the reactions of independent and dependent teams to collateral marginally reliable alarms. Based on prior literature, we expected dependent teams to show slower but more appropriate alarm reactions and poorer ongoing task performances. Eighty general psychology students (40 two-person teams) independently or dependently performed a psychomotor task while reacting to two alarm systems; one that was 80% reliable, and one that was 40%, 60%, or 80% reliable. Participants responded more frequently to alarms of higher reliability, and less appropriately to those of medium reliability. Generally, dependent teams made more appropriate alarm reactions. Our results suggest that designers and trainers should promote team interdependence and communication when operators are faced with marginally reliable signals.

Introduction

Since the early 1980s researchers have investigated alarm mistrust during complex tasks such as aviation (Bliss, 1997), mining (Mallett, Vaught, & Brnich, 1992), ship handling (Kerstholt, Passenier, Houttuin, & Schuffel (1996) and automobile driving (Nohre, MacKinnon, & Sadalla, 1998). Generally, they have found that participants responded slower, less frequently, and less appropriately to less reliable alarms. In certain cases, they also showed degraded performance on the ongoing task. However, researchers have considered only those situations where individual operators are responsible for reacting to alarms. They have not studied the impact of marginally reliable alarm signals on teams of operators.

Teamed alarm responding is common in aviation and other complex task environments. In critical care units, nuclear power plants, surgical theatres, and air traffic control centers, task operators frequently share the responsibility for reacting to emergency signals. Such cooperation may include sharing information, troubleshooting systems, deciding about signal priority, and allocating responsibility for system response functions. However, the degree to which team members function interdependently varies with the task and the environment (Thompson, 1967). In aviation, alarms during high workload activities such as flight planning, takeoff, and landing typically require the flight crew to coordinate reactions. In cruise flight, however, it is not uncommon for flight crew members to react independently. Bliss, Bowens, Krefting, Byler, and Gibson (in press) recently showed that dependent teams react to alarms more appropriately, but more slowly.

One goal of the current research was to replicate the findings of Bliss et al. (in press) by investigating the reactions of dependent and independent teams to alarm signals of various reliability levels. This experiment supplements earlier work by introducing collateral alarms. Alarm task interdependence and pressure alarm reliability were manipulated using a 2 X 3 mixed design, where interdependence was manipulated between groups and pressure alarm reliability within groups. In the experiment, dyads responded to two separate alarm systems. One of those systems (temperature alarms) had a fixed alarm reliability of 80% true alarms. The other system's reliability fluctuated; pressure alarms had a reliability of 40%, 60% or 80% true alarms.

Method

Design

Alarm task interdependence and alarm reliability were manipulated using a 2 X 3 mixed design. Interdependence was manipulated between two groups. Independent team members required no interaction to react appropriately to the alarms. Dependent team members, however, required interaction to react appropriately. Pressure alarm system reliability (40%, 60%, and 80% true alarms within each session) was manipulated within groups. Teams experienced all reliability levels over three sequential task sessions. Temperature alarm system reliability remained constant at 80%.

Participants experienced both pressure and temperature alarms in each session.

Ongoing task measures included gauge monitoring accuracy and dual-axis tracking accuracy. Alarm reaction measures included speed to react (in seconds), appropriateness of reactions (responding to true alarms and canceling false alarms were appropriate reactions), and response frequency (the percentage of alarms to which participants responded within each experimental session).

Participants

Eighty undergraduate and graduate students (40 two-person teams) from Old Dominion University participated in this study for course credit. The ages of the participants ranged from 18 to 43 years. There were 10 same-sex, dependent teams (1 male, 9 female), 11 same-sex, independent teams (2 male, 9 female), 10 different-sex, dependent teams and 9 different-sex, independent teams. A ten-dollar performance bonus was promised to the team with the highest score on the primary and alarm tasks.

Materials

The Multi-Attribute Task (MAT) battery (Comstock & Arnegard, 1992) was used as the ongoing experimental task. The MAT battery is a microcomputer-based task designed to simulate the demands required by piloting aircraft. It measures cognitive and spatial abilities through dual-axis compensatory tracking, gauge monitoring, and resource management tasks. The continuous compensatory tracking task is particularly suitable for measuring operator attention shifts in multiple-task situations. The MAT battery was independently presented to both team members on IBM compatible 486 computers, using 14" color VGA monitors. The participants used the mice and keyboards to make responses.

Participants performed the MAT tasks back-to-back while auditory and visual alarms were presented using a Macintosh PowerMac 603 and a 14" VGA monitor. The alarm stimuli were modified fire bells digitized from a Boeing 757/767 simulator. The Macintosh was positioned ninety degrees to the side, relative to the primary task computers. When a temperature alarm activated, participants determined whether the MAT gauges TEMP1 and TEMP2 were out of tolerance. If both TEMP1 and TEMP2 were out of tolerance, the alarm was true and required participants to hit the Macintosh F12 key (marked "R" for "RESPOND") and to reset the alarm (in that order). If none or only one of the TEMP1 and TEMP2 gauges were out of tolerance, the alarm was false and required participants to hit the Macintosh F9 key (marked "C" for "CANCEL") and resume the primary tracking task. For pressure alarms, the procedure was similar, except that participants monitored the PRES1 and PRES 2 gauges.

Interdependent team members were required to communicate to determine alarm validity, because one team member monitored TEMP1 and PRES1 and the other monitored TEMP2 and PRES2. Independent team members monitored all gauges, and so were not required to communicate to react appropriately. The alarm stimuli were presented at 60 dB(A) (ambient sound level was 45 dB(A)).

Procedure

Participants completed an Informed Consent Form prior to participating. Then the experimenters carefully presented detailed experimental instructions to the participants. Independent team members were told that they had all of the necessary information on their primary task (MAT) screens to make responses to the alarms and did not necessarily have to communicate with each other. Dependent team members were told to communicate with their teammates to determine the validity of each alarm.

After the initial instructions, participants received familiarization on each element of the primary task (MAT). They practiced tracking, monitoring, and managing resources during two 120-second sessions. Participants also received familiarization on the alarms, as well as instructions about how to respond to them. Following primary and alarm task familiarization, participants completed a joint 200-second practice session, during which they completed the MAT task while responding to alarms.

After the practice sessions, participants began the first of three experimental sessions, separated by 5-minute breaks. Ten alarms were presented during each session. The reliability, or true alarm rate, of pressure alarms during each session was 40%, 60%, or 80% (randomly counterbalanced). The reliability of the temperature alarms during each session was 80%. The reliability of both alarm systems was told to the participants before they began each session. The appropriateness of reactions was reflected by a team score, present at all times on the Macintosh screen. Appropriate alarm reactions increased the score and inappropriate reactions decreased the score. After completing three experimental sessions, participants were debriefed and dismissed.

Results

We calculated several 2 X 3 mixed ANOVAs to test our hypotheses. There was no interaction between team interdependence and pressure alarm reliability level for alarm response frequency; and no main effect of interdependence. However, there was a reliability main effect, $F(2,76)=64.398$, $p<.001$. Trend analyses indicated that response frequency increased with alarm reliability in a linear fashion, $F(1,38)=129.600$, $p<.001$.

There was a significant interaction between interdependence and reliability for alarm reaction appropriateness, $F(2,76)=10.193$, $p<.001$. We also found a main effect for interdependence, suggesting that dependent teams made more appropriate reactions to alarms, $F(1, 38)= 4.000$, $p=.05$. The main effect for reliability was also significant, $F(2,76)=12.135$, $p<.001$. Further tests indicated a quadratic trend, with participants showing less appropriate reactions to alarms that were 60% reliable, $F(1,38)=19,563$, $p<.001$.

Although there was no significant interaction or interdependence main effect for alarm reaction time, there was a main effect for reliability, $F(2,76)= 3.708$, $p=.029$. The data followed a linear trend, with participants reacting more quickly to alarms that were more reliable, $F(1,38)=8.181$, $p=.007$.

We also analyzed task performance for the ongoing MAT tasks. However, we found no significant interaction or main effects for any of the task measures (MAT task tracking accuracy monitor activation frequency, or pump activations), $p>.05$.

Discussion

Although the results of this research were similar to past efforts, an interesting difference concerned primary task performance. Whereas Bliss et al. (in press) had noted performance fluctuations according to alarm reliability, the current research did not reveal such differences. One possible explanation is that the presence of two alarm systems with unique reliability levels increased participants' workload so that they focused exclusively on alarm reactions (Bliss & Dunn, 2000).

At an applied level, the current research findings suggest that alarm designers consider the effects of multiple alarm reactions on primary task performance. Complex tasks may be redesigned so that time-critical alarm reaction decisions are handled independently, but that team members should be encouraged to collaborate when reaction appropriateness is of great importance. A deeper examination of the results of this study raises questions about the role of trust in complex task situations.

Dimensions of Trust

In virtually all psychological experiments, participants make a decision to trust the experimenter when they sign the informed consent form. This basic trust is inspired by the content of the form, the experimenter's status, and the behavior that the experimenter exhibits prior to and during the experiment. Usually, unless there are reasons to suspect otherwise, participants will likely grant their trust to the experimenter.

In the current experiment, other representations of trust were present as well. As with typical task performance simulations, individual participants exhibited some semblance of trust in the computerized primary task and alarm systems. They trusted the primary task to perform well because of their prior conceptualizations of computer based tasks, and because they had no reason to not trust the task. Their trust in the secondary alarm task, however, was manipulated by the experimenter, who suggested that the alarms may not always be trustworthy, to various degrees. As a result, the performances of the teams reflected mistrust.

Adding to the team members' mistrust of the alarm systems was mistrust of each other. From an examination of the demographic data, it was clear that most of the team members did not know each other prior to participating in this experiment. For independent teams, this was probably not a major influence on the task performance data. Individual team members could react according to the information that they saw on their own screens, without confirming the validity of that information with their teammate. However, dependent team members were forced to consider the other team member much like the way they considered the alarms: an unknown quantity that may or may not be offering truthful information. It is likely that such skepticism contributed to the extra time required for dependent teams to react to the alarms.

Innocence Lost: The Cry-Wolf Effect as The Betrayal of Alarm Trust

In past alarm mistrust experiments, it has been shown that mistrust of automated alarms becomes translated into degraded performance (Bliss, 1993; Getty, Swets, Pickett, & Gonthier, 1995). There are several ways in which this degradation is manifested. First, when confronted with an alarm system that is unreliable, some participants will exhibit a complete lack of trust, responding to none of the alarms that sound. This happens even when they are made aware of the specific reliability level before the experiment begins. Second, some participants facing an unreliable alarm system will choose to trust it anyway, responding to all of the alarms generated. The third group, representing about 80% of the particpants, will respond in a manner that mirrors the reliability of the alarm system. In past research, this has been termed "probability matching" (Bliss, 1993).

An interesting aspect of these behaviors is that they do not appear immediately. In fact, available research shows that response patterns take about three sessions (approximately 30 alarm exposures) to develop (Bliss, Jeans, & Prioux, 1996). Empirical work is needed to determine whether alarm trust is regained over a similar time scale as it is lost. Such information has implications for interventions designed to improve alarm reactivity. Because trust is a concept that is usually attributed to human-human relationships, perhaps it would be prudent to examine those variables that facilitate the growth of trust among humans, and apply them to human-automation relationships.

Strategies for Increasing Team Trust in Automated Alarms

For alarm designers and operator trainers, it is of paramount importance to ensure that operators exhibit trust in the alarm systems. There are several ways that this might be achieved. Past research has shown the utility of maximizing alarm reliability (Bliss, 1993), advertising high alarm reliability rates (Bliss, Dunn, & Fuller, 1995), adding redundant sources of alarm information (Bliss, Jeans, & Prioux, 1996), and augmenting alarm stimuli and response options (Bliss, 1997). Each of these strategies is effective because operator trust is increased. By manipulating interaction etiquette, other possibilities exist for optimizing team trust.

Traditionally, humans have shown fear, resentment, and suspicion when confronted with unfamiliar technology (Brod, 1984) Therefore, endowing the automated alarm systems with human qualities may result in more reliable responses. Other possibilities might be to make alarm stimuli emotional (through vocal inflection) or representative of their creator (by altering pitch or auditory spectral characteristics).

Other strategies could include manipulating the trustworthiness of an operator's teammate, along a contiuum from automated to semiautomated to human. Such research may inform about operator attitudes toward automation as well as other operators.

References

Bliss, J.P. (1993). The cry-wolf phenomenon and its effect on alarm response. Unpublished doctoral dissertation. University of Central Florida, Orlando.

Bliss, J.P. (1997). Alarm reaction patterns by pilots as a function of reaction modality. The International Journal of Aviation Psychology, 7(1), 1-14.

Bliss, J.P., Bowens, L., Krefting, R., Byler, A., & Gibson, A. (in press). Collective mistrust of alarms. Proceedings of the 2002 Human Factors and Ergonomics Society Proceedings.

Bliss, J.P., & Dunn, M.C. (2000). Behavioural implications of alarm mistrust as a function of task workload. Ergonomics, 43(9), 1283-1300.

Bliss, J.P., Dunn, M., & Fuller, B.S. (1995). Reversal of the cry-wolf effect: an investigation of two methods to increase alarm response rates. Perceptual and Motor Skills, 80, 1231-1242.

Bliss, J.P., Jeans, S.M., & Prioux, H.J. (1996). Dual-task performance as a function of individual alarm validity and alarm system reliability information. Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting, 1237-1241. Santa Monica, CA: Human Factors and Ergonomics Society. October 2-8: Philadelphia, PA.

Brod, C. (1984). Technostress: The human cost of the computer revolution. Reading, MA: Addison-Wesley.

Comstock, J.R., & Arnegard, R.J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research. (Technical Memorandum No. 104174). Hampton, VA: National Aeronautics and Space Administration.

Kerstholt, J.H, Passenier, P.O., Houttuin, K., & Schuffel, H. (1996). The effect of a priori probability and complexity on decision making

in a supervisory control task. Human Factors, 38(1), 65-78.
Mallett, L.G., Vaught, C. & Brnich, M.J. (1992). Sociotechnical communication in an underground mine fire: A study of warning messages during an emergency evacuation. Safety Science, 16, 709-728.
Nohre, L., MacKinnon, D.P., & Sadalla, E.K. (1998). The effects of warning false alarms on driving behavior. Draft Manuscript. Arizona State University.
Thompson, J.D. (1967). Organizations in action. New York: McGraw-Hill.