Emerging Technologies and Future of Work, Vol. 117, 2023, 178–187
https://doi.org/10.54941/ahfe1004417

Effects of System Reliability on Workload and Performance in Image Recognition Tasks

Xiaodong Xu[1], Liang Ma[1], Yun Zhang[3], and Cheng Xu[2,3]

[1]Department of Industrial Engineering, Tsinghua University, Beijing, China
[2]Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China
[3]Beijing Electro-Mechanical Engineering Institute, Beijing, China

ABSTRACT

Autonomy has found wide-ranging applications, yet its imperfect nature necessitates human oversight and intervention. Investigating autonomy's impact on the operator is pivotal for enhancing human-machine system performance and safety. This study analyzes the effects of autonomous system reliability on operator task performance and mental workload in the context of vehicle type recognition. Experimental findings reveal that autonomy with 90% reliability significantly reduces task completion time and lessens subjective workload. Autonomy with 70% reliability supports the participants, while 50% reliability hampers them, although insignificantly. The reliability threshold for autonomy to have no effect on the participants is around 55%. Autonomy reliability's influence on the operator lies in altering task completion strategies — an all-or-none approach that accelerates task processing speed without improving overall response accuracy. The experiment yielded insights applicable to the design of assistive autonomous systems and the allocation of human-machine functions in real-world tasks.

Keywords: Autonomy, Reliability, Mental workload, Performance

INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are experiencing a surge in usage across diverse fields, such as health, industry, agriculture, and the military (Ayamga, Akaba, & Nyaaba, 2021). Image recognition and classification is a mission-critical aspect of modern human-UAV system interaction. Integrating autonomous systems, like computer vision technology, is pivotal in reducing manpower and personnel costs, enhancing efficiency, and minimizing variability (Balfe, Sharples, & Wilson, 2015). Nevertheless, since autonomous systems are not infallible, operator confirmation of the autonomous recognition results remains imperative to ensure the system's overall performance and prevent potential safety and liability incidents (Parasuraman & Riley, 1997).

The relationship between autonomy and operator mental workload is an essential consideration to efficiency and safety (Kantowitz & Campbell,

1996). Mental workload refers to "the portion of operator information processing capacity or resources that is required to meet system demands" (Eggemeier, Wilson, Kramer, & Damos, 2020), which is commonly assessed via subjective reports, the primary and/or the secondary task performance and physiological metrics (Young, Brookhuis, Wickens, & Hancock, 2015). The effects of the reliability of autonomy on operator mental workload and performance during human-machine collaboration are not entirely clear. Some studies have suggested that mental workload and performance improve as the degree of autonomy reliability increases (Balfe et al., 2015; Chavaillaz, Wastell, & Sauer, 2016). However, it has also been reported that more reliable autonomous systems can lead to over-reliance and complacency (Parasuraman, Mouloua, Molloy, & Hilburn, 1996), potentially reducing the operator's situation awareness and resulting in lower task performance and increased mental workload (Oakley, Mouloua, & Hancock, 2003), which is especially pronounced during sudden abnormal situations or phases when the task complexity is already high (Frazier, McComb, Hass, & Pitts, 2022).

A study conducted by Wickens and Dixon (2007) presented a quasi-meta-analysis that demonstrated when autonomous systems operate at reliability levels below 70%, they result in poorer system performance compared to situations where no autonomy is present. Conversely, higher reliability yields more favorable outcomes. However, the review did not accurately define the concept of reliability, such as misses, false alarms, or mixed errors. Additionally, it also failed to provide a comprehensive distinction between various task domains and types, which encompasses multiple domains, including aviation, driving, and the military, coupled with diverse task types such as monitoring, diagnosing, and controlling. Consequently, there exists a necessity for further validation of the determined threshold's applicability and universality. Furthermore, most pertinent literature focuses on concurrent tasks, thus posing challenges when attempting to directly evaluate the influence of autonomy reliability on a specific task. Since intricate decision processes can be deconstructed into a sequence of atomic binary decisions, the conclusions drawn from simple independent binary decision tasks can be more seamlessly extrapolated to other decision-making contexts (Yu, Berkovsky, Taib, Zhou, & Chen, 2019).

In summary, this study addressed two critical questions based on the dichotomous task of vehicle type recognition: (1) How does the introduction of autonomy affect the operator's mental workload and performance? (2) What is the relationship between the effects and the autonomy's reliability (specifically, false alarms)? Examining the effects of imperfect autonomous systems on operator mental workload and performance carries significant implications for vital aspects like the design of assistive autonomous systems and the allocation of human-machine functions.

METHODOLOGY

Task and Materials

In this experiment, the participants were exclusively tasked with a singular objective: discerning whether the vehicle depicted in the image is a passenger

or non-passenger vehicle. They were then required to provide corresponding keypress responses ("C" for passenger vehicles and "N" for non-passenger vehicles) based on their identification decision. Passenger vehicles include taxis, private cars, and buses, while non-passenger vehicles include engineering and military vehicles. All experimental images were sourced from pertinent YouTube videos, and the experimental platform (see Figure 1), named ATLP, was developed using MATLAB 2022b.

Figure 1: Diagram of the experimental platform. On the left, a vehicle image awaits participant identification. The image was converted to grayscale to simulate the UAV's perspective, with a red box highlighting the target vehicle for localization simulation. The right side presents essential data: participant ID, experimental phase, current autonomy reliability level, and elapsed round time. Below, conveniently accessible response buttons, operable via the keyboard, offer participant interaction.

Design and Procedure

The experimental study focused on the independent variable of assisted autonomous system reliability, with a total of four distinct levels: 0%, 50%, 70%, and 90%. 0% indicates that no autonomous system was involved, and all judgments needed to be completed by the participants themselves. In contrast, the three other reliability levels entailed the presentation of images pre-filtered by the autonomous system, and each finally presented to the participants was consistently labeled as the passenger vehicle. Nonetheless, owing to the autonomy's reliability not reaching 100%, the prospect remained that non-passenger vehicle images might be presented. In the experimental conditions involving the autonomy, the participants could either directly entrust the autonomy's recognition outcome and press the key of "C" upon image display or, alternatively, re-evaluate the image and respond by pressing a key corresponding to their own recognition results.

The reliability of the autonomous system was attained and calibrated through random number generation. For instance, in the case of the

autonomous system operating at 70% reliability, the system first generated a random number within the range of 0 to 1 before each presentation of a vehicle image to the participants. A non-passenger vehicle was displayed if this random number exceeded 0.7; otherwise, a passenger vehicle was shown. Hence, the three levels mentioned above of autonomy reliability represent conceptual averages rather than static constants. The incorporation of randomness serves to more faithfully replicate the utilization of image recognition algorithms within real-world scenarios.

This study utilized a within-group design, requiring all participants to complete the recognition and response task under the four experimental conditions mentioned earlier. Each condition consisted of 30 rounds, with independent content in every round and randomized image presentation sequences. A Latin-square design was employed for the three autonomy-assisted conditions to eliminate the impact of the condition order. However, all participants must complete the experiment without autonomy assistance first. Before each experimental phase, the participants were informed about the autonomy reliability level. The entire experiment took approximately 30 minutes.

Dependent Measures

The experiment's dependent variables comprised task performance and subjective workload perception. Task performance was evaluated through task completion time and image recognition accuracy, both of which were automatically recorded by the ATLP system. Participants' subjective workload perception was assessed using the NASA-TLX (Hart, 2006), a widely utilized scale that evaluates operator workload across six distinct dimensions. This scale was filled out upon the completion of each experimental condition.

Participants

A total of 36 students from Tsinghua University participated in this experiment (age: $M = 24.4$, $SD = 2.8$), including 15 male and 21 female students. All participants had normal or corrected-to-normal vision. Before the experiment, each participant received an initial introduction to the experiment's purpose and procedures. They were instructed to complete the experiment with optimal speed and accuracy. Subsequently, informed consent was obtained through signed consent forms. Following task completion, each participant received a reward of ¥30 (approximately $5).

Data Processing

Due to the participation of over 30 individuals and the implementation of a within-group design with an equal distribution of participants across each experimental condition, the parametric tests exhibited robustness. Consequently, repeated measures ANOVA and paired t-tests with Bonferroni adjustments were combined to assess the distinctions among the four experimental conditions. Furthermore, an initial examination of the relationship between autonomy reliability and operator performance, and mental workload was conducted through linear regression analysis.

RESULTS

Figure 2 illustrates the findings regarding the participants' task performance and subjective workload perception across reliability conditions (R0, R50, R70, and R90). Examination of recognition accuracy indicated no significant differences among all groups ($F(3, 105) = 0.82$, $p = .48$, $\eta^2 = 0.01$), and all groups exhibited recognition accuracy exceeding 95% (Table 1). However, task completion time significantly differed across reliability conditions ($F(3, 105) = 22.83$, $p < .01$, $\eta^2 = 0.10$). Post hoc analyses showed that those in the R90 condition achieved notably shorter task completion time (Table 2), while differences among the other three conditions were not statistically significant.

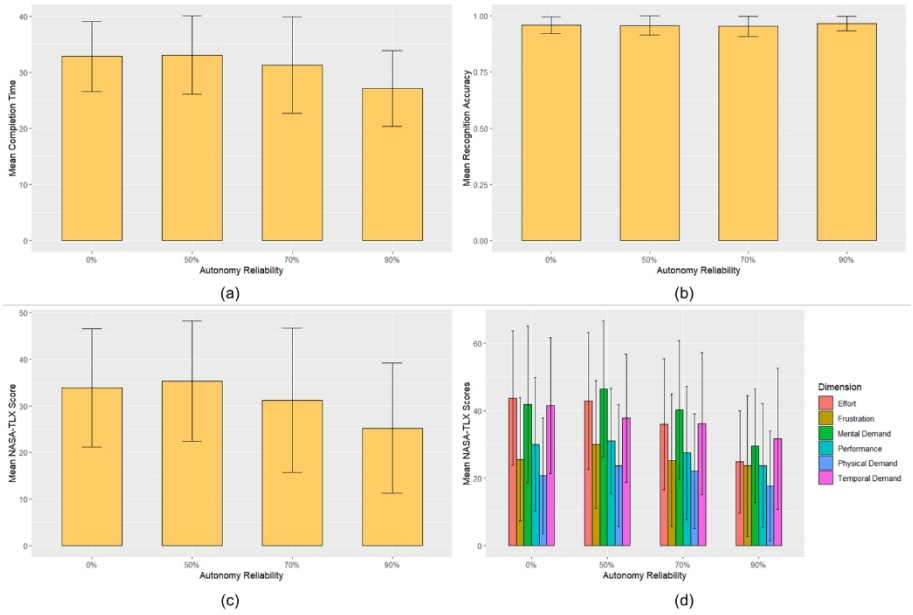

Figure 2: Performance and mental workload across experimental conditions: (a) Mean task completion time; (b) Mean recognition accuracy; (c) Mean NASA-TLX score; (d) Mean scores on the six dimensions of the NASA-TLX.

Subsequent analysis of subjective workload data revealed significant differences among all groups concerning the overall NASA-TLX score ($F(3, 105) = 10.60$, $p < .01$, $\eta^2 = 0.08$) and its four subscales, including mental demand ($F(3, 105) = 11.58$, $p < .01$, $\eta^2 = 0.08$), physical demand ($F(3, 105) = 6.83$, $p < .01$, $\eta^2 = 0.02$), temporal demand ($F(3, 105) = 3.61$, $p = .02$, $\eta^2 = 0.03$), and effort ($F(3, 105) = 15.52$, $p < .01$, $\eta^2 = 0.14$). Further post hoc assessments indicated that the R90 group attained the lowest scores on all the dimensions mentioned above. Remarkably, their scores were significantly lower than the other three groups in mental demand and effort, significantly lower than the R0 and R50 groups in the NASA-TLX, and only significantly lower than the R50 group in physical demand. The R50 group

obtained the highest scores on the NASA-TLX, mental and physical dimensions, with significant difference observed solely in the NASA-TLX. More details on post hoc analysis are shown in Table 2.

Table 1. Descriptive statistics for performance and mental workload: *M(SD)*.

Reliability	0% (R0)	50% (R50)	70% (R70)	90% (R90)
Accuracy (%)	95.9 (3.7)	95.7 (4.3)	95.4 (4.5)	96.6 (3.2)
Time (s)	32.8 (6.3)	33.0 (7.0)	31.3 (8.6)	27.1 (6.8)
NASA-TLX	33.9 (12.7)	35.3 (12.9)	31.2 (15.5)	25.2 (14.0)
Mental Demand	41.8 (23.2)	46.4 (20.2)	40.3 (20.5)	29.6 (16.9)
Physical Demand	20.7 (17.2)	23.8 (18.1)	22.1 (17.0)	17.6 (16.3)
Temporal Demand	41.5 (20.1)	37.8 (19.0)	36.1 (21.1)	31.7 (21.0)
Performance	30.0 (19.8)	31.1 (15.5)	27.5 (19.8)	23.8 (18.4)
Effort	43.8 (19.9)	42.9 (20.3)	35.9 (19.4)	24.9 (15.1)
Frustration	25.6 (18.2)	30.0 (18.9)	25.3 (19.6)	23.6 (20.8)

Table 2. The significant part of post hoc test results.

	Condition 1	Condition 2	t	p.adj
Time	R0	R90	5.95	<.01
	R50	R90	7.72	<.01
	R70	R90	5.12	<.01
NASA-TLX	R0	R90	3.95	<.01
	R50	R70	2.98	.03
	R50	R90	4.61	<.01
Mental Demand	R0	R90	4.40	<.01
	R50	R90	5.36	<.01
	R70	R90	3.35	.01
Effort	R0	R90	6.06	<.01
	R50	R90	5.28	<.01
	R70	R90	3.79	<.01
Physical Demand	R50	R90	4.47	<.01

Using R0 as the baseline condition, the differences between the experimental outcomes of the other three conditions and the baseline condition were quantified. Assuming the simplest linear relationship between autonomy reliability and operator performance and mental workload, the experimental results were linearly fitted, with task completion time and NASA-TLX score differences serving as functions of autonomy reliability, respectively (Figure 3). The regression equation for task completion time (Time) in relation to autonomy reliability (Reliability) was $Time = -14.88 \times Reliability + 8.09$ ($F = 24.39$, $p < .01$). Since neither variable followed a normal distribution, Spearman's correlation yielded $r = -0.45$ (p <. 01), indicating an 18% variance attributable to reliability. The regression equation for NASA-TLX score (Score) in relation to autonomy reliability (Reliability) was $Score = -25.35 \times Reliability + 14.42$ ($F = 13.85$, $p < .01$).

Similarly, a correlation of $r = -0.35$ ($p < .01$) emerged, with reliability accounting for 11% of the variance. Notably, the predicted crossover points where *Time* = 0 lay at *Reliability* = 0.544, and *Score* = 0 lay at *Reliability* = 0.569.

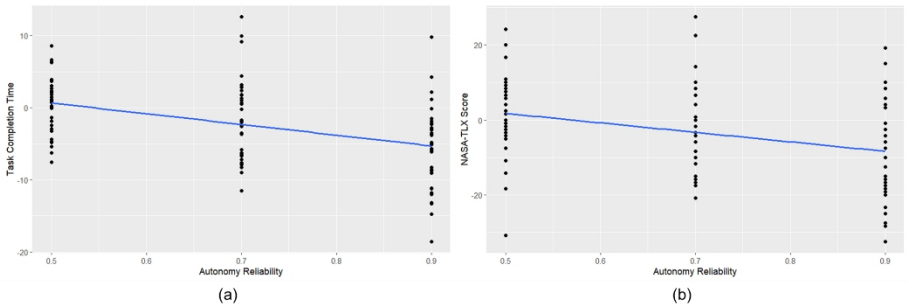

Figure 3: Regression analysis of (a) task completion time and (b) NASA-TLX score on autonomy reliability.

DISCUSSION

The participants did not attain 100% accuracy across all experimental conditions, indicating that the presence of autonomy and its reliability are not the primary factors influencing the participants' image recognition accuracy. Some participants noted post-experiment that they struggled to discern the vehicle type in certain images and had to make random selections. Additionally, instances of continuous key presses and erroneous inputs were reported during the experiment. Participant unfamiliarity with the experimental materials and misoperation under time constraints may underlie the fluctuations in image recognition accuracy. Importantly, this factor remains constant and unaffected by the experimental conditions.

Compared to the other experimental conditions, the R90 group exhibited significantly shorter task completion time and significantly lower NASA-TLX scores. This implies that autonomy with 90% reliability significantly reduces the participants' mental workload, predominantly in the dimensions of mental demand and effort. The participants achieved the same near-perfect recognition accuracy as the baseline condition (R0) with fewer attention resources. Both objective performance and subjective scores for the R70 and R50 groups did not significantly differ from those of the R0 group. However, examining the mean values, both task completion time and NASA-TLX scores were comparatively higher for the R50 group in contrast to the baseline condition and lower for the R70 group. This suggests that autonomy with 50% reliability had a disruptive effect on the participants, while autonomy with 70% reliability had a slightly assistive impact, although neither reached statistical significance.

The effects of autonomy on operator performance and mental workload may stem from altering the operator's task completion strategy. When presented with an image, the participants can opt to trust the system's judgment

and respond by pressing "C" (Path 1). Alternatively, the participants can disregard the system's results and respond based on their own judgment (Path 2). Another option is to combine the system's judgment and their own before responding (Path 3). Of course, there are additional paths that might not be immediately recognizable. Under different paths, the extent to which the operator completely disregards or attentively considers certain signals, as well as the number of signals involved, varies. Various autonomy reliabilities influence the proportion of the operator selecting these task-processing paths. Higher autonomy reliability may lead the participants to rely more on the machine, overlooking more image-related information, and reacting solely to pivotal and uncomplicated machine-generated judgments, hence favoring Path 1. Conversely, the participants might lean towards Paths 2 and 3 under lower autonomy reliability. Autonomy reliability impacts the operator's reliance and trust in the machine, consequently reshaping resource allocation and task execution strategy. This all-or-none approach enables quicker task completion at higher reliability levels by allowing the user to address more rounds (Ferraro & Mouloua, 2021). However, it does not enhance overall response accuracy.

Furthermore, the presence of autonomy might introduce additional, albeit unnecessary, compelling information that the operator must attend to. Since this experiment is dichotomous, the significance of autonomy's presence becomes negligible when autonomy's reliability is at 50%—in other words, experiments R0 and R50 are nearly indistinguishable. Despite this, the experimental results reveal a negative impact of the autonomous system. This implies that the operator cannot entirely disregard the low-reliability autonomy's presence and act with complete independence in forming judgments. This underscores the importance of prioritizing autonomy reliability when integrating autonomy aids into real-world tasks.

The linear regression results exhibit a notable inverse correlation between autonomy reliability and operator task completion time and NASA-TLX scores. The threshold marking the boundary between the disruptive and assistive impacts of autonomy is approximately 55%. Interestingly, this threshold is significantly lower than what the previous review suggested (Wickens & Dixon, 2007), which might relate to variations in task type, quantity, and the specific definition of reliability.

There are two limitations to this experiment. Firstly, the task duration may have been too brief, or the number of experimental rounds too limited. Establishing a stable perception of the machine requires a certain time frame, allowing continuous adjustments to culminate in a more steadfast task-processing strategy. The task strategy's volatility could contribute to the lack of significant differences between the R50, R70, and R0 groups. Existing literature suggests that around 30 rounds are needed to solidify a stable task-processing strategy (Yu et al., 2019), though this value may be specific to the task. In future, preliminary experiments and pilot studies could be employed to determine the required number of rounds or time duration for strategy stabilization prior to conducting formal experiments.

Secondly, the experimental design encompassed a limited number of independent variable levels. This hindered a direct comparison of operator task

performance and subjective workload perception across various reliability levels. Moreover, the data's sparsity posed challenges for accurate linear fitting to model the relationship between reliability and mental workload. This complexity extended to exploring the reliability level of assisted autonomy when achieving baseline performance equivalency. For future investigations, increasing the number of reliability levels, incorporating physiological measurements, and employing nonlinear models for fitting experimental results could yield more precise relationship models and threshold delineations. Moreover, endeavors could be made to identify the operator's task completion strategies based on physiological metrics, verifying assumptions about underlying influence mechanisms.

CONCLUSION

Given that human judgment and decision-making flexibility surpasses that of computers, humans will continue to be an integral facet of unmanned systems. Hence, studying human capabilities, constraints, and modes of human-machine interaction within such systems is necessary. This investigation delves into the effects of autonomous systems on operator task performance and mental workload during a vehicle type recognition task. Results indicate no significant difference in operator recognition accuracy across experimental conditions. However, a pronounced negative correlation emerges between autonomy reliability and operator task completion time and NASA-TLX scores. Autonomy with 90% reliability significantly reduces task completion time and subjective workload; autonomy with 70% reliability partially assists the participants, and autonomy with 50% reliability disrupts the participants' performance and mental workload, although insignificantly. The autonomy reliability threshold without any effect on the participants is approximately 55%. Autonomy reliability's influence on the operator hinges on modifying their task completion strategies, encompassing attention allocation and path selection ratios. This all-or-none approach enhances task processing speed but fails to improve response accuracy. This research highlights the significance of emphasizing autonomy assistance reliability during its integration into real-world tasks. Future considerations encompass increasing task round numbers and independent variable levels for more direct autonomy reliability effect comparisons. Moreover, exploiting physiological metrics to discern diverse operator task processing strategies could validate effect mechanism assumptions. A more precise functional relationship model concerning autonomy's impact on the operator could emerge by combining assorted data types. This study informs assistive autonomy system design and human-machine function allocation in real-world tasks, while paving the way for exploratory adaptive autonomy research.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from Science Foundation of Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, and National Natural Science Foundation

of China with the Research Project on Human Behavior in Human-Machine Collaboration (Project Number: 72192824).

REFERENCES

Ayamga, M., Akaba, S., & Nyaaba, A. A. (2021). Multifaceted applicability of drones: A review. Technological Forecasting and Social Change, 167, 120677. Doi: https://doi.org/10.1016/j.techfore.2021.120677.

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. Applied Ergonomics, 47, 52–64. Doi: https://doi.org/10.1016/j.apergo.2014.08.002.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. Applied Ergonomics, 52, 333-342. Doi: https://doi.org/10.1016/j.apergo.2015.07.012.

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (2020). Workload assessment in multi-task environments. In Multiple task performance (pp. 207–216): CRC Press.

Ferraro, J. C., & Mouloua, M. (2021). Effects of automation reliability on error detection and attention to auditory stimuli in a multi-tasking environment. Applied Ergonomics, 91, 103303. Doi: https://doi.org/10.1016/j.apergo.2020.103303.

Frazier, S., McComb, S. A., Hass, Z., & Pitts, B. J. (2022). The Moderating Effects of Task Complexity and Age on the Relationship between Automation Use and Cognitive Workload. International Journal of Human–Computer Interaction, 1–19.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Kantowitz, B. H., & Campbell, J. L. (1996). Pilot workload and flightdeck automation. Automation and human performance: Theory and applications, 117–136.

Oakley, B., Mouloua, M., & Hancock, P. (2003). Effects of Automation Reliability on Human Monitoring Performance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47(1), 188–190. Doi:10.1177/154193120304700139.

Parasuraman, R., Mouloua, M., Molloy, R., & Hilburn, B. (1996). Monitoring of automated systems. In Automation and human performance (pp.91–115): CRC Press.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human factors, 39(2), 230–253. Doi:10.1518/001872097778543886.

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. Doi:10.1080/14639220500370105.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. Ergonomics, 58(1), 1–17. Doi:10.1080/00140139.2014.956151.

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do i trust my machine teammate? an investigation from perception to decision. Paper presented at the Proceedings of the 24th International Conference on Intelligent User Interfaces.