Global Issues Challenge: Challenges in AI at the Human Level, Vol. 234, 2026, 33–47
https://doi.org/10.54941/ahfe1007456

Human–AI Collaboration in Automated X-Ray Screening: Effects of Alarm Types and Reliability Levels on Operator Performance in Subway Security

Xin Zhou

Department of Industrial Engineering, Tsinghua University, China

ABSTRACT

Security screening of passenger baggage is critical to public safety in metro systems, yet the performance of the screening system as a whole depends heavily on how automated advice is communicated to human operators. This study examined how the format of diagnostic advice provided by an automated diagnostic aid system (ADAS) should be matched to the system's reliability in a simulated subway X-ray screening task. Three alarm types—binary alarm, likelihood alarm, and automated decision—were crossed with three reliability levels (70%, 80%, and 90%) in a mixed factorial design, with alarm type manipulated within subjects and reliability between subjects (18 valid datasets, $n = 6$ per reliability group). Operator performance was measured with signal detection sensitivity (d') and response times for target-present and target-absent decisions; subjective trust was assessed on five dimensions. A two-way mixed analysis of variance on d' revealed a significant alarm type × reliability interaction, $F(2, 15) = 7.05$, $p = .007$, $\eta p^2 = .48$, whereas neither main effect was significant. Operators' sensitivity increased monotonically with reliability under the binary alarm, but under the likelihood alarm it peaked at 80% and deteriorated at 90% reliability, falling below the sensitivity of the ADAS alone. Subjective trust tracked the objective pattern: the likelihood alarm was preferred when reliability was low, whereas the automated decision attracted the highest trust when reliability was high. The findings suggest that alarm transparency should be adapted to automation reliability.

Keywords: Human–AI collaboration, X-ray screening, Diagnostic aid, Alarm design, Trust in automation, Signal detection theory

INTRODUCTION

X-ray security checkpoints are widely deployed at subway entrances to detect dangerous items in passengers' baggage and to maintain public safety. Typically, a screener judges whether a piece of baggage contains forbidden items by inspecting its X-ray image. Because subway traffic is tremendous, both the efficiency and the accuracy of the screening operation are essential. The identification and decision work, however, is mainly performed by human screeners and can be characterized as a visual search task (Yu et al., 2017). Such tasks require screeners to remain vigilant over long periods, which easily induces visual fatigue and degrades accuracy

Received June 14, 2026; Revised June 29, 2026; Accepted July 7, 2026; Available online July 20, 2026

(Hancock et al., 2013). With recent technological advances, automated diagnostic aid systems (ADAS) have been introduced into the screening task. Although an ADAS can substantially enhance screeners' performance and reduce their workload, it also raises new questions about human–automation interaction. In the remainder of this introduction, we first summarize previous research on visual search in baggage screening, then discuss the relationship between human operators and automated aids, and finally outline the present study.

Visual Search in X-Ray Baggage Screening

Visual search in baggage screening requires the screener to inspect X-ray images, search for forbidden items, and make a binary decision (Schwaninger, 2005). When performing this task, the screener must overcome four challenges: limited target visibility, an unknown target set, the possible presence of multiple targets, and low target prevalence (Biggs & Mitroff, 2015). To identify forbidden items correctly, the screener has to memorize the target set and its X-ray appearance in advance (Schwaninger, 2005). Humans are at a disadvantage relative to automation in several respects: screeners must remain vigilant to detect rarely occurring targets, and prolonged screen exposure causes visual fatigue, so the efficiency and accuracy of screening are ultimately bounded by human performance. Automated systems, by contrast, excel at memorization and are immune to fatigue, and can therefore compensate for human weaknesses. However, introducing automation creates new problems of its own, most notably misuse and disuse of the aid. Consequently, the relationship between the human operator and the automated system must be considered explicitly.

Automated Diagnostic Aid Systems

An automated assistance system is a computer system that supports or replaces tasks performed by humans (Parasuraman & Wickens, 2008), and the diagnostic aid is one major category of such systems. By issuing alarms or alerts in various perceptual forms, a diagnostic aid changes how operators allocate attention (Cullen et al., 2013). In subway screening, the system indicates potentially threatening objects in the X-ray images of passenger baggage and provides diagnostic advice in the form of target-present or target-absent judgements. The operator may accept or reject the advice before making the final decision. To date, such systems have mainly been investigated in laboratory studies with student participants rather than in field operations (Rice & McCarley, 2011). The principal expected benefit of automation is a considerable reduction in operators' mental and physical workload and, in turn, in human error. Yet automation can also introduce new problems and may even paradoxically increase mental workload (Wiener & Curry, 1980). Research on human–automation interaction has shown that operators'

cognitive tasks change as automation is introduced, and that—contrary to intuition—the human role becomes more rather than less important as automated systems evolve (Parasuraman & Wickens, 2008): fewer operators end up bearing greater responsibility for more difficult final decisions.

Signal Detection Theory

Signal detection theory (SDT) provides a framework for analysing decisions between two overlapping states of the world, signal and noise (Wickens et al., 2013), and is widely used to evaluate detection performance achieved by human operators, by machines, or by human–machine combinations. Each detection trial yields a binary response ("yes", the signal is present, or "no", the signal is absent), and the combination of the response with the true state of the world produces one of four outcomes: hit, miss, false alarm, or correct rejection (see Figure 1). Detection performance is summarized by the sensitivity index

$$d' = z(\mathrm{H}) - z(\mathrm{FA}) \tag{1}$$

where H and FA denote the hit rate and the false-alarm rate, respectively. Sensitivity corresponds to the distance between the means of the noise and signal-plus-noise distributions; because the two distributions inevitably overlap, no detection system is perfect, and better systems are characterized by larger d' values.

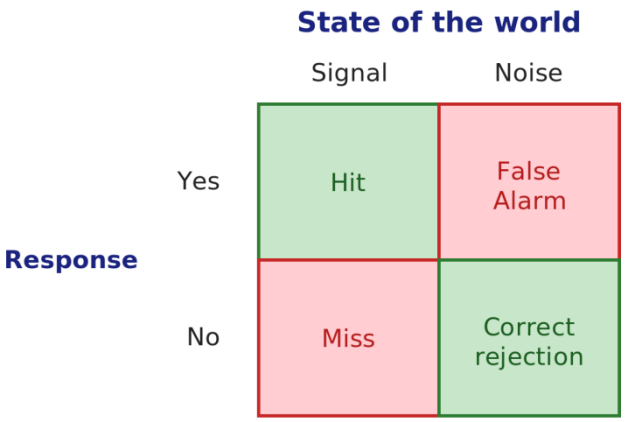

Figure 1: Four outcomes of a signal detection judgement.

When an ADAS is introduced into the detection task, the operator's cognitive process changes. Without the aid, the operator works only from the raw image and must inspect every X-ray picture carefully. With the aid, the operator receives a diagnostic reference that lowers mental and physical workload, but a new factor emerges: the operator's attitude toward the automation strongly influences how the aid is used and, consequently, the joint performance of the human–machine system (see Figure 2).

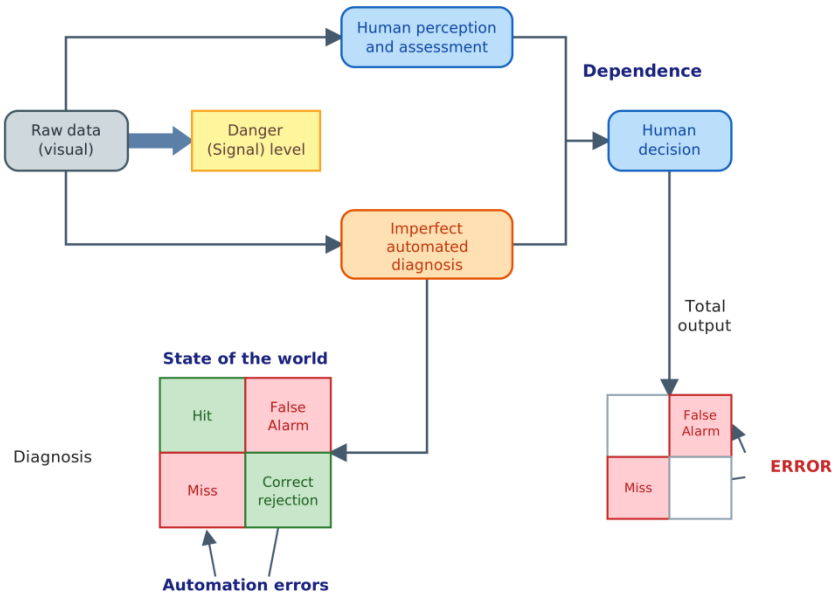

Figure 2: Model of the human–machine interaction system.

The Present Study

Although diagnostic aids are increasingly available, it remains unclear how the format of their advice should be matched to their reliability. The present study addressed this question by comparing three alarm types—binary alarm, likelihood alarm, and automated decision—across three reliability levels (70%, 80%, and 90%) in a simulated subway X-ray screening task. Specifically, we examined two research questions. First, how do alarm type and ADAS reliability jointly affect operators' detection sensitivity and response times? Second, how does operators' subjective trust in the three alarm types vary with the reliability level? Answers to these questions can inform the design of alarm policies for human–AI teaming in safety-critical screening operations.

METHOD

Participants

Twenty-one graduate students were recruited from the community of Tsinghua University and were paid 30 RMB each for taking part. One participant's accuracy fell below that of the ADAS, suggesting that he either did not understand or did not comply with the instructions, and his data were excluded. The final sample therefore comprised 18 valid datasets (10 males, 8 females; mean age = 23.8 years), with six participants in each reliability group. All participants reported normal or corrected-to-normal vision.

Apparatus and Materials

Stimulus presentation and response collection were controlled by a PC running E-Prime. Stimuli were presented on a 13.3-inch LCD monitor with a resolution of 1024 × 768 pixels and a refresh rate of 60 Hz. The stimulus materials were colored X-ray images of baggage containing everyday objects, rendered on a white background. The forbidden items included knives, guns, scissors, and cutters, and different items corresponded to different levels of search difficulty. The target prevalence was set to 30%, that is, 30% of the images contained a forbidden item, consistent with the base event rate used in previous benchmarking studies.

Alarm Configuration

Three types of ADAS advice were implemented. The binary alarm provided two messages, "Danger" and "Safe", according to whether the system judged a forbidden item to be present. The likelihood alarm provided two additional graded messages, "Warning" and "Possible safe", which signalled that the system had lower confidence in its judgement; images receiving these two messages were comparatively harder to search. The automated decision mode was identical to the binary alarm except that images judged "Safe" were dimmed and hidden from the operator, who therefore only needed to recheck the images flagged as "Danger". In the binary and likelihood alarms, "Danger" and "Safe" conveyed the same information. The meanings of all messages are listed in Table 1.

Table 1: Meanings of the alarm messages.

Alarm Message	Meaning
Danger (BA, LA, AD)	The system detects a forbidden item and is confident in its judgement.
Warning (LA)	The system detects a possible forbidden item; its confidence is lower than for "Danger".
Possible safe (LA)	The system fails to find a forbidden item but is not certain that the baggage is safe.
Safe (BA, LA)	The system is confident that the baggage is safe.

Note. BA = binary alarm; LA = likelihood alarm; AD = automated decision.

To parameterize the ADAS with SDT, two quantities were fixed in advance: the sensitivity of the system and its decision criteria. Three reliability levels (70%, 80%, and 90%) were defined as the percentage of correct system judgements among all judgements; the corresponding system sensitivities computed from Equation 1 are $d' = 1.05$, 1.68, and 2.56, respectively. For the likelihood alarm, "Danger" and "Warning" were classified as target-present advice, and "Possible safe" and "Safe" as target-absent advice. The first criterion $c1$, which separated "Danger" from "Safe", was set to 1.0 for both the binary and the likelihood alarm; two further criteria, $c2$ and

*c*3, separated "Danger" from "Warning" and "Possible safe" from "Safe", respectively. Given these parameters and the 30% target prevalence, the expected frequencies of each message for 100 trials were computed and are shown in Table 2.

Table 2: Alarm configuration settings (expected frequencies per 100 trials).

Reliability	Alarm Type	Message	Target Present	Target Absent
70%	Binary	Danger	21	21
		Safe	9	49
	Likelihood	Danger	9	4
		Warning	12	17
		Possible safe	7	28
		Safe	2	21
80%	Binary	Danger	24	14
		Safe	6	56
	Likelihood	Danger	17	5
		Warning	7	9
		Possible safe	4	16
		Safe	2	40
90%	Binary	Danger	27	7
		Safe	3	63
	Likelihood	Danger	24	3
		Warning	3	4
		Possible safe	2	7
		Safe	1	56

Different colors were used to draw the participants' attention to the diagnostic messages; the color sets of the binary and likelihood alarms are shown in Figure 3. During the task, the alert message and its color bar appeared at the bottom of the screen, and when the ADAS detected a forbidden item it marked the item with a red rectangle. A sample display of the simulated screening checkpoint is shown in Figure 4.

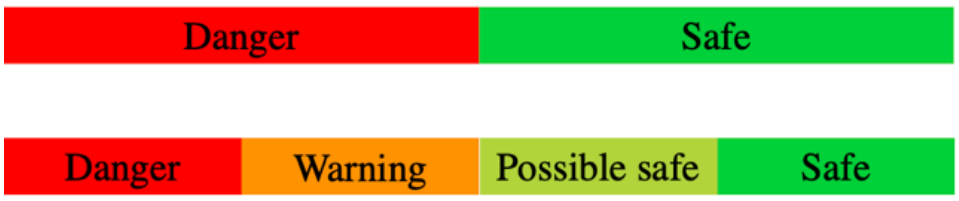

Figure 3: Color sets of the binary and likelihood alarms.

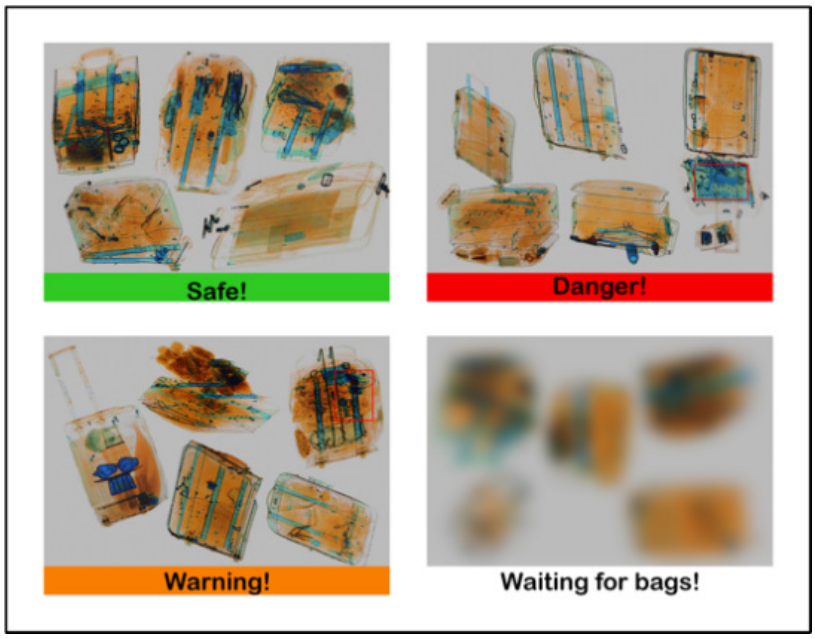

Figure 4: Sample display of the simulated screening scenario.

Experimental Design and Measures

The experiment adopted a 3 × 3 mixed factorial design. Alarm type (binary vs. likelihood vs. automated decision) was manipulated within subjects, and reliability level (70% vs. 80% vs. 90%) between subjects, yielding nine treatment conditions; each participant completed 100 trials under each alarm type at a single reliability level. The order of the three alarm types was counterbalanced across participants according to a Latin square.

Three categories of dependent variables were collected. First, detection sensitivity d' was computed for each participant in each condition from the observed hit and false-alarm rates; because these rates could equal 0 or 1 in some conditions, a correction of 0.5 was added to all raw hit and false-alarm counts before computing d' (Hautus, 1995). Second, response times were recorded separately for target-present (RTp) and target-absent (RTa) decisions. The two response types reflect different cognitive processes: target-present responses are a more direct measure of search performance, whereas target-absent responses are more strongly influenced by the decisional processes that establish a criterion for terminating an unsuccessful search (Drury & Chi, 1995). Trials with incorrect responses were excluded from the response time analysis. Third, after each alarm-type block, participants completed a human–computer trust questionnaire measuring five dimensions: perceived reliability, technical competence, understandability, faith, and personal attachment (Madsen & Gregor, 2000). The Cronbach's alpha values for the five dimensions were .809, .741, .760, .747, and .746, indicating acceptable internal consistency. Because the automated decision mode hides part of the image set and thereby changes the operator's decision space, it

is not directly comparable to the other two alarm types in SDT terms; the objective analyses therefore compared the binary and likelihood alarms only, whereas the subjective analyses covered all three alarm types.

Procedure

Upon arrival, participants were informed of the reliability level of the automated system to which they had been assigned. They then read instructions describing the search task and completed a practice block of 10 images, using material identical in kind to that of the formal experiment; participants who were unfamiliar with the procedure or the forbidden items could repeat the practice block. Before the experiment, the experimenter reviewed the forbidden items with the participants and explained their characteristic appearance in X-ray images. Before each block, on-screen instructions explained the properties of the upcoming alarm type, and participants were reminded: "You should be as accurate as possible when making your response. You are given 9 seconds for each picture."

The participants' task was to search each baggage X-ray image and decide whether it contained a forbidden item. Each trial began with a central fixation mark presented against a blank white background. Participants responded by pressing "J" on the keyboard if they believed a forbidden item was present and "F" if they believed it was absent. After finishing each block, they completed the trust questionnaire for the alarm type they had just used, and a short interview was conducted at the end of the session. The whole experiment lasted approximately 45 minutes. The experimental scenario is shown in Figure 5.

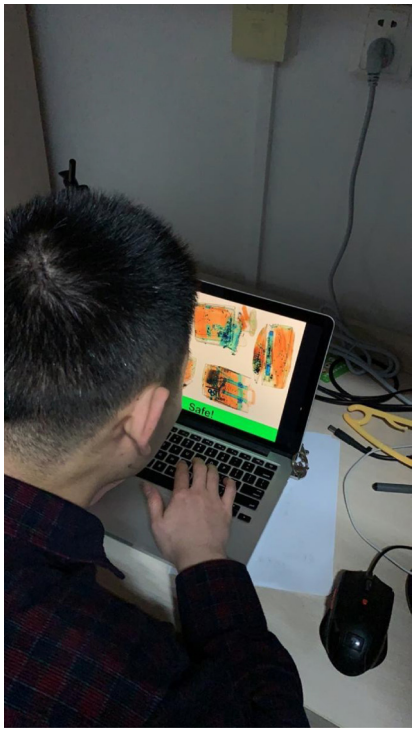

Figure 5: Experimental scenario.

Data Analysis

Sensitivity values, target-present response times, and target-absent response times were each submitted to a two-way mixed analysis of variance (ANOVA) with alarm type (binary vs. likelihood) as the within-subject factor and reliability level (70% vs. 80% vs. 90%) as the between-subject factor, using Type III sums of squares. Significant interactions were followed up with simple main effect analyses. Because the trust ratings were ordinal and the subgroup sizes were unbalanced, the effect of alarm type on each trust dimension was tested separately within each reliability level using the nonparametric Kruskal–Wallis test. The significance level was set at .05 for all analyses, and all statistical analyses were performed in R.

RESULTS

Sensitivity

Figure 6 shows the mean d' values for all combinations of reliability level and alarm type, together with the sensitivity of the ADAS alone (d' = 1.05, 1.68, and 2.56 at the 70%, 80%, and 90% levels, respectively). Under the binary alarm, participants' sensitivity increased monotonically with the reliability of the ADAS. Under the likelihood alarm, sensitivity first increased from the 70% to the 80% level and then decreased at the 90% level, where performance was worst and the average sensitivity (d' = 1.87) fell below that of the ADAS alone (d' = 2.56).

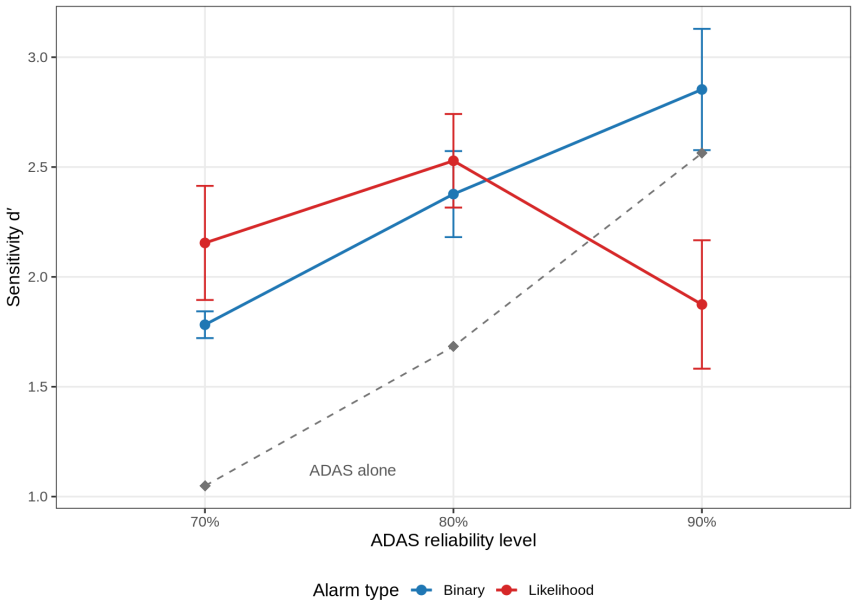

Figure 6: Mean d′ as a function of reliability level and alarm type.

Note. Error bars represent ±1 standard error. The dashed grey line indicates the sensitivity of the ADAS alone.

The two-way mixed ANOVA on d' (Table 3) revealed a significant alarm type × reliability interaction, $F(2, 15) = 7.05$, $p = .007$, $\eta p^2 = .48$, whereas neither the main effect of alarm type, $F(1, 15) = 0.93$, $p = .351$, nor that of reliability level, $F(2, 15) = 1.95$, $p = .177$, reached significance. Thus, the effect of alarm type on operators' sensitivity depended on the reliability of the system.

Table 3: Two-way mixed ANOVA on sensitivity d′ (type III sums of squares).

Effect	SS	df	F	p	ηp^2
Reliability	1.595	2, 15	1.95	.177	.206
Alarm type	0.207	1, 15	0.93	.351	.058
Reliability × Alarm type	3.151	2, 15	7.05	.007	.485

Simple main effect analyses were conducted to decompose the interaction. The effect of alarm type was significant only at the 90% reliability level, $F(1, 15) = 12.86$, $p = .003$, where the binary alarm outperformed the likelihood alarm; at the 70% and 80% levels the difference between alarm types was not significant. Conversely, the effect of reliability level was significant for the binary alarm, $F(2, 15) = 7.30$, $p = .006$, but not for the likelihood alarm, $F(2, 15) = 1.63$, $p = .23$. Relative to the sensitivity of the ADAS alone, the joint human–machine performance under the binary alarm represented gains of approximately 70%, 41%, and 11% at the 70%, 80%, and 90% reliability levels, respectively, indicating that the added value of the human operator diminished as the automation approached the operator's own ability.

Response Time

Figure 7 shows the mean response times, classified into target-absent and target-present decisions. No effect reached significance in the two-way mixed ANOVAs of RTa or RTp, although two marginal trends emerged: target-absent responses tended to be faster with the likelihood alarm than with the binary alarm, $F(1, 15) = 4.03$, $p = .063$, and the alarm type × reliability interaction on target-present responses approached significance, $F(2, 15) = 3.40$, $p = .061$, reflecting slower target-present responses under the likelihood alarm at the 90% level. The limited sample size may explain the absence of reliable response time effects, and these trends should therefore be interpreted with caution.

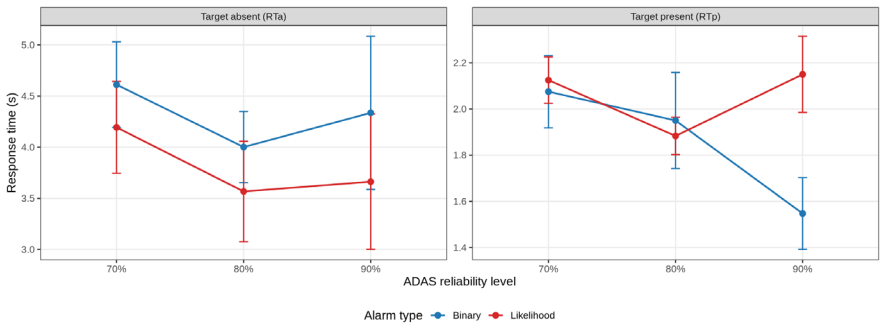

Figure 7: Mean response times for target-absent (left) and target-present (right) decisions.

Note. Error bars represent ±1 standard error.

Subjective Trust

The mean ratings of the five trust dimensions for each alarm type are shown in Figures 8–10, separately for the three reliability levels. Kruskal–Wallis tests were conducted on each dimension within each reliability level. Only the perceived reliability dimension at the 90% level differed significantly across the three alarm types, $\chi^2(2) = 7.37$, $p = .025$; no other comparison reached significance, so the following between-alarm differences should be interpreted as descriptive patterns.

At the 70% reliability level, participants' evaluations of competence, faith, and perceived reliability were highest for the likelihood alarm. The automated decision received the lowest faith rating: knowing that 30% of the forbidden items would be missed by the hidden "safe" triage made participants reluctant to rely on it, even though they found its instructions easy to follow. At the 80% level, the binary alarm received the highest ratings on all five dimensions, and the automated decision again received the lowest faith rating. At the 90% level, by contrast, the automated decision was rated highest on all dimensions, including faith, indicating that participants felt at ease following its instructions and were willing to delegate the triage of "safe" images to a highly reliable system.

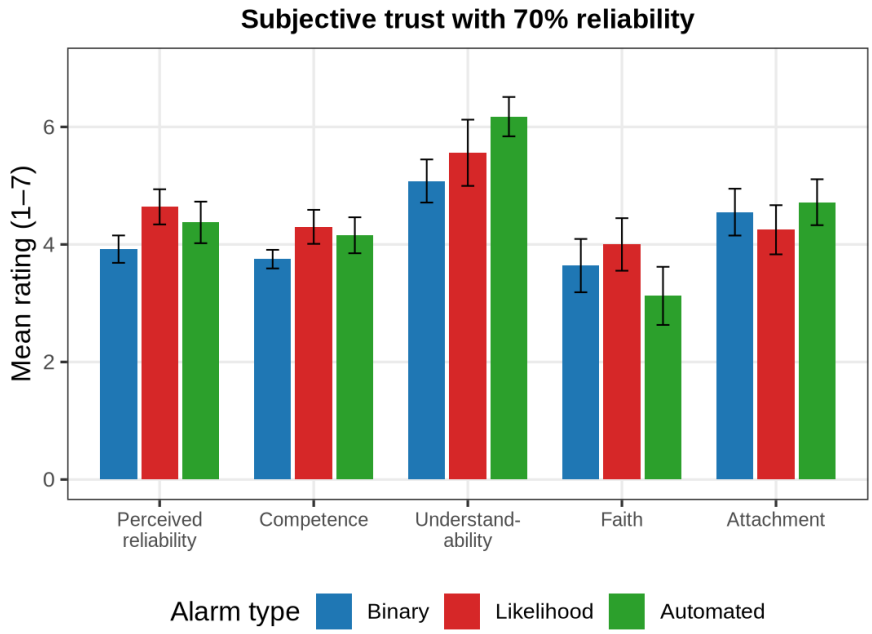

Figure 8: Mean trust ratings for the three alarm types at the 70% reliability level.

Note. Error bars represent ±1 standard error.

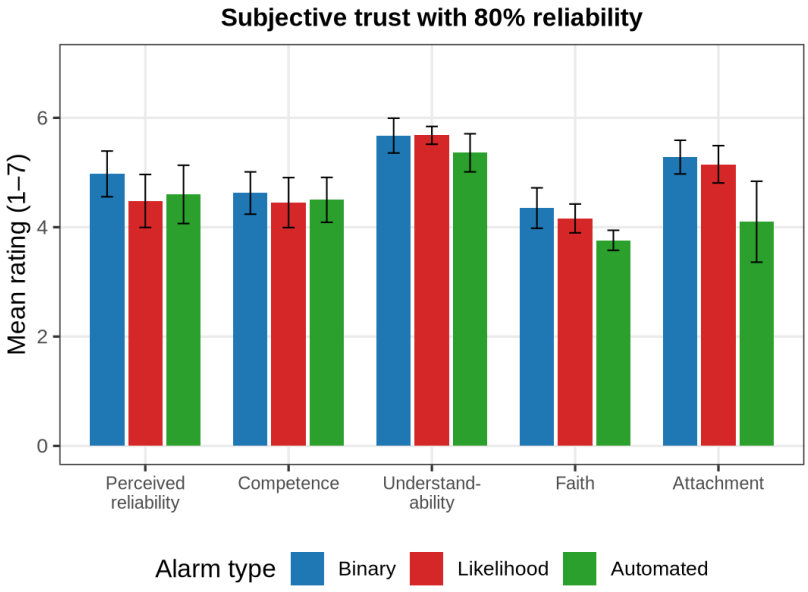

Figure 9: Mean trust ratings for the three alarm types at the 80% reliability level.

Note. Error bars represent ±1 standard error.

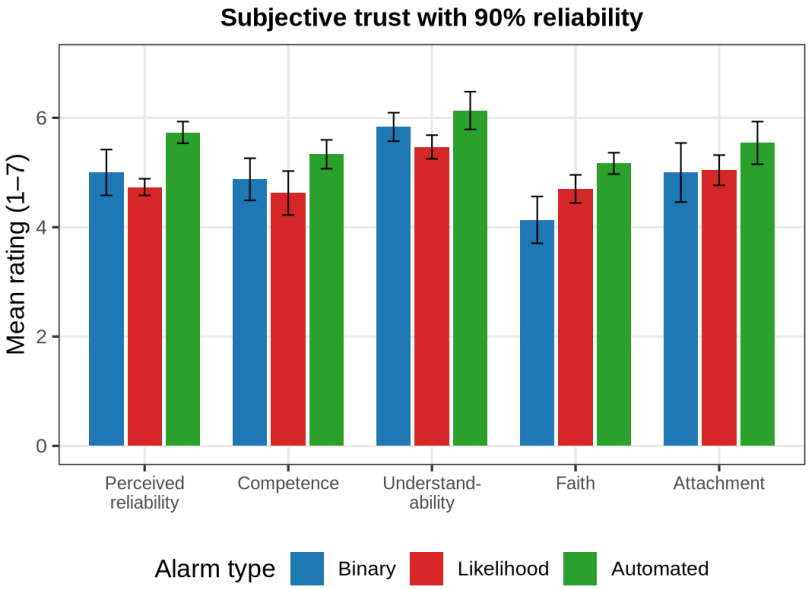

Figure 10: Mean trust ratings for the three alarm types at the 90% reliability level.

Note. Error bars represent ±1 standard error.

Interview Findings

The post-experiment interviews complemented the questionnaire results. Participants in the 70% reliability group reported little confidence in the system: frequent errors were expected, and they accepted that they had to rely

on their own judgement. Under this circumstance, they wanted the system to provide more information even at the cost of longer response times, felt compelled to check every "safe" image carefully, and regarded the automated decision mode as unacceptable; the likelihood alarm was considered the better choice. In contrast, participants in the 90% reliability group expressed willingness to follow the system's advice and reported growing confidence in the automated decision mode as they experienced its consistent performance.

DISCUSSION

Our first objective was to determine whether operators' performance differs across reliability levels and alarm types. The results indicate that an ADAS improves human performance in a complex and indirect way. When the reliability level was comparatively low, the joint human–machine performance exceeded the machine-alone benchmark by a wide margin, but the gain shrank as reliability increased (from approximately 70% at the 70% level to 11% at the 90% level under the binary alarm). This ceiling effect admits a straightforward explanation: most participants' unaided search ability corresponds to a sensitivity of roughly $d' = 2$, between the 80% and 90% reliability levels of the ADAS. When the system's reliability is well below the operator's own ability, the human contribution is large; when the system approaches the operator's ability, the joint performance can be only slightly better than the system alone.

The central finding is the significant interaction between reliability level and alarm type. Under the binary alarm, sensitivity increased with reliability, whereas under the likelihood alarm it first increased and then dropped sharply at the 90% level—significantly below the binary alarm and even below the ADAS alone. This pattern indicates that alarm format should be chosen according to the system's reliability. A likelihood alarm is better suited to a system of lower reliability, presumably because its graded, more ambiguous advice conveys additional diagnostic information and recruits more of the operator's attention; consistent with this interpretation, target-present responses tended to be slower under the likelihood alarm at the high reliability level, although target-absent responses tended to be terminated faster. The poor performance of the likelihood alarm under 90% reliability is not surprising in hindsight. When the automation is sufficiently accurate, operators could achieve satisfactory performance by simply following its advice, and the finer gradations of the likelihood alarm lose their informational value: distinguishing "Possible safe" from "Safe" contributes little when the system is almost always right, while the additional ambiguous messages increase stress, confusion, and the opportunity for error. Under high reliability, a binary alarm—or even the automated decision mode—is therefore the better choice.

Regarding subjective trust, participants' evaluations of the alarm types were consistent with the objective measurements, although most between-alarm differences were descriptive rather than statistically significant. When reliability was not high, participants tended to distrust the automated decision mode. Taking the 70% level as an example, the alarm configuration

implies that nine forbidden items per hundred images would be missed by the hidden triage; in the security domain, the cost of missing a dangerous item far outweighs the cost of checking thoroughly, and the interviews confirmed that participants would rather spend more time to improve their judgement accuracy. It is noteworthy, however, that at the 90% level participants expressed the greatest faith in, and preference for, the automated decision mode, whose performance exceeded that of the human operator in this condition. Repeated interaction with a system that consistently functions well appears to strengthen operators' faith in it, supporting the view that trust calibration tracks experienced reliability but depends on the form in which the automation communicates its assessments.

Several limitations should be acknowledged. First, the sample was small (six participants per reliability group), so the null results—particularly for response times—are likely underpowered, and the marginal trends require confirmation in larger samples. Second, the participants were graduate students rather than professional screeners, and the task was a laboratory simulation with a fixed 30% target prevalence, which is far higher than operational prevalence; generalization to field settings should therefore be made with caution. Third, the trust ratings were analysed with independent-samples nonparametric tests because the within-subject pairing could not be fully exploited, which is a conservative approach. Future work should validate the present findings with professional screeners, lower target prevalence, and adaptive alarm policies that switch alarm type as the system's estimated reliability changes.

CONCLUSION

To maximize the performance of the human–machine screening system, the alarm type of an automated diagnostic aid should be matched to its reliability level. When the system's accuracy is not high, a likelihood alarm is preferable: its graded advice provides more information, recruits the operator's attention, and supports informed human override. When the system is sufficiently reliable, a binary alarm or an automated decision mode is the better choice, as it minimizes unnecessary decisional load and enables efficient triage while sustaining operator trust. These results provide actionable guidance for the design of alarm policies in human–AI collaboration for safety-critical screening operations.

REFERENCES

Biggs, A. T., & Mitroff, S. R. (2015). Improving the efficacy of security screening tasks: A review of visual search challenges and ways to mitigate their adverse effects. *Applied Cognitive Psychology, 29*(1), 142–148. https://doi.org/10.1002/acp.3083

Cullen, R. H., Rogers, W. A., & Fisk, A. D. (2013). Human performance in a multiple-task environment: Effects of automation reliability on visual attention allocation. *Applied Ergonomics, 44*(6), 962–968. https://doi.org/10.1016/j.apergo.2013.02.010

Drury, C. G., & Chi, C.-F. (1995). A test of economic models of stopping policy in visual search. *IIE Transactions, 27*(3), 382–393. https://doi.org/10.1080/07408179508936754

Hancock, P. A., Mercado, J. E., Merlo, J., & Van Erp, J. B. F. (2013). Improving target detection in visual search through the augmenting multi-sensory cues. *Ergonomics, 56*(5), 729–738. https://doi.org/10.1080/00140139.2013.771219

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers, 27*(1), 46–51. https://doi.org/10.3758/BF03203619

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50*(3), 511–520. https://doi.org/10.1518/001872008X312198

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied, 17*(4), 320–331. https://doi.org/10.1037/a0024243

Schwaninger, A. (2005). Increasing efficiency in airport security screening. *WIT Transactions on the Built Environment, 82*, 405–416.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (4th ed.). Pearson.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics, 23*(10), 995–1011. https://doi.org/10.1080/00140138008924809

Yu, R., Yang, L., & Wu, X. (2017). Risk factors and visual fatigue of baggage X-ray security screeners: A structural equation modelling analysis. *Ergonomics, 60*(5), 680–691. https://doi.org/10.1080/00140139.2016.1192226