ATTITUDES TOWARD UNRELIABLE DIAGNOSTIC AIDING IN DANGEROUS TASK ENVIRONMENTS

Faerevaag, C. L., Nguyen, B. A., Jimenez, C. A., & Jentsch, F.
University of Central Florida

It has often been cited that diagnostic aiding technology which falls below 70% reliability is not useful, and will harm overall task performance. This *reliability threshold* is based on tasks humans are capable of performing unaided. However, future robotic teammates may be capable of acting and gathering information and helping build situation awareness in environments that are too difficult or too dangerous for humans. However, if initial reliability is low, there may be resistance to introducing the technology. The current study investigates the perception of reliable and unreliable diagnostic aiding automation (robots) in both benign and dangerous environments. Undergraduate participants read a description of an autonomous robotic teammate working in either benign or dangerous environments, sending diagnostic aiding information to a human teammate with either high (80%) or low (50%) reliability. Participants in the dangerous environment conditions reported more positive overall perception of, and a stronger willingness to work with a robot, even at very low (50%) reliability. Results suggest that in dangerous environmental conditions, people may perceive unreliable robots more positively and be more willing to work with them. Implications for the introduction of new diagnostic aiding technologies, as well as strategies to support SA under conditions of unreliable diagnostic aiding are discussed.

INTRODUCTION

In an increasing number of civilian and military domains, humans are performing tasks with the help of automation (Parasuraman & Wickens, 2008). In complex operational environments, human operators must integrate information from a large number of sources in order to form an accurate situation assessment, and build and maintain situation awareness (SA; Endsley, 1995). Diagnostic aiding automation may be used to support SA by gathering and analyzing information from the environment (Horrey, Wickens, Strauss, Kirlik, & Stewart, 2006).

In real world applications, information acquisition and analysis are rarely perfect, whether performed by automation or humans (St. John & Manes, 2002). Even imperfect automation usually improves system performance (Yeh, Merlo, Wickens, & Brandenburg, 2003). However, the literature has stated that when automation reliability levels falls below 70% (e.g., Madhavan, Wiegmann, & Lacson, 2006), performance tends to decrease to levels lower than those of manual task performance alone (i.e., without automation assistance). This *reliability threshold* of 70%, as identified by Wickens and Dixon (2007), is based on a review of studies in which the automation performed tasks which humans could also perform unaided. In these types of tasks, if reliability falls too low, automation may be perceived as useless, or even as being detrimental to task performance. Thus, making human operators less likely to rely on automated systems, and instead, execute tasks manually while disregarding any information coming from an automated system that is perceived as unreliable.

As artificial intelligence technology develops, automation will be expected to perform new or increasingly complex tasks, but, initially at least, reliability of this new technology may be low (Schuster, Jentsch, Fincannon, & Ososky, 2013). Common wisdom suggests that, if reliability falls below 70%, there may be resistance to introducing the technology into real-world applications. However, if a human cannot perform a task unaided, or especially if the environment is deemed too risky for a human to begin with, there may still be an advantage, or at least a willingness, to receiving diagnostic aiding information from automation, even if it is less than 70% reliable (Maltz & Shinar, 2003). We propose that the threshold below which diagnostic aiding technology should not be introduced will vary based, among other things, on environmental risks and the human operator's ability to build and maintain SA unaided in the task environment. Hence, there is also a need to develop methods of supporting SA of human operators when working with less reliable automation.

The current research investigated the perceived value of reliable and unreliable diagnostic aiding information received from a robot in both benign and dangerous environments. Implications for the introduction of diagnostic aiding technologies and supporting SA when diagnostic aiding is unreliable are discussed.

Reliability Threshold

An approximation of a *reliability threshold* below which the benefits of an automated system disappear, is a practical parameter for use in computational modeling of system performance (Wickens & Dixon, 2007). As Wickens and Dixon note, any value used as a *reliability threshold* will not be exact, since it will be moderated by numerous contextual factors, such as: human operator knowledge, task difficulty, concurrent task load, etc. It is important that factors moderating the *reliability threshold* are well understood, since system managers could use a threshold as a rule-of-thumb for deciding whether to introduce diagnostic aiding technology. This study explored the effect that two types of environments (benign versus dangerous) may have on perceived reliability

and the willingness to use an automated aid. In extreme environments where it would be too risky, or impossible for humans to perform unaided, any amount of work automation can perform will be useful. Under these circumstances, less reliable automation may still be preferable to none (Maltz & Shinar, 2003; St. John & Manes, 2002).

SA and Diagnostic Aiding

Parasuraman, Sheridan, and Wickens (2000) proposed a four-stage model of automation in which automation may be applied at varying degrees to any of the following stages: (1) information acquisition, (2) information analysis, (3) decision selection, and (4) action implementation. The first two stages of the four-stage model of automation are known as *diagnostic aiding* (Horrey et al., 2009), and also map cleanly on to the first two levels of Endsley's model (1995) of SA: (1) perception of elements in the environment; and (2) comprehension of their meaning.

When automation is used for diagnostic aiding, it can be a useful tool in building and maintaining SA, especially when it extends the perceptual capabilities of unaided humans (Endsley, 1996). If reliability is low, it could negatively impact SA by providing incorrect information to the operator. However, if an unreliable diagnostic aid is the only option for gathering information in a dangerous task environment, there is a need to develop mitigating strategies to support SA of operators working in these environments.

Imperfect Automation

Rarely, if ever, is automation 100% reliable, nor is it capable of operating 100% independently. Even a more advanced autonomous system will occasionally need human intervention (Burke, Murphy, Coovert, & Riddle, 2004). When failures do occur, the ability of the human operator to recover will be affected by the dynamic within the human-robot team, and the nature of the error (Groom & Nass, 2007).

Diagnostic aiding automation designed to detect elements in the environment can produce two types of errors: misses and false alarms. The system designer may program a system to be more prone to false alarms or misses by setting the response criterion to be liberal or conservative, respectively, as described by Signal Detection Theory (Green & Swets, 1988). Designers may have a good reason to set a liberal response bias (if, for example, a miss would put human lives at risk), but research in this area has generally found false alarms to be more detrimental to overall task performance than misses (Dixon & Wickens, 2006). In the current study, the reliability of the automation was indicated by false alarm rate.

Current Study

Participants were asked to rate a hypothetical autonomous robotic teammate at varying levels of reliability on characteristics such as trustworthiness, competence, and intelligence. It was hypothesized that positive perceptions of a robotic teammate would be higher at lower levels of reliability in dangerous environments than it would in more benign conditions. It was also hypothesized that participants will report that they would be more likely to want to work with the robot in dangerous environments, regardless of reliability level.

Participants were presented with information about a fictitious autonomous robot called SearchBot capable of providing Stage 2 (information analysis) diagnostic aiding. Participants were told that SearchBot is capable of working alongside humans as a member of a search and rescue team, and that its main function is to identify human forms, even if they are immobile. When SearchBot detects a human form in the environment, it relays the location to a remote human teammate on a visual display. Two between-subjects manipulations were used. The operational environment was either benign or dangerous, and reliability was either high (80%) or low (50%).

METHOD

Participants

Undergraduate students were recruited from introductory Psychology classes at the University of Central Florida, and compensated with course credit. A power analysis indicated that 179 participants would yield a medium effect size. Data was collected from 198 participants, and 18 participants were excluded due to failed manipulation checks.

The 180 participants included in the analyses ranged in age from 18 to 35 years-old (M = 19.83, SD = 2.35), and 43.3% were male. The demographic questionnaire also asked for participants' military experience, first language, and a self-reported level of familiarity with robotic technology on a scale from 1 (Not at all familiar) to 6 (Very familiar). No significant differences were found based on these characteristics.

Design

Independent Variables. A 2X2, between-subjects factorial design was used, with independent variables of reliability (factors: low and high) and environment (factors: benign and dangerous). Reliability was set at either 80% (high), or 50% (low), which was indicated by a false alarm rate of either 20% or 50%, respectively. Environment was manipulated by showing participants sets of images of areas in which SearchBot is capable of operating: environments easily accessible to humans (benign), or ones inaccessible to humans (dangerous).

Dependent Variables. A questionnaire, *Subjective Ratings of Autonomous Robots* (S-RAR), was developed specifically for the purposes of this study, and asked participants to rate SearchBot on nine traits using a 6-point semantic differential scale. Bipolar adjectives included: unreliable/reliable, untrustworthy/trustworthy, undependable/dependable, unintelligent/intelligent, unpredictable/predictable, incompetent/competent, useless/useful, incapable/capable, and unhelpful/helpful. The scale demonstrated good internal consistency, with a Cronbach's alpha of .84. Responses on the nine items were

averaged to indicate the participants' overall perception of SearchBot on a scale from 1 to 6. A higher S-RAR score indicated a more positive perception. An additional question asked participants to rate on a 6-point Likert scale how likely they would be to use SearchBot to assist them in a hypothetical search and rescue task (1 = Very unlikely, 6 = Very likely). Participants were then asked to select three scenarios in which they believed humans working with the aid of SearchBot would be more effective than humans working alone. They were given a list of 21 scenarios to choose from, as well as the option to provide their own response.

Materials

The stimuli included reading material and pictures describing the purpose, capabilities, reliability, and operational environment of a fictitious autonomous robot called SearchBot. Also included were a picture of SearchBot, and a picture of the visual display used by the human teammate to receive diagnostic aiding information from SearchBot (see Figure 1). SearchBot was an edited photo of the iRobot SUGV 5. The visual display was a picture of a Durabook TA10 (a "rugged" tablet designed for military and other field operations), edited to display a floor plan schematic with red dots, indicating "signals" from SearchBot. The pictures of benign and dangerous environments were chosen to look as similar as possible in complexity, but differ in accessibility to humans. For example, both conditions contained a picture of a staircase, but in the dangerous condition, the staircase was inside of a burning building (see Figure 2). Other environment images contained uneven surfaces and obstacles.

The passages read by participants were 653 words long, and scored a 10.4 on the Flesch-Kincaid grade level test in Microsoft Word. The text described SearchBot's capabilities and limitations, as well as its reliability level. SearchBot was described as being able to detect human forms, and relay their location to a human teammate. The text provided examples of SearchBot's reliability, and this varied by condition. For example, in the low reliability condition, it was explained that for every ten signals SearchBot sent, the human teammate would investigate the area and find a human five times, and the other five times would find an object vaguely resembling a human form.

Figure 1. SearchBot and visual display.

Figure 2. Staircase image in benign (left) and dangerous environment condition (right).

Procedure

This study was reviewed and approved by the university's Institutional Review Board. Participants were randomly assigned to one of four conditions before they arrived in the research lab. Participants were told that they would be reading and viewing information about an autonomous robot, capable of working as a teammate, and then answering a questionnaire. The researcher explained the relevant terms (e.g. autonomous robot, human-robot teaming) and then presented the participant with the stimuli. After being given time to study the text and pictures, the researcher answered any questions participants had, and then gave them the questionnaires. When all questionnaires were completed, the researcher explained the purpose of the study, and thanked participants for their time.

RESULTS

A 2 (Environment: Benign or Dangerous) x 2 (Reliability: Low or High) between-subjects MANOVA was used to calculate group differences for S-RAR score, and the likelihood rating that the participant would utilize SearchBot in a search and rescue task. Using Pillai's trace, there was a significant effect of environment, $V = 0.80$, $F(2, 175) = 7.63$, $p = .001$, and reliability, $V = 0.18$, $F(2, 175) = 19.26$, $p < .001$ on the S-RAR score and the likelihood of use. Follow up univariate ANOVAs showed a significant main effect of environment on both the S-RAR, $F(1, 176) = 14.56$, $p < .001$, partial $\eta^2 = .076$, and the likelihood of use, $F(1, 176) = 5.84$, $p = .017$, partial $\eta^2 = .032$. S-RAR scores were higher in the dangerous environmental condition ($M = 5.09$, $SD = .49$), than in the benign condition ($M = 4.78$, $SD = .72$). The same was true for the likelihood of use: scores were higher in the dangerous environmental condition ($M = 5.20$, $SD = .78$), than in the benign condition ($M = 4.87$, $SD = 1.04$). The follow up univariate ANOVAs also showed a significant effect of reliability on the S-RAR score, $F(1, 176) = 14.56$, $p < .001$, partial $\eta^2 = .17$. As expected, participants in the high reliability condition rated SearchBot higher ($M = 5.19$, $SD = .48$) than participants in the low reliability condition ($M = 4.69$, $SD = .68$). There was no significant effect of reliability on the likelihood of using SearchBot, $p > .05$, indicating that participants were similarly likely to want to use SearchBot

whether reliability was high (M = 5.09, SD = .97) or low (M = 4.98, SD = .90).

An a priori analysis was conducted to test the hypothesis that participants' S-RAR scores of low reliability SearchBot would be different in dangerous and benign environments. The test shows a significant difference in the scores, $t(72.05) = 3.60$, $p = .001$. Participants in the low reliability - dangerous environment condition rated SearchBot higher (M = 4.93, SD = .45) than participants in the low reliability - benign environment condition (M= 4.45, SD = .78).

When participants chose three scenarios in which they think a human-robot team would be more effective than humans alone, chi-square tests showed no significant differences across groups. The most common responses were Fire (42.1%), Earthquake (34.1%), Carbon Monoxide Leak (32.3%), and Detecting Enemies in War Zones (25.55%). A chi-square test showed that the 25 participants from whom data was collected in the two weeks following a shooting at the Pulse Nightclub in Orlando, FL on June 12, 2016 were more likely to choose Active Shooter, χ^2 (1, N = 50) = 6.349, p = .025, and Police Raids, χ^2 (1, N = 50) = 5.711, p = .037, regardless of condition, when compared to the previous 25 participants from whom data was collected immediately before the shooting.

DISCUSSION

Results of the present study provide evidence of the likelihood of humans' willingness to use unreliable systems, especially in dangerous environments. Participants had a more positive perception of SearchBot when it was portrayed as operating in a dangerous environment. They also indicated a greater willingness to use the robot in the dangerous environment condition, regardless of reliability level. Within the low reliability condition, participants rated SearchBot more positively when it was portrayed as operating in a dangerous environment vs. a benign one. This reinforces the notion that humans would perceive automated systems more favorably under dangerous circumstances, even if its reliability is considered low.

As was expected, participants had a higher positive perception of SearchBot when its reliability was high. However, they did not indicate a significantly greater inclination to want to use it than the low reliability SearchBot. One possible explanation as to why participants were willing to use low and high reliability SearchBot alike may be due to the nature of the error, which was false alarms. False alarms, while potentially detrimental to performance, may not always be as problematic as misses. In situations such as search and rescue, it may be acceptable for a system to erroneously identify a random shape as a potential human shape than a system failing to identify actual humans.

History Effects

Due to the research university's proximity to the Pulse Nightclub shooting in Orlando, FL, local media was saturated with reporting on the shooting, potentially more than in other parts of the country. Following the shooting, participants were more likely to choose Active Shooter and Police Raids as scenarios in which Human-Robot teams would be more effective than humans alone.

At the time, it was reported that a robot was used to search the scene for explosives after the shooter had been neutralized, but before first responders entered the building (Santora, 2016). This is the type of scenario in which a robot gathering information from a dangerous environment can protect human lives, and sending in humans would generally be considered too risky if an automated alternative is available. It is also a situation in which a liberal response bias would be considered acceptable, since a miss would potentially cause the death of first responders.

This is a real-world example in which using a robot is preferable, perhaps even if it falls below the 70% *threshold*. However, at too high a level of unreliability, the misleading diagnostic aid may delay rescue efforts, leading to more casualties. This highlights the importance of studying acceptable *reliability thresholds* for diagnostic aiding technology.

Limitations and Future Research

This study investigated how environment affects perceptions of, and willingness to work with robotic teammates. Future research is needed to determine effects on task performance and the *reliability threshold*.

With a reliability rate of only 50%, the unreliable version of SearchBot fell well below the 70% threshold identified by Wickens and Dixon (2007). The fact that in no condition did S-RAR score or likelihood of use fall below 3.5 (the neutral point) suggests that a lower reliability level should be included in future studies.

If humans are likely to rely on unreliable automation in dangerous environments, it is important to investigate how this will impact SA. Mitigating strategies to support SA will be needed in future human-robot teams, particularly if the automated aid is unreliable. There may be a benefit to using two or more aids in order to cross-check information. For example, in some instances, additional information from an unmanned aerial vehicle (UAV) supports SA of human operators working with ground robots (Chadwick, 2008). In environments where UAV operation is not possible, multiple ground robots could be used. For example, in dark, enclosed, or cluttered environments, views from different perspectives may be used to overcome perceptual difficulties and provide an integrated assessment of the environment. Other strategies may include training for operators working with unreliable diagnostic aids in order to build accurate mental models of the robot's capabilities and limitations, and user interface which conveys information about how and when information may be unreliable.

CONCLUSIONS

The results of this study suggest that, in dangerous environments, humans may be more willing to use unreliable diagnostic aiding. In real world applications, less reliable diagnostic aiding may need to be used if it is the only way to

build and maintain SA in a dangerous area. We propose that in any task environment, automation which can perform better than an unaided human will be useful. If a human is unable to gather information from a dangerous environment, an automated diagnostic aid may be of some help, even at very low levels of reliability.

In these situations, reliance on imperfect automation may be higher compared to benign environments. Understanding the dynamics of trust and reliance in human-automation interaction will become increasingly important as human-robot teams are utilized to perform increasingly dangerous tasks, where risk of failure carries higher consequences. If humans are to utilize robotic teammates in the safest and most effective ways, further research is needed on ways to support SA in human-robot teams.

ACKOWLEDGEMENT

Portions of the research reported here were performed in connection with Contract Number W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation heron.

REFERENCES

Burke, J. L., Murphy, R. R., Coovert, M. D., & Riddle, D. L. (2004). Moonlight in Miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human–Computer Interaction, 19*(1-2), 85-116.

Chadwick, R. A. (2008). Considerations for use of aerial views in remote unmanned ground vehicle operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52*, 252–256.

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.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*, (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum.

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics.* Los Altos, CA: Peninsula Publishing.

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies, 8*(3), 483-500.

Horrey, W. J., Wickens, C. D., Strauss, R., Kirlik, A., & Stewart, T. R. (2009). Supporting situation assessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement on judgment skill. In A. Kirlik (Ed.), *Adaptive perspectives on human-technology interaction: Methods and models for cognitive engineering and human-computer interaction* (pp. 55-70). Oxford University Press.

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.

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors, 45*(2), 281-295.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30*(3), 286-297.

Parasuraman, R., & Wickens, C.D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50*(3), 511-520.

Santora, M. (2016, June 12). Last Call at Pulse Nightclub, and Then Shots Rang Out. New York Times. Retrieved from: http://www.nytimes.com/2016/06/13/us/last-call-at-orlando-club-and-then-the-shots-rang-out.html?_r=0

Schuster, D., Jentsch, F., Fincannon, T., & Ososky, S. (2013). The impact of type and level of automation on situation awareness and performance in human-robot interaction. In D. Harris (Ed.), *Lecture Notes in Computer Science: Vol. 8019. Engineering Psychology and Cognitive Ergonomics*, (pp. 252-260).

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*, 332-336.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201-212.

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.