The Effects of Automation Reliability and Multi-tasking on Trust and Reliance in a Simulated Unmanned System Control Task

Svyatoslav Guznov(✉), Alexander Nelson, Joseph Lyons, and David Dycus

Air Force Research Laboratory, WPAFB, Dayton, USA
{svyatoslav.guznov.ctr, alexander.nelson.2, joseph.lyons.6, david.dycus}@us.af.mil

Abstract. This study examined the effects of automation reliability and multi-tasking on trust and reliance in a simulated unmanned system scenario. Participants performed an insurgent search task with the help of an automated aid that provided information about targets with varying levels of reliability (high, medium, and low). In addition, a multi-tasking condition was implemented in which a radio communication assignment designed to increase cognitive demand was performed. Results indicated that participants were not able to accurately assess the true reliability of the automated aid in any condition, and were unable to discriminate between low and medium reliability. Results from the multi-tasking manipulation show that participants were more reliant upon the automated aid when the secondary task was present. Overall, this study provides insight into the patterns of trust calibration errors that may negatively affect performance in human-machine teams, particularly when additional task pressure is present.

Keywords: Unmanned systems · Automation reliability · Trust · Reliance · Multi-tasking

1 Introduction

In recent years, military systems have become more technologically complex and are incorporating higher levels of automation than ever before. As the function of the operator in these human-machine systems has evolved to more of a supervisory and decision-making role, the importance of understanding the factors that influence trust and reliance has dramatically increased. Most automated systems are not infallible and it is critical that the human partner is able to calibrate their trust for appropriate reliance. When trust is not calibrated correctly, errors in over-reliance and under-reliance can occur, leading to *misuse or disuse* of the automation [12].

The major factor that influences the reliance calibration process is trust [7]. Trust in a system is a belief that the trustee will accomplish a certain objective and the willingness of a trustor to accept vulnerability and uncertainty [6]. Several studies

C. Stephanidis (Ed.): HCII 2015 Posters, Part II, CCIS 529, pp. 616–621, 2015.
DOI: 10.1007/978-3-319-21383-5_103

examined the influence of automation reliability on trust and reliance showing that increased automation reliability increases trust and reliance [1]. However, unmanned systems often incorporate multi-tasking demands where the operator, in addition to the primary task (e.g., target search), needs to communicate with teammates, interact with a control panel, or accomplish other tasks. Such environments might result in overreliance errors and misuse due to the diversion of cognitive resources away from the evaluation of automation performance [11]. These types of overreliance errors can have severe consequences when the automation used to accomplish the task is imperfect [6].

Although previous studies have examined automation reliability and multi-tasking as factors that affect trust and reliance, none of them looked at the joint effects of these two factors. In this study, we examined the effects of three levels of automation reliability and task type on trust and reliance. The participants performed an insurgent search task in the Mixed Initiative eXperimental (MIX) [2] testbed and had an Automated Aid (AA) that provided information about insurgent and other combatant locations on a map. In the multi-task condition, participants were asked to perform a communication task (Coordinate Response Measure (CRM) [3] in addition to the search task.

An interaction was expected between reliability and multi-tasking factors. The participants were expected to calibrate their trust and reliance appropriately to the level of automation reliability when posed with the insurgent search task only. However, the participants were expected to over trust and overrely on low reliability automation when asked to perform the CRM task concurrently with the insurgent search task.

2 Methodology

2.1 Participants

Forty eight participants were recruited for this experiment (28 men and 20 women). Participants ranged in age from 18 to 59 years (M = 36.67, SD = 11.22). All participants reported normal or corrected-to-normal vision.

2.2 Design

The experiment employed a 3 (Automation Reliability) × 2 (Task Type) mixed design. The Automation Reliability was a between-subjects factor including high reliability (HR), medium reliability (MR), and low reliability (LR) levels with the reliability values of 93 %, 75 %, and 55 % respectively. The Task Type was a within-subjects factor including single task and multi-task levels. The Task Type factor levels were counter-balanced to control for potential carry-over effects. The dependent variables for the study were insurgent search performance and CRM task performance; reliance, trust state, and perceptual accuracy with the regard to the AA; and perceived workload.

2.3 Apparatus and Materials

The experiment was conducted using two computers that ran the MIX testbed and the CRM task. The MIX testbed simulated a UGV task. The MIX interface consisted of a video feed window that showed the UGV camera view and the AA window. In the simulation, the UGV moved along a pre-determined path while the operator monitored the video feed screen searching for insurgents. The AA provided the participants with a map that showed the locations of the combatants. Depending on the condition (i.e., low, medium, or high reliability), the AA made respectively seven, four, or one classification errors. A classification error occurred when the AA marked an insurgent as a non-insurgent or vice versa. Participants were asked to press either *Accept* or *Reject* buttons in the AA interface when they agreed or disagreed with the AA.

The CRM software was used to simulate a military radio communication assignment. Each participant was assigned the call sign "Arrow" and was asked to follow commands associated with their call sign by pressing a color- and number-coded button as quickly as possible on a touch screen monitor. The program logged the accuracy of the selections made by the participants.

In this study, the following metrics were used. The Perceptual Accuracy metric [9] estimated participants' accuracy at evaluating the reliability of the AA. The Trust Scale [8] measured participants' trust state. The NASA-Task Load Index (NASA-TLX) [5] was used to measure participants' perceived workload.

2.4 Procedure

Upon arrival, the participants were trained on how to perform the experimental task in the MIX testbed and the CRM task. Next, the participants performed the experimental task consisting of two phases: single task (insurgent search in the MIX simulator alone) or multi-task (insurgent search and the CRM task). In the single task condition, the participants were asked to search for the insurgents using the AA. The participants were also asked to accept or reject the AA's classification suggestions. The task was paused three times to administer the Trust Scale, Perceptual Accuracy, and NASA-TLX questionnaires. In the multi-task condition, the participants performed the task identical to that of the single task condition, but were also asked to simultaneously perform the CRM task.

3 Results

3.1 Perceived Reliability

Mixed-model ANOVA showed a significant main effect for the Automation Reliability factor, F (2, 33) = 29.85, $p < .001$, partial $\eta^2 = .64$. Post hoc comparisons using the Tukey HSD criterion for significance showed that there was no significant difference between LR and MR levels. However, both LR (M = 59.79, SD = 9.06) and MR (M = 64.44, SD = 9.51) levels had significantly lower perceived reliability ratings when compared to HR (M = 84.31, SD = 8.02) level with $p < .001$ for both comparisons.

3.2 Trust State

Mixed-model ANOVA revealed a significant main effect for the Automation Reliability factor, F (2, 33) = 14.04, $p < .001$, partial $\eta^2 = .46$. Post hoc comparisons using the Tukey HSD criterion for significance showed that there was no significant difference between LR and MR levels. However, both LR ($M = 2.04$, $SD = .7$) and MR ($M = 2.53$, $SD = .68$) levels were significantly lower in trust ratings when compared to HR ($M = 3.68$, $SD = .73$) level with $p < .001$ for both comparisons.

3.3 Reliance

The reliance scores were calculated as a sum of the total number of agreements with the AA. Mixed-model ANOVA revealed a significant interaction between Automation Reliability and Task Type factors, F (2, 33) = 3.48, $p < .04$, partial $\eta^2 = .17$. In addition, there was a significant main effect for Automation Reliability F (2, 33) = 51.89, $p < .001$, partial $\eta^2 = .76$. Post hoc comparisons with the Tukey HSD criterion for significance showed LR Single Task condition ($M = 7.41$, $SD = .97$) produced significantly lower reliance when compared to LR Multi-task condition ($M = 8.45$, $SD = .1.26$), $p < .05$ (Fig. 1).

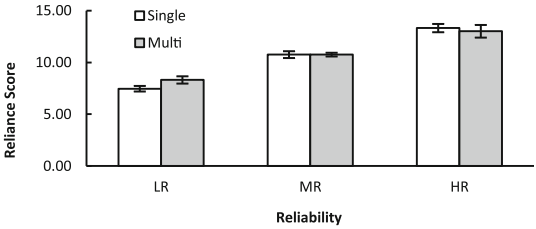

Fig. 1. Reliance across automation reliability conditions. Error bars are standard errors.

3.4 Insurgent Search Task and CRM Task Performance

No significant main effects for the Automation Reliability and Task Type for both the insurgent search and the CRM tasks were observed. In addition, there was also no significant interaction between these two factors for both tasks.

3.5 Global Workload

Mixed-model ANOVA revealed a significant main effect for the Task Type factor, F (1, 44) = 4.88, $p < .05$, partial $\eta^2 = .1$. Participants in the Multi-Task condition ($M = 29.31$, $SD = 16.09$) reported significantly higher workload when compared to the Single Task condition ($M = 26.06$, $SD = 17.27$).

4 Discussion

The main focus of the study was to examine the joint effects of automation reliability and multi-tasking on trust calibration and reliance towards the AA. The overall results confirmed some, but not all, of the original hypotheses. While LR and MR conditions were similar to each other in perceived reliability, each was significantly lower when compared to the HR condition. In addition, the participants in LR condition overestimated the reliability by approximately 6 % and participants in MR and HR conditions underestimated the reliability by approximately 11 %. A similar pattern was observed for the trust state scores: the participants in LR and MR conditions rated their trust to be significantly lower when compared to the HR condition. However, there was no difference between LR and MR conditions showing that the participants were not sensitive to reliability manipulation when the reliability levels are low (50 %–70 %). For the AA reliance, a significant interaction was observed showing that participants relied on the LR automation more when asked to perform the CRM task confirming the original hypothesis that increasing task demand would divert cognitive resources away from evaluating the automation. Finally, participants found the multi-tasking condition more challenging when compared to the single task condition indicating that the CRM task indeed induced additional mental demand.

Generally, the results indicate the complex nature of the interaction between different levels of automation reliability and multi-tasking. Participants were neither accurate in their judgment of automation reliability levels nor were they able to discriminate between low and medium levels of automation reliability. It appears that they "averaged" low and medium levels of reliability by overestimating one and underestimating the other showing low trust resolution [6]. The results related to underestimation of reliability correspond well with previous findings of underestimation of imperfect automation [13]. These findings indicate that human perception of system reliability is not linear and possibly require additional features to help the operators correctly judge its magnitude. In addition, while the participants were not affected by the CRM task in their perceptual accuracy and trust ratings, the behavioral outcome (i.e., reliance) was affected in the low reliability condition. This shows that even if the participants estimated the aid to be equally reliable, they still have a tendency to agree with the automation more, possibly due to a lack of the cognitive resources to adequately interact with the automation as suggested by [10]. Overall, this study provides insight into the patterns of trust calibration errors that may negatively affect performance in human-machine teams, particularly when additional task pressure is present.

There are limitations associated with the experiment that the authors would like to address in future studies. The performance data indicated that the results could have been more dramatic if the insurgent search task and the secondary tasks were higher in difficulty or longer in duration. In addition to addressing these errors, future studies would benefit from the integration of psychophysiological metrics (e.g. EEG, eye-tracking) that give additional information about the participants' trust and workload states.

References

1. Bailey, N.A., Scerbo, M.W.: Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience and operator trust. Theor. Issues Ergon. Sci. **8**, 321–348 (2007)
2. Barber, D., Davis, L., Nicholson, D., Chen, J.Y.C., Finkelstein, N.: The mixed initiative experimental (MIX) testbed for human robot interactions with varied levels of automation. In: Proceedings of the 26th Annual Army Science Conference, December 1–4, ADA505701 (2008)
3. Bolia, R.S., Nelson, W.T., Ericson, M.A., Simpson, B.D.: A speech corpus for multitalker communication research. J. Acoust. Soc. Am. **107**, 1065–1066 (2000)
4. de Visser, E.J., Parasuraman, R., Cosenzo, K.: Effects of imperfect automation on human supervision of multiple uninhabited vehicles. Paper presented at the Annual Meeting of Division 21 of the American Psychological Association, George Mason University, Fairfax, VA, March 2007
5. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload. North Holland Press, Amsterdam (1988)
6. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)
7. Lee, J.D., Seppelt, B.D.: Human factors in automation design. In: Nof, S. (ed.) Springer Handbook of Automation, pp. 417–436. Springer, New York (2009)
8. Merritt, S.M., LaChapell, J., Lee, D.: The perfect automation schema: Measure development and validation. Technical report submitted to the Air Force Research Laboratory, Human Effectiveness Directorate, 30 June 2012
9. Merritt, S.M., LaChapell, J., Lee, D.: Continuous calibration of trust in automated systems-phase 2. Technical report submitted to the Air Force Research Laboratory, Human Effectiveness Directorate, 31 May 2013
10. Parasuraman, R., Manzey, D.: Complacency and bias in human use of automation: an attentional integration. Hum. Factors **52**, 381–410 (2010)
11. Parasuraman, R., Molloy, R., Singh, I.L.: Performance consequences of automation-induced "complacency". Int. J. Aviat. Psychol. **3**, 1–23 (1993)
12. Parasuraman, R., Riley, V.: Humans and automation: use, misuse, disuse, abuse. Hum. Factors **39**, 230–253 (1997)
13. Wiegmann, D.: Agreeing with automated diagnostic aids: a study of users' concurrence strategies. Hum. Factors J. Hum. Factors Ergon. Soc. **44**(1), 44–50 (2002)