

The Effects of Automation Error Types on Operators' Trust and Reliance

Svyatoslav Guznov¹✉, Joseph Lyons¹, Alexander Nelson¹, and Montana Woolley²

¹ Air Force Research Laboratory, WPAFB, Ohio, USA
{svyatoslav.guznov.1, joseph.lyons.6,
alexander.nelson.2}@us.af.mil

² CSRA, Virginia, VA, USA
montana_woolley@sra.com

Abstract. This study examined the joint effects of automation error type and error severity on operators' trust and reliance during a simulated unmanned aerial vehicle (UAV) mission. Participants were asked to search for improvised explosive devices (IEDs) with the help of an automated aid (AA). Four combinations of error types (miss and false alarm) and severity (mild and severe) were used, but with the same rate of error across all conditions. The results did not confirm the original hypothesis that severe false alarms would result in the lowest levels of trust and reliance, while the mild miss condition would result in the highest levels. No significant differences in self-reported trust were found among the conditions and the mild false alarm condition resulted in the lowest levels of reliance on the AA. In addition, reported perception of reliability of the AA was significantly lower in the severe miss condition compared to all other conditions. Finally, the mild false alarm condition produced the worst IED search task performance. Overall, results indicate a complex interaction between error types and error severity.

Keywords: Unmanned systems · Automation errors · False alarms · Misses · Trust · Reliance · Multi-tasking

1 Introduction

Operators' interaction with unmanned systems is becoming increasingly complex. In order to maintain optimal performance of human-machine teams, automation is heavily implemented. Multiple studies demonstrate the benefits of automation reflected in improved decision making, performance, and reduced workload [1]. At present, these systems are not perfectly accurate and may commit errors thus requiring the operators to maintain supervisory control over them. The operators must be "in the loop" to make the final decision on when to trust or distrust the automation, consequently leading to their decision to use or not use the automation (i.e., rely on automation). The major factor that forms the decision to rely on the system is trust [2] which is defined as an individual's (1) belief that the system will accomplish a certain objective and (2) willingness to accept vulnerability and uncertainty [3]. Trust is particularly important in a situation when the

operators need to use imperfect automation (e.g., low reliability automation) to accomplish the task [3].

A key characteristic of system reliability is the number of errors it commits. Previous studies show that more reliable systems (i.e., the fewer errors made) result in higher trust and reliance [4]. These studies primarily concentrated on the effects of error rates on human interaction with automation [5]; however, few studies have explored such effects of different types of errors. One study [6] examined the effects of error difficulty on trust and reliance. They found that obvious errors made by automation produced larger reduction in trust and reliance. Additionally, [7] found that more severe automation errors – the errors with the largest negative outcome – resulted in a larger decrease in trust. In addition to error severity, research by [5, 8, 9] suggested that false alarms (FAs) and misses differentially affect reliance behavior: FAs had a more negative effect on trust than did misses, which is possibly due to FAs being more noticeable. Finally, [10] examined the effects of error type on automation reliance. In cases when automation produced FAs, participants (1) did not rely on automation during states when alerts (i.e., automation suggestions) were provided and (2) over relied on automation when alerts were not provided. The impact of an automation that produced misses had the opposite effect.

Previous studies examined the effects of error types and severity on trust and reliance. However, none of them explored their joint effect in a realistic task environment. For example, in a military mission an automated system that monitors and classifies improvised explosive devices (IEDs) and non-IEDs can commit errors in multiple ways: by committing false alarms or misses and by making errors with different levels of severity. Thus, the goal for this study was to examine the effects of these types of errors on operators' trust and reliance in the automated system within a generic IED search mission. During the study, the participants performed an IED search task in the Mixed Initiative eXperimental (MIX) testbed [11, 15], which simulated controlling an Unmanned Aerial Vehicle (UAV). The participants were given an Automated Aid (AA) that provided information about the location of the IED and non-IED units on a separate map. Depending on the condition, the AA committed distinct error types including IED miss (severe miss), non-IED miss (mild miss), IED false alarm (severe false alarm), and non-IED false alarm (mild false alarm). It was expected that severe false alarm errors would result in the lowest levels of trust and reliance, while the mild miss errors would result in the highest levels of trust and reliance.

2 Methodology

2.1 Participants

Sixty-eight participants were recruited for this experiment (23 men and 45 women). Participants ranged in age from 18 to 52 years ($M = 23.79$, $SD = 8.61$). All participants reported normal or corrected-to-normal vision and hearing, normal color vision, and no history of neurological disorders.

2.2 Design

The experiment employed a 2 (Error Type) \times 2 (Error Severity) between-subjects design. The Error Type factor included Miss (M) and False Alarm (FA) levels. The Error Severity factor included Mild Error (ME) and Severe Error (SE) levels. Thus, 17 participants were assigned at random to one of the four experimental conditions (MME, MSE, FAME, and FASE). The dependent variables for the study were IED search performance, reliance on the AA, state trust, perceived reliability of the AA, and perceived workload. Additionally, trait trust was measured.

2.3 Apparatus and Materials

The experiment was conducted using a computer that ran the MIX testbed which simulated a UAV control task. The MIX interface was comprised of the UAV videofeed window that showed a UAV camera view, the AA window which classified the objects by placing icons on the map, and the Ground Troop Report (GTR) window which provided feedback on the decision made by the participants (Fig. 1).

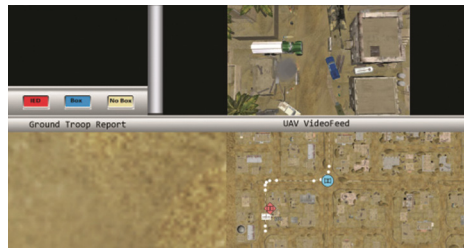


Fig. 1. MIX testbed screenshot

During the task, the UAV moved autonomously along a pre-determined path. The operator monitored a limited area through a UAV videofeed and searched for IEDs using the AA suggestions. The path contained IEDs and non-IEDs (similar looking but harmless objects) further referred to as *Boxes*. The UAV periodically stopped and the participants were prompted to make a decision by pressing one of the three buttons to classify IEDs, Boxes, or no IED/Box situations in the GTR window. The simulator logged the UAV coordinates, events on the AA map, and user mouse clicks within the MIX interface. Throughout the course of the search task, the AA provided the participants with suggestions about the locations of the IEDs and Boxes by placing corresponding icons in the AA map. The AA was not entirely reliable, producing errors in 15 % of suggestions for all conditions. However, the conditions varied based on error type and severity. Miss and FA error types contained two separate severity levels of errors, thus exposing participants to four error conditions. The AA committed Miss error types by either reporting no IED icon when there was an IED in the environment (severe error) or reporting no Box icon when there was a Box in the environment (mild error). Similarly, the AA committed FA error types by either reporting an IED icon when there was no IED (severe error) or reporting a Box icon when there was no Box (mild error). The AA

made a total number of 20 classifications in each condition. Specifically, the AA in the MSE condition contained seven IED icons, ten Box icons, and three miss errors. The AA in the MME condition contained ten IED icons, seven Box icons, and three miss errors. The AA in the FASE condition contained ten IED icons (three of which were FA errors) and ten Box icons. Finally, the AA in the FAME condition contained ten IED icons and ten Box icons (three of which were FA errors). The environment contained either 17 or 20 total objects depending on the condition. Specifically, the MSE and MME conditions contained 10 IEDs and 10 Boxes, the FASE condition contained 7 IEDs and 10 Boxes, and the FAME condition contained 10 IEDs and 7 Boxes. The participants received feedback in the GTR window immediately after making a decision. The length of the route was set to 20 min. In this study, the following metrics were used: the Perceived Reliability metric measured participants' accuracy at evaluating the reliability of the AA [12], the Reliance Intentions Scale (RIS) measured participants' state trust [13], the Propensity to Trust Machines (PTM) metric measured trait trust [12], and the NASA-Task Load Index (NASA-TLX) measured perceived workload [14].

2.4 Procedure

Upon arrival, each participant was asked to read and sign the informed consent forms followed by completion of a color blindness test. Next, the participants filled out the demographics questionnaire and PTM metric. The participants were trained on how to perform the experimental task in the MIX testbed. In particular, they were trained on how to search for the IEDs and how to use the AA that assisted in the IED search. Participants were informed that in the actual experimental task the AA might make errors; however, no other information regarding number or type of errors was provided. Near the end of the training phase, the participants performed a short scenario to practice the skills. This short training scenario's AA made no errors. Next, the participants performed the experimental task in the MIX testbed. At the end of the experimental phase, the participants filled out the RIS, Perceived Reliability Scale, and NASA-TLX questionnaires. Upon completion of the experiment, the participants were debriefed and dismissed.

3 Results

3.1 Perceived Reliability

Between-subjects ANCOVA with the PTM variable as a covariate showed a significant interaction between the Error Type and Error Severity factors, $F(1, 63) = 7.41, p < .05, \eta^2 = .11$. Post-hoc comparisons using the Tukey HSD criterion for significance showed significantly lower perceived reliability for MSE ($M = 51.47, SD = 26.73$) when compared to MME ($M = 72.53, SD = 13.22$) and FASE ($M = 70.70, SD = 17.40$) conditions with $p < .05$ for both comparisons (Fig. 2).

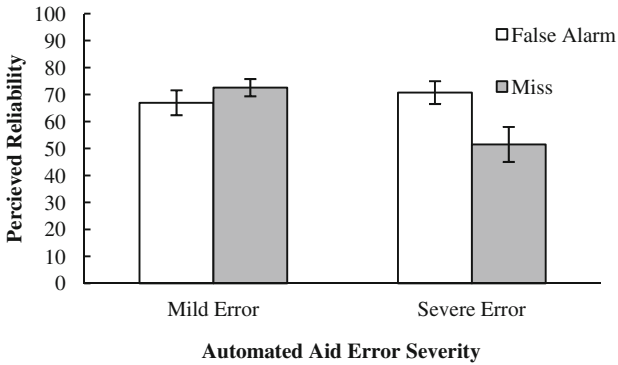


Fig. 2. Perceived reliability of the AA across the error types. Error bars are standard errors

3.2 State Trust

Between-subjects ANCOVA with the PTM variable as a covariate showed no significant main effects or interaction for the Error Type and Error Severity factors for the state trust. The average trust level across all conditions was 36.12 ($SD = 10.46$), with the maximum possible score of 70. The Cronbach’s alpha for the metric was .83.

3.3 Reliance

Reliance percentages (i.e., percentage of agreements with the AA) were calculated for each condition. Between-subjects ANCOVA with the PTM variable as a covariate revealed a significant interaction between Error Type and Error Severity factors, $F(1, 63) = 19.30, p < .001, \eta^2 = .24$. In addition, there was a significant main effect for Error Type, $F(1, 63) = 22.93, p < .001, \eta^2 = .27$ and Error Severity, $F(1, 63) = 12.91, p < .001, \eta^2 = .17$. Post-hoc comparisons with the Tukey HSD criterion for significance showed FAME condition ($M = 76.99, SD = 4.51$) produced significantly lower reliance

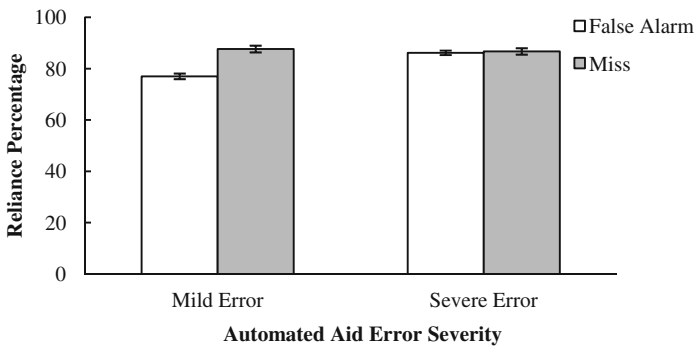


Fig. 3. Reliance on the AA across error types. Error bars are standard errors

when compared to MSE ($M = 86.69$, $SD = 5.15$), MME ($M = 87.62$, $SD = 5.42$), and FASE ($M = 86.17$, $SD = 3.61$) conditions $p < .001$ (Fig. 3).

3.4 IED Search Performance

Percentages of hits (i.e., percentage of correctly identified IEDs) were calculated for each condition. Between-subjects ANCOVA with the PTM variable as a covariate revealed a significant interaction between Error Type and Error Severity factors, $F(1, 63) = 17.70$, $p < .001$, $\eta^2 = .22$. In addition, there was a significant main effect for Error Type, $F(1, 63) = 8.74$, $p < .05$, $\eta^2 = .12$ and Error Severity, $F(1, 63) = 4.87$, $p < .05$, $\eta^2 = .07$. Post-hoc comparisons with the Tukey HSD criterion for significance showed FAME condition ($M = 52.94$, $SD = 15.72$) produced significantly lower hit percentage when compared to MSE ($M = 72.35$, $SD = 19.85$), MME ($M = 81.18$, $SD = 15.36$), and FASE ($M = 78.15$, $SD = 16.06$) conditions $p < .05$ (Fig. 4).

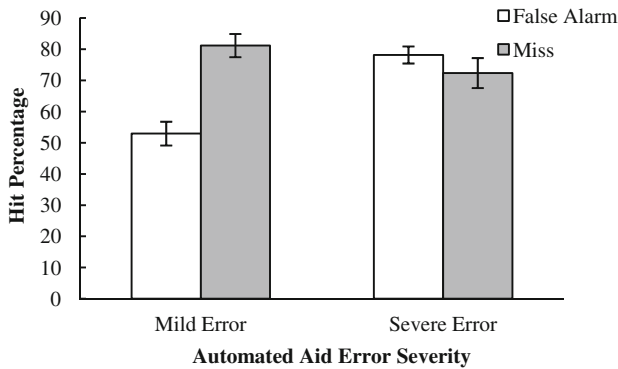


Fig. 4. IED hits across the error types. Error bars are standard errors

Percentages of false alarms (i.e., percentage of objects incorrectly identified as IEDs) were calculated for each condition. Between-subjects ANCOVA with the PTM variable as a covariate revealed a significant interaction between Error Type and Error Severity factors, $F(1, 63) = 9.66$, $p < .05$, $\eta^2 = .13$. In addition, there was a significant main effect for Error Type, $F(1, 63) = 7.30$, $p < .05$, $\eta^2 = .10$ and Error Severity, $F(1, 63) = 17.28$, $p < .01$, $\eta^2 = .22$. Post-hoc comparisons with the Tukey HSD criterion for significance showed FAME condition ($M = 4.76$, $SD = 1.60$) produced significantly higher false alarm percentage when compared to MSE ($M = 1.00$, $SD = 1.86$), MME ($M = 1.63$, $SD = 3.15$), and FASE ($M = 0.88$, $SD = 1.51$) conditions $p < .01$ (Fig. 5).

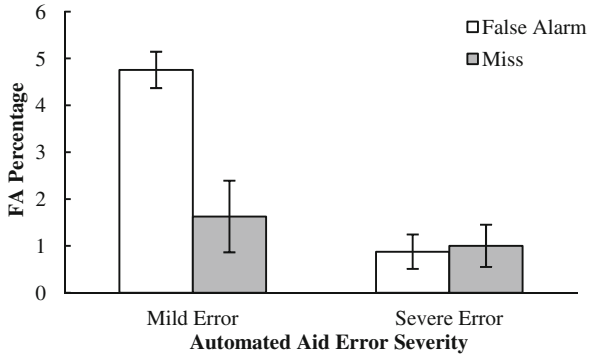


Fig. 5. IED false alarms across error types. Error bars are standard errors

3.5 Global Workload

Between-subjects ANCOVA with the PTM variable as a covariate showed no significant main effects and no interaction for the Error Type and Error Severity factors for the workload observed. The average level of workload across all of the conditions was 44.12 ($SD = 12.70$).

4 Discussion

The main focus for the study was to examine the joint effects of automation error type (FA and misses) and error severity (mild and severe) on operators' trust and reliance. It was predicted that severe false alarm errors would result in lowest trust and reliance while mild misses would produce the highest trust and reliance as found in previous studies by [5, 8, 9]. The results did not confirm the original hypothesis. The participants did not report any significant differences in trust for all conditions, with trust being moderate on average. The participants relied on the AA the least in the FAME condition when compared to the other three conditions. In this condition, participants relied on the AA in approximately 75 % of the cases (however, not necessarily correctly) while the other three conditions had reliance levels of slightly above 85 %. The lowest level of perceived reliability was observed in MSE condition, not in the FASE condition as suggested by previous research, with perceived reliability rated slightly above 50 %. The participants' evaluations of reliability were closer to the actual AA reliability (85 %) for the other three conditions. The MSE reliability value was significantly lower when compared to the MME and FASE conditions. Finally, in terms of performance, the FAME condition produced the lowest number of hits and highest number of FAs indicating unexpected consequence of a mild error on performance.

Overall, results from our study indicated complex interactions between error types and error severity. Trust literature suggests that there is a one-directional relationship between perceptions of automation trustworthiness (in this study, measured by perceived reliability metric), trust in automation, and reliance on the automation (agreement with

the suggestions of the AA) [3]. For example, the perception of an automated system being low in reliability is expected to be accompanied by low trust and, consequently decreased use of the system. The results indicate that there may not necessarily be a positive correlation between trustworthiness, subjective trust, and trust outcome reliance behavior. In this study FAME was rated the lowest in perceived reliability; however, we did not observe matching outcomes in trust and reliance. Moreover, it appears that mild error FAME produced lowest system reliance as well as the poorest performance; however, it was not rated in the lowest reliability and trust. This finding indicates that errors irrelevant to the primary task (IED search) might have a detrimental effect on the task performance. The participants tended to under-rely appropriately (meaning agree with the AA when it was correct). Minor failures can cause a negative impact on reliance calibration and performance. Ultimately, this study provides insight into the patterns of trust and reliance on the systems that commit different types of errors which may negatively affect performance in human-machine teams. These results suggest that human-machine systems should be tailored to account for automation errors.

There are limitations associated with the experiment. The effects of error and severity types can be task-dependent. For example, [5] used a collision avoidance task in a simulated agricultural environment. In that study, the consequences for both types of errors (Misses and FAs) were equated to remove selection bias. We attempted to equate the consequences for both error types by using verbal instructions during the training. However, if the instructions were insufficient, the participants might have had their bias set to weight misses as more important when compared to FAs, thus affecting the results. The duration of the task could have been longer to allow for the capturing of changes in trust as participants become more familiar with the aid. Also, it would be beneficial to introduce pauses during the task to measure changes in trust and reliance as the task progresses. In addition, psychophysiological measurements such as eye-tracking and electroencephalogram could provide additional information on the neurophysiological processes occurring in relevance to the task.

Acknowledgments. The authors would like to thank Dr. Nathan Bowling, Wright State University (Dayton, Ohio) for support of this research and for comments on this manuscript.

References

1. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man, Cybern.* **30**, 286–297 (2000)
2. 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, NY (2009)
3. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. *Hum. Factors* **46**(1), 50–80 (2004)
4. 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)
5. Sanchez, J.: Factors that affect trust and reliance on an automated aid (Unpublished doctoral dissertation). Georgia Institute of Technology, Atlanta, GA (2006)

6. Madhavan, P., Wiegmann, D., Lacson, F.: Automation failures on tasks easily performed by operators undermine trust in automated aids. *Hum. Factors* **48**(2), 241–256 (2007)
7. Khasawneh, M.T.: Effect of error severity on human trust in hybrid systems. In: *Human Factors and Ergonomics Society 48th Annual Meeting Proceedings*. New Orleans, LA, pp. 439–433 (2004)
8. Dixon, S., Wickens, C.D.: Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high workload. *Hum. Factors* **48**, 474–486 (2006)
9. Geels-Blair, K., Rice, S., Schwark, J.: Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *Int. J. Aviat. Psychol.* **3**, 245–266 (2013)
10. Sanchez, J., Rogers, W.A., Fisk, A.D., Rovira, E.: Understanding reliance on automation: Effects of error type, error distribution, age, and experience. *Theor. Issues Ergon. Sci.* **15**(2), 134–160 (2014)
11. 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)
12. 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 (2012)
13. Lyons, J.B., Koltai, K.S., Ho, N.T., Johnson, W.B., Smith, D.E., Shively, J.R.: Engineering trust in complex automated systems. *Ergonomics in Design* (in press)
14. 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)
15. Guznov, S., Nelson, A., Lyons, J., Dycus, D.: The effects of automation reliability and multitasking on trustworthiness, trust, and reliance in a simulated unmanned system control task. In: *Proceedings of 17th Human-Computer Interaction Conference*, pp. 616–621 (2015)