Ergonomics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/terg20

Effects of false alarms and miscues of decision support systems on human–machine system performance: a study with airport security screeners

David Huegli, Alain Chavaillaz, Juergen Sauer & Adrian Schwaninger

To cite this article: David Huegli, Alain Chavaillaz, Juergen Sauer & Adrian Schwaninger (11 Feb 2025): Effects of false alarms and miscues of decision support systems on human–machine system performance: a study with airport security screeners, Ergonomics, DOI: 10.1080/00140139.2025.2453546

To link to this article: https://doi.org/10.1080/00140139.2025.2453546

Published online: 11 Feb 2025.

ERGONOMICS
https://doi.org/10.1080/00140139.2025.2453546

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Effects of false alarms and miscues of decision support systems on human–machine system performance: a study with airport security screeners

David Huegli[a,b,c], Alain Chavaillaz[b], Juergen Sauer[b] and Adrian Schwaninger[a,c]

[a]School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Olten, Switzerland; [b]Department of Psychology, University of Fribourg, Fribourg, Switzerland; [c]Center for Adaptive Security Research and Applications (CASRA), Zurich, Switzerland

ABSTRACT

Decision support systems such as explosives detection systems for cabin baggage (EDSCB) at airport security checkpoints help screeners detect bombs by highlighting areas in X-ray images that might contain explosives. However, these systems are not perfect and can produce false alarms (i.e. alarm when no target is present) and miscues (i.e. a non-target is cued but the actual target is located elsewhere in the image). This study investigated the consequences of such automation errors in 112 professional airport security screeners who were supported by a simulated EDSCB with realistic X-ray images of cabin baggage. They had to detect bombs, guns, and knives under one of three experimental conditions: miscue prone, false alarm prone, or multiple failures (false alarms and miscues). Results showed that screeners missed more knives when the EDSCB provided miscues. We conclude that on-screen alarm resolution of EDSCB alarms in primary screening has the disadvantage that miscues can result in missing prohibited articles at airport security checkpoints. To avoid this problem, automated decision or clear instructions to screeners should be considered.

Practitioner Statement: Airport security screeners inspect X-ray images of cabin baggage through visual search and decision making with the help of explosives detection system for cabin baggage screening (EDSCB). The present experiment addresses whether EDSCB miscues affect operator performance and whether miscues are a problem when conducting EDSCB on-screen alarm resolution.

HIGHLIGHTS

- We tested 115 professional airport security screeners using realistic X-ray images of cabin baggage.
- Screeners were supported by an explosives detection system for cabin baggage (EDSCB) to help them detect bombs.
- Besides correct explosives alarms, the EDSCB made false alarms on target-absent images or miscues on images containing guns or knives that were localised elsewhere on the image.
- Screeners missed more knives when the EDSCB provided miscues than when the EDSCB provided no miscues.
- Instead of on-screen alarm resolution of EDSCB alarms in primary screening, automated decision or clear instructions to screeners should be considered.

ARTICLE HISTORY
Received 27 September 2023
Accepted 6 January 2025

KEYWORDS
Visual search; airport security; decision support systems; miscues; false alarms

Automated decision support systems are one type of automation that supports human operators by providing information about a particular state of the world (Mosier and Manzey 2020). Typical examples are alarm systems or detection systems (Rieger, Heilmann, and Manzey 2021) in security contexts (Goh, Wiegmann, and Madhavan 2005; Hättenschwiler et al. 2018; Huegli, Merks, and Schwaninger 2020), process control (Chavaillaz, Wastell, and Sauer 2016, Chavaillaz et al. 2019; Chavaillaz and Sauer 2017; Sauer, Chavaillaz, and Wastell 2016), or medical screening tasks (Alberdi et al. 2004; Drew, Cunningham, and Wolfe 2012; Xiao et al. 2021). One function of decision support systems in visual inspection tasks is to provide direct cues that operators must resolve. Direct cues attract the attention of human operators by indicating the exact location of a possible target (Chavaillaz et al. 2018; Darnell and Lamy 2022; Posner, Snyder, and Davidson 1980).

CONTACT David Huegli david_huegli@me.com University of Applied Sciences and Arts Northwestern Switzerland (FHNW), School of Applied Psychology (APS), Institute Humans in Complex Systems, Riggenbachstrasse 16, Olten 4600, Switzerland

Such direct cues have two goals: First, they should support the operator's search during a visual inspection task (Chavaillaz et al. 2018; Goh, Wiegmann, and Madhavan 2005; Wiegmann et al. 2006) by guiding attention towards the cued area (Wolfe 2021); and second, they should improve the operator's decision making regarding the detection of critical events or targets (Wickens et al. 2015). However, the reliability of most decision support systems is imperfect, and system designers must decide how to set their decision criterion. Usually, the priority is to set a more liberal decision criterion (Sorkin and Woods 1985) to detect more targets (automation hits). But a liberal decision criterion comes at the cost of producing more alarms where no target is present (automation false alarms) or producing automation miscues. Based on previous studies (Chavaillaz et al. 2020; Goh, Wiegmann, and Madhavan 2005), we define automation miscues as alarms in the wrong location when a target is located elsewhere in the image. Because of the decision support system's imperfect reliability, operators should assess the validity of its alarms and act appropriately. Such behaviour would improve human–machine system performance (Meyer and Kuchar 2021). Unfortunately, operators often do not react optimally to the decision support system's presence, resulting in insufficient human–machine system performance (Bartlett and McCarley 2017, 2019; Boskemper, Bartlett, and McCarley 2022).

Although direct cues can improve human–machine system performance (Chavaillaz et al. 2018), they can also have adverse effects when false alarms and miscues become frequent. It has been well-researched that operators will ignore automation alarms if they experience many false alarms, a so-called 'cry wolf' effect (Bliss, Gilson, and Deaton 1995; Dixon and Wickens 2006; Parasuraman and Wickens 2008; Zirk, Wiczorek, and Manzey 2020). Several studies have found this phenomenon and described it as a lack of operator compliance (Bliss, Gilson, and Deaton 1995; Huegli, Merks, and Schwaninger 2020, 2023; Parasuraman, Sheridan, and Wickens 2000; Zirk, Wiczorek, and Manzey 2020). We define compliance operationally as the proportion of target-present responses given by operators on trials with automation alarms—regardless of whether true or false (Dixon and Wickens 2006; Manzey, Gérard, and Wiczorek 2014). Compliance has been described as a behavioural expression of trust in automation (Hoff and Bashir 2015; Lee and See 2004). Although there is some evidence that compliance somehow relates to subjective trust in automation (Avril et al. 2022; Chancey et al. 2017; Lee and Moray 1992), subjective trust and objective compliance do not always correlate strongly (Chavaillaz, Wastell, and Sauer 2016, Chavaillaz et al. 2019; Rovira, McGarry, and Parasuraman 2007). Eye-tracking research has shown that the marked areas are the ones that are mainly inspected by operators (Drew, Cunningham, and Wolfe 2012), which implies strong attentional guidance (Wolfe 2021) by direct cues. However, if the marked area is a miscue, this can become a problem if operators get distracted by it and visual search for prohibited articles in other areas of the X-ray image is impaired. Alberdi et al. (2004) and Kunar et al. (2017) analysed the performance data of radiologists detecting breast cancer cells with the help of an automated decision support that provided direct cues. Radiologists benefitted from the alarms when they located cancer cells correctly, but less when miscues were given. Goh, Wiegmann, and Madhavan (2005) tested student participants in a simulated baggage X-ray screening task in which they had to detect knives in greyscale images with the help of a decision support system. Participants benefitted from direct cues, especially when automation reliability was high, but they missed most of the targets when miscues were present. We extend this research by using highly realistic colourised X-ray images with professional airport security officers supported by an explosive detection system (EDSCB) for cabin baggage screening. Moreover, whereas previous studies investigated whether miscues impair the detection of the same type of target located elsewhere in the X-ray image, we investigated spill-over effects from a cueing system for one type of target (EDSCB for explosive) on detection of different types of targets (guns and knives).

X-ray baggage screening at airports aims to prevent passengers from bringing prohibited articles (e.g. guns, knives, bombs) onto an aircraft (Harris 2002; Petrozziello and Jordanov 2019). Bombs are the most dangerous prohibited articles in passenger baggage and are technically called improvised explosive devices (IEDs). Screeners can detect IEDs when they are trained well (Halbherr et al. 2013; Koller et al. 2008, Koller, Drury, and Schwaninger 2009; Schuster et al. 2013; Schwaninger and Hofer 2004) and when they are supported with explosives detection systems for cabin baggage (EDSCB) screening (Hättenschwiler et al. 2018; Huegli, Merks, and Schwaninger 2020, 2023). EDSCB support screeners by providing direct cues in areas in the X-ray image that might contain explosive material. When the EDSCB is used as a decision support system in primary screening, the screeners conduct on-screen alarm resolution. That is, the screeners visually inspect the X-ray image and decide whether the EDSCB alarm is correct or incorrect (Hättenschwiler et al. 2018;

Schwaninger and Merks 2019). When the screeners decide that the bag could contain a prohibited article, the bag is sent to secondary screening where another screener inspects the bag by applying explosives trace detection or manual baggage opening (Sterchi and Schwaninger 2015). An alternative to on-screen alarm resolution in primary screening is automated decision: Bags on which EDSCB has alarmed are sent directly to secondary screening for further inspection (Hättenschwiler et al. 2018; Huegli, Merks, and Schwaninger 2020). State-of-the-art multi-view EDSCB technologies (EDSCB of Standard C2) achieve automation hit rates in the range of 75%–90% and automation false alarm rates of 6%–20% (Hättenschwiler et al. 2018). Because such false alarm rates can impair checkpoint efficiency (Sterchi and Schwaninger 2015), some countries still use EDSCB with on-screen alarm resolution. EDSCB false alarms can either occur on target-absent images not containing prohibited articles (genuine false alarms) or on images containing other prohibited articles such as guns or knives (miscues).

The goal of the present study was to investigate whether the failure proneness of a decision support system affects operator performance during a visual inspection task. More specifically, we wanted to examine spill-over effects from a cueing system for one type of target (EDSCB for explosives) on human detection of different types of targets (guns and knives). Besides the theoretical relevance of the study, we also wanted to answer the practically relevant question whether miscues are a problem that needs to be considered when conducting EDSCB on-screen alarm resolution in primary screening. The present study used highly realistic X-ray images in colour with professional airport security officers supported by EDSCB with a realistic automation reliability. The screeners had to detect prohibited articles (guns, knives, and IEDs). Screeners were tested in three experimental conditions: a *false alarm prone* condition where all EDSCB alarms were false alarms, a *miscue prone* condition where all EDSCB alarms were miscues, and a *multiple failures* condition where EDSCB false alarms and miscues appeared equally often. The dependent variables were performance measures (hit rate, false alarm rate, response time), behavioural trust (operator compliance), and subjective trust perception. Basic research on visual search has shown that salient cues capture attention, which can result in lower detection of targets, particularly in difficult visual search tasks (Gaspelin, Ruthruff, and Lien 2016; Luck et al. 2021; Ruthruff et al. 2020). Therefore, we expected lowest performance in the miscue prone condition, because screeners get distracted by the miscue and visual search for prohibited articles in other areas of the X-ray image gets impaired. Because knives are more difficult to detect than guns (Halbherr et al. 2013; Koller et al. 2008, Koller, Drury, and Schwaninger 2009), we expected to find the negative effect of miscues for detecting knives and less, if at all, for detecting guns. Because well trained screeners achieve a high detection of IEDs (Halbherr et al. 2013; Koller et al. 2008, Koller, Drury, and Schwaninger 2009), we expected that screeners would comply more with correct EDSCB alarms than with incorrect EDSCB alarms. That is, they should recognise such automation failures.

Method

Participants

Participants were 112 cabin baggage screeners from an international airport who were tested during their regular working hours. They signed up voluntarily to participate in the experiment during normal working hours as part of a typical working day. They received monetary compensation in the form of their regular paycheque based on their hourly salary. All screeners had been trained and certified according to the standards of the European Regulation (European Parliament 2015). After excluding seven screeners with unusually low detection rates (-2.5 *SD*) and six screeners with incomplete trials, the final sample consisted of 99 screeners (43 females, 52 males, and four others) with a mean age of 40.57 years ($SD = 9.53$ years, range = 25–63 years). The number of participants was determined through an a priori power analysis (alpha = 0.05 and beta = 0.85 with a small to medium effect size and a correlation of .5 among repeated measures). The study complied with the American Psychological Association Code of Ethics and was approved by the Internal Review Board of the Department of Psychology, University of Fribourg.

Procedure

Participants were tested in the training facilities of the airport in groups of a maximum of four screeners. Screeners were randomly assigned to one of the experimental conditions. The experiment took place in a quiet room and under supervision before the beginning of their working shift. Screeners sat approximately 60 mm in front of 21.5 BENQ 'GL2250' monitors connected to Dell 'Optiplex 3080' computers with Intel Core i5 processors. Before the experiment, participants were informed about the study procedures and goals

and gave written informed consent. Then they received oral and written instructions about the study, the user interface, and the EDSCB hit and false alarm rate. They were also told that an EDSCB alarm could indicate either a true alarm on an IED present, a false alarm without a prohibited article in the X-ray image, or a false alarm with either a gun or a knife elsewhere in the X-ray image—that is, a miscue. Participants were told about the overall false alarm rate of the EDSCB but not about the probability of the different types (false alarms without a prohibited article in the image vs. miscues). After the instructions, participants could start the experiment individually. They first conducted 66 practice trials with feedback given after each trial to allow participants to familiarise themselves with the task, and because exposure to the automation leads to a calibration of cognition towards the true reliability of the automation (Parasuraman and Wickens 2008) and knowing automation reliability is important for trust calibration (Mosier and Manzey 2020). Participants were supported by an EDSCB with the same reliability level and error proneness during the practice trials as during the main experiment. They then performed the main experiment in two test blocks containing 267 trials without feedback displayed in random order. European regulations (European Parliament 2015) mandate a break of at least 10 min after about 20 min of continuous X-ray image screening. Therefore, screeners took a break of 10 min after the first test block. After the X-ray screening task, participants completed the questionnaire to assess trust in EDSCB. The whole session took approximately 90 min.

Materials

The X-ray images and simulators were provided by the Centre for Adaptive Security Research and Applications (CASRA). Together with two X-ray image experts (former screeners) of CASRA, we selected 600 unique X-ray images of passenger bags displayed with multi-view imaging (66 practice and 534 test trials) containing no prohibited articles from a pool of 7,000 X-ray images recorded during regular airport security cabin baggage screening at several international airports. Also, we selected 72 unique images of IEDs (8 for practice, 64 for test trials), 36 of guns, and 36 of knives (4 of each for practice and 32 of each for test trials) that X-ray image experts of CASRA had recorded previously. The two X-ray image experts created target-present images by merging the prohibited articles into 144 of the 600 X-ray images using validated image merging algorithms (Mendes, Schwaninger, and Michel 2011; von Bastian, Schwaninger, and Michel 2008). This resulted in an overall target prevalence of 24%, with 12% of all images containing IEDs, 6% containing guns, and 6% containing knives. This complied with ratios of different prohibited article categories used in threat image projection (TIP) at this airport, a technology that projects pre-recorded X-ray images of prohibited articles into X-ray images of real passenger bags during baggage screening (Hofer and Schwaninger 2005; Meuter and Lacherez 2016; Skorupski and Uchroński 2016). Each target-present image contained one prohibited article. EDSCB alarms were set manually with the support of the X-ray image experts to achieve realistic trials. Figure 1 shows two multi-view X-ray images containing either a gun (a) or a knife (b).

Design

The experiment used a 3 × 3 mixed design with EDSCB failure proneness (miscue prone, false alarm prone, and multiple failures) as a between-subject factor and prohibited article category (guns, knives, IEDs) as a within-subject factor. Dependent variables were hit rate, false alarm rate, target present response time, target absent response time, operator trust, and compliance. Participants were randomly assigned to one of the three EDSCB failure proneness conditions. EDSCB supported screeners in the detection of explosive threats (IEDs). We chose guns as easy and knives as more difficult nonexplosive threats.

Task

Screeners worked on the task using a mouse to point and click on the screen. They were presented with one X-ray image per trial containing either a prohibited article (target-present) or not (target-absent). In all failure type conditions, red frames indicated areas in the X-ray images that might contain explosive material (EDSCB alarm, see Figure 1). No X-ray image contained more than one alarm. In every failure type condition, 48 out of 64 IEDs were correctly alarmed by the EDSCB (75%). Failure type conditions differed only in the nature of EDSCB alarms. In the false alarm prone condition, 32 EDSCB false alarms occurred on target-absent images. In the miscue prone group, 32 EDSCB false alarms occurred on images containing guns or knives (16 alarms each). In the multiple failures condition, 16 EDSCB false alarms occurred on images not containing prohibited articles, and the other EDSCB false alarms were on images containing guns or knives (eight each). Thus, the EDSCB had a hit rate of 75%, a false alarm rate of 7%, and overall 91% correct decisions, which is very realistic compared to multi-view EDSCB

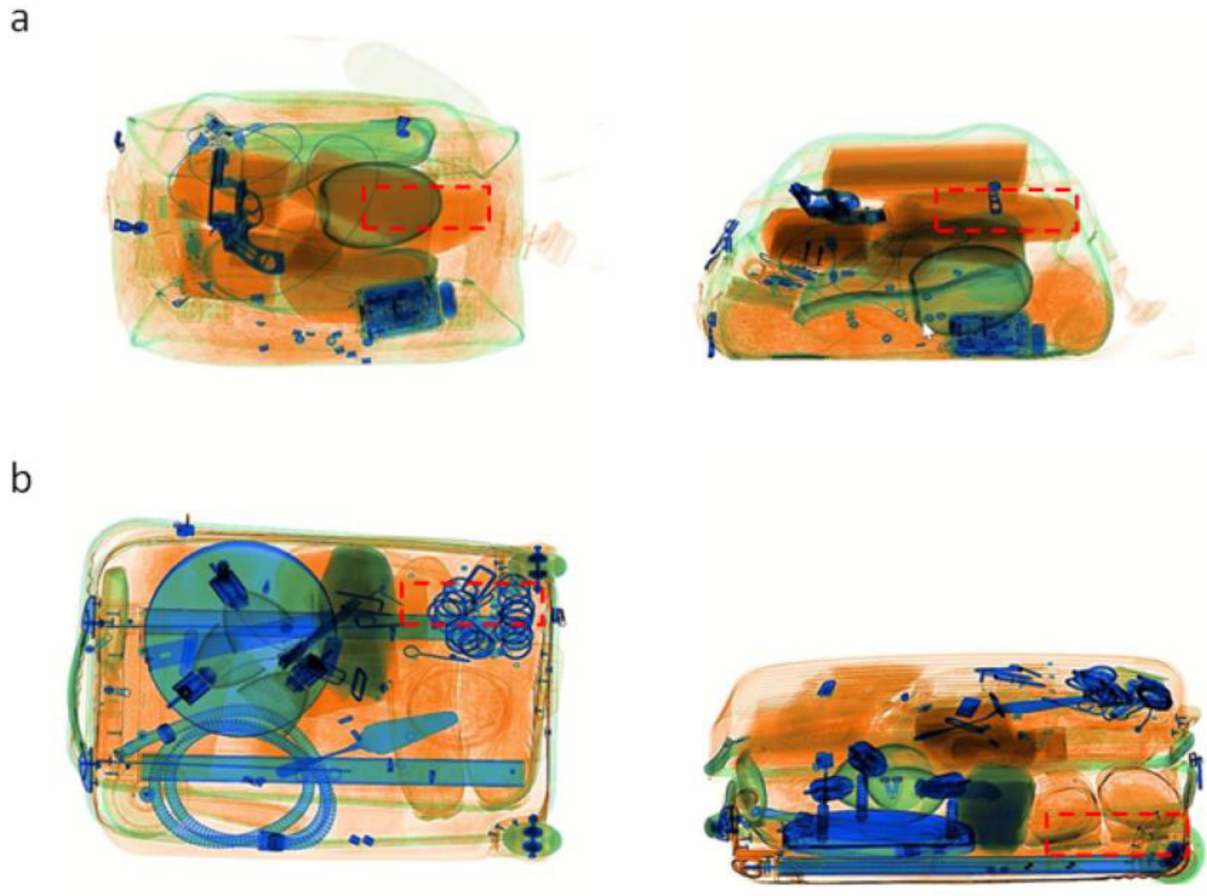

Figure 1. X-ray images from a multi-view EDSCB machine providing a miscue (red frame). The same image is shown from two viewpoints differing by about 90°. Image a contains a gun (easy); Image b, a knife (difficult).

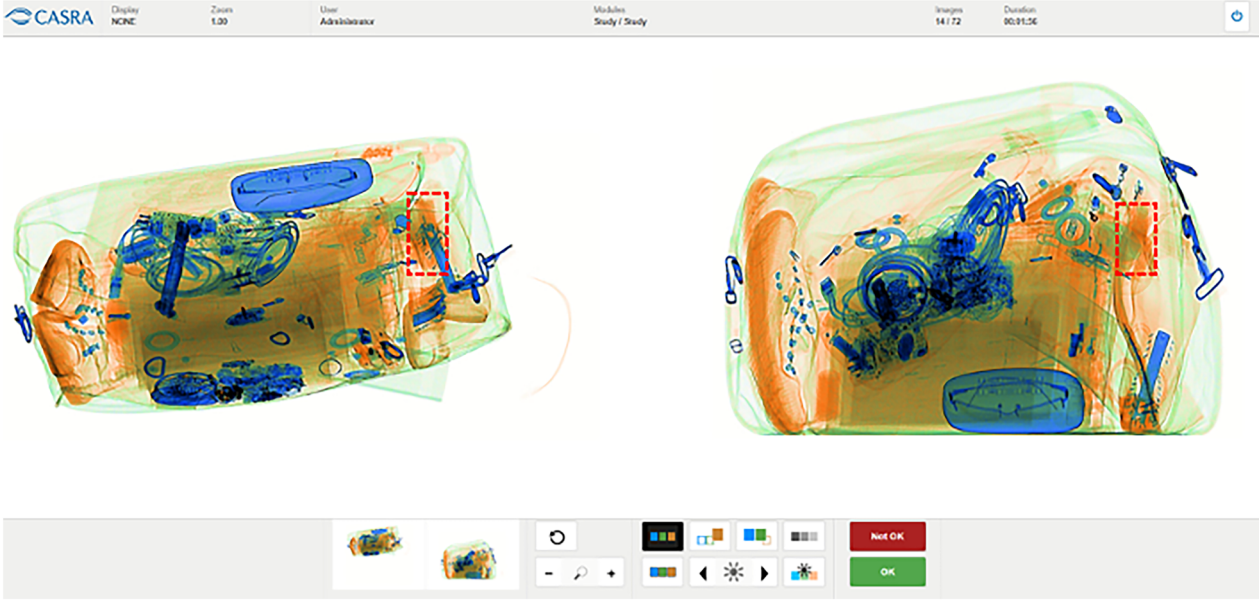

Figure 2. Interface of the simulator that was used for the experiment (X-Ray Tutor version 4, XRT4). The X-ray images of passenger bags contain a miscue (red dotted lines) and a knife as the actual target elsewhere in the image.

in operation at airports (Hättenschwiler et al. 2018; Huegli, Merks, and Schwaninger 2020). Figure 2 shows the interface of the simulator that was used for the experiment (X-Ray Tutor version 4, XRT4). Participants had 15 s to decide whether an X-ray image contained a target by clicking on an 'OK' button (target-absent) or a 'NOT OK' button (target-present). Before evaluating an X-ray image as 'NOT OK,' screeners had to mark the object they classified as a target by clicking on it. Many airports apply such a 15-s time limit in cabin baggage screening. After each trial, participants rated how confident they were about their decision on a 5-point scale ranging from 1 (*not confident*) to 5 (*very confident*). Then the subsequent trial started.

Dependent variables

We examined the following dependent variables as measures of detection performance: (a) the human–machine system *hit rate* (percentage of correctly marked threats on target-present trials), (b) the mean of the median[1] response times on target-present trials in ms, (c) the human–machine system *false alarm rate* (percentage of target-present responses on target-absent trials), (d) the mean of the median response times on target-absent trials. We measured (e) *trust perception* as the subjective trust in the automated system (Chavaillaz et al. 2019) assessed with the 12-item Checklist for Trust between People and Automation (CTPA; (Jian, Bisantz, and Drury 2000) on a 7-point scale ranging from 1 (*not at all*) to 7 (*agree totally*). An item example is 'The system is reliable.' As a behavioural trust measure, we chose (f) the *compliance rate* (% of target-present responses and the marking of alarm in trials in which the decision support system alarmed, whether correct or incorrect) consistent with Manzey, Gérard, and Wiczorek (2014) and Pharmer et al. (2021) and the definition by Dixon and Wickens (2006) and Meyer (2001, 2004). We also assessed operator confidence on each trial, but this was not analysed for the present study.

Statistical analyses

As mentioned in the introduction, we aimed to investigate spill-over effects from a cueing system for one type of target (EDSCB for explosive) on detection of different types of targets (guns and knives). Some trials with nonexplosive targets (guns and knives) contained miscues, trials with explosive targets (IEDs) did not contain miscues. Therefore, we conducted separate analyses of variance (ANOVAs) for target present trials with nonexplosive targets (guns and knives), for trials with explosive targets (IEDs), and for target absent trials. For nonexplosive targets, we ran 3 (failure type condition: false alarm prone, miscue prone, multiple failures) × 2 (nonexplosive prohibited articles category: guns, knives) mixed-design ANOVAs. For trials containing explosive targets and target-absent trials, we conducted one-way between-subjects ANOVAs. A Huyn–Feldt correction was applied to address any violation of the sphericity assumption. All ANOVAs and post hoc comparisons were calculated with R version 4.03 (R Core Team, 2022). Alpha was set at 0.05, and Holm–Bonferroni corrections were applied (Holm 1979) for post hoc *t* tests to correct for family-wise errors. We report effect sizes of ANOVAs using η_p^2 (partial eta-squared) with values of 0.01, 0.06, and 0.14 being interpreted as small, medium, and large effects respectively (Cohen 1988,p. 368). We computed basic bootstrapped 95% confidence intervals (1000 iterations) to assess the precision of estimated means.

Results

Detection of nonexplosive prohibited articles (guns and knives)

Figure 3 displays the human–machine system's hit rate by failure type condition and prohibited article category. The ANOVA confirmed that detection performance for nonexplosive prohibited articles depended on the failure proneness of EDSCB. There were significant main effects of failure type, $F(2, 95) = 4.82$, $p = .01$, partial $\eta^2 = 0.05$, and of prohibited article category, $F(1, 95) = 129.38$, $p < .001$, partial $\eta^2 = 037$. Additionally, the interaction between failure type and prohibited article category was significant, $F(2, 95) = 4.10$, $p = .02$, partial $\eta^2 = 0.04$. Further, post hoc comparisons revealed that knives were detected better in the false alarm prone than in the miscue prone ($p = .04$) or the multiple failures conditions ($p = .02$). There was no evidence of a difference in the detection of knives between the miscue prone group and the multiple failures group ($p > .99$). Figure 3 shows that differences for guns were less pronounced. The post hoc test for guns did not reveal a significant difference between the false alarm prone and the miscue prone condition ($p > .99$). Interestingly, screeners in the false alarm prone condition detected guns more accurately than in the multiple failures condition ($p = .02$) whereas the mean difference to screeners in the miscue prone condition was not significant ($p = .08$).

Figure 4 shows the target presence response times by failure type condition and prohibited target category. The ANOVA showed that the main effect of failure type was not significant, $F(2, 95=0.24$, $p = .79$, partial $\eta^2 = 0.00$, indicating that failure type had no effect on

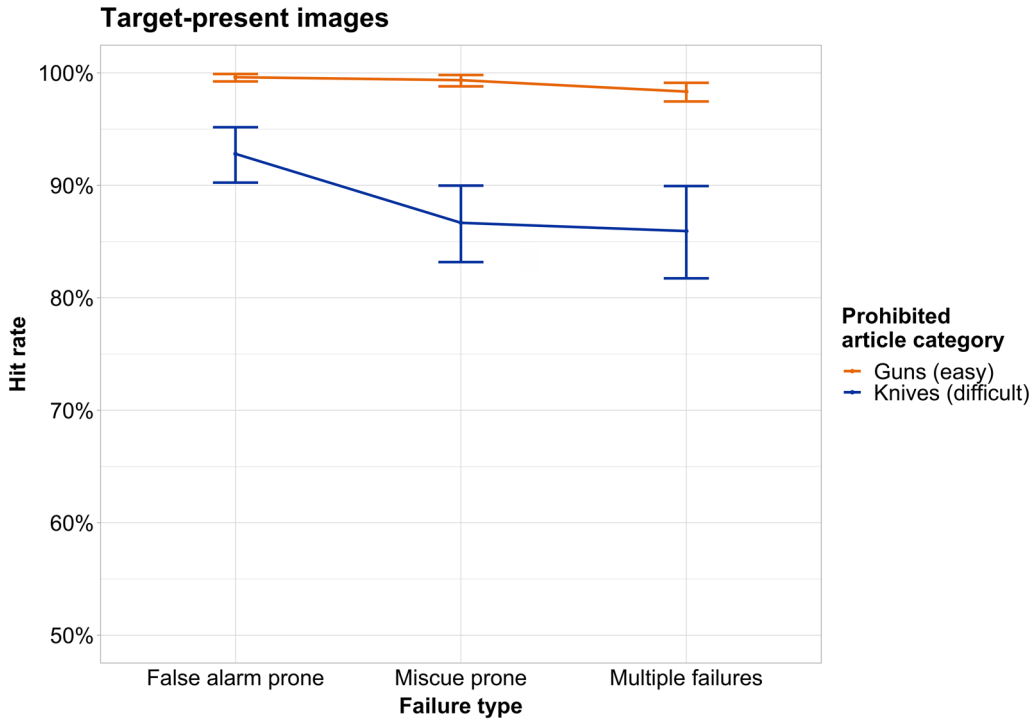

Figure 3. Mean human–machine system hit rate by failure type condition and prohibited article category. Error bars show bootstrapped 95% confidence intervals (1000 iterations).

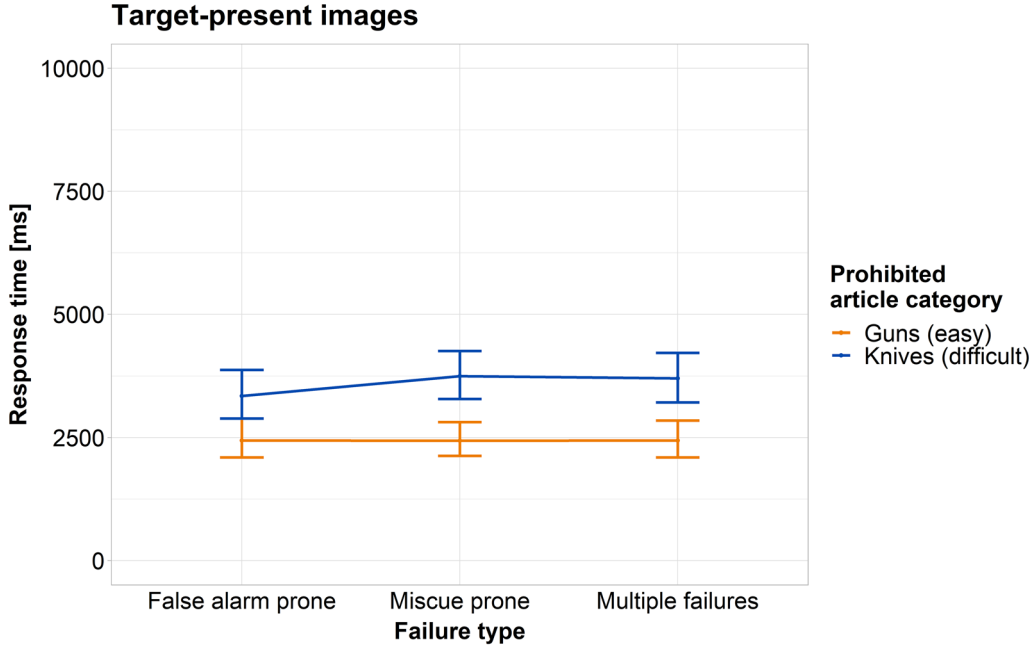

Figure 4. Target-present response times (mean of medians) by failure type condition and prohibited article category. Error bars show bootstrapped 95% confidence intervals (1000 iterations).

response times. Guns were detected faster than knives ($p < .001$, as there was a significant main effect of prohibited article category, $F(1, 95) = 185.00$, $p < .001$, partial $\eta^2 = 0.16$, a medium sized effect. The interaction between failure type and prohibited article category was not significant, $F(2, 95) = 2.26$, $p = .11$, partial $\eta^2 = 0.00$.

Detection of explosive prohibited articles (IEDs)

Explosives (IEDs) were detected very well in all test conditions, the hit rates were as follows: false alarm prone ($M = 0.93$, BCa 95% CI [0.91, 0.94]), miscue prone ($M = 0.92$, BCa 95% CI [0.90, 0.93]), and multiple failures ($M = 0.90$, BCa 95% CI [0.88, 0.92]). Moreover, a one-way

between-subjects ANOVA to compare the hit rate in different failure type conditions did not reveal any significant difference, $F(2, 95) = 0.51$, $p = .60$, partial $\eta^2 = 0.01$. A one-way between-subjects ANOVA to compare the target present response times for trials containing explosives (IEDs) in different failure type conditions did not reveal any significant differences, $F(2, 95) = 0.06$, $p = .94$, partial $\eta^2 = 0.00$. More specifically, results were as follows: false alarm prone ($M = 3334$ ms, BCa 95% CI [2819, 4021]), miscue prone ($M = 3442$ ms, BCa 95% CI [2960, 4052]), and multiple failures ($M = 3311$ ms, BCa 95% CI [2924, 3901]).

Target-absent trials

Figure 5 shows mean target absent response times for the failure type conditions (false alarm prone, miscue prone, multiple failures). A one-way between-subjects ANOVA to compare the failure type conditions, $F(2, 95) = 2.41$, $p = .09$, partial $\eta^2 = 0.05$, did not reveal any significant differences in the human–machine system false alarm rate.

A one-way between-subjects ANOVA to compare different failure type conditions (false alarm prone, miscue prone, multiple failures), did not reveal any significant differences in target-absent RTs, $F(2, 95) = 1.07$, $p = .34$, partial $\eta^2 = 0.02$ (see Figure 6).

Operator trust and compliance

For trust perception, a one-way ANOVA with the factor failure type condition, $F(2, 91) = 0.17$, $p = .84$, partial $\eta^2 = 0.00$, revealed no significant differences between conditions: false alarm prone condition ($M = 2.86$, BCa 95% CI [2.49, 3.33]), miscue prone condition ($M = 3.02$, BCa 95% CI [2.62, 3.55]), and multiple failures condition ($M = 3.04$, BCa 95% CI [2.66, 3.46]).[2] Figure 7 shows operator compliance with correct EDSCB and incorrect EDSCB alarms in each experimental group. The two-way ANOVA showed a significant main effect of failure type, $F(2, 95) = 5.70$, $p = .004$, partial $\eta^2 = 0.06$. There was also a significant main effect of alarm validity, $F(1, 95) = 7687.20$, $p < .001$, partial $\eta^2 = 0.97$, indicating a very large effect. Additionally, the interaction between failure type and prohibited article category was significant, $F(2, 95) = 6.87$, $p = .002$, partial $\eta^2 = 0.07$. Operators showed less compliance with incorrect EDSCB alarms in the miscue prone condition than in the false alarm prone condition and the multiple failures condition (both $ps < .01$). Operator compliance

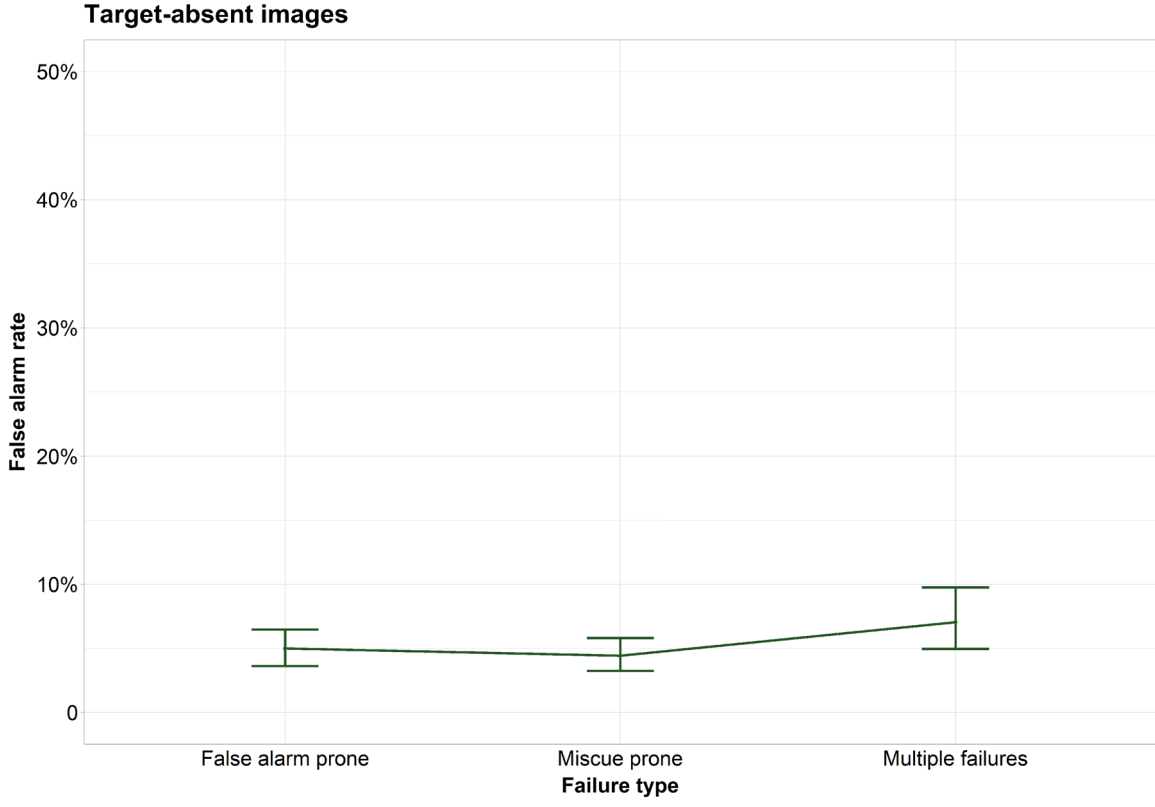

Figure 5. Human–machine system false alarm rate by failure type condition. Error bars show bootstrapped 95% confidence intervals (1000 iterations).

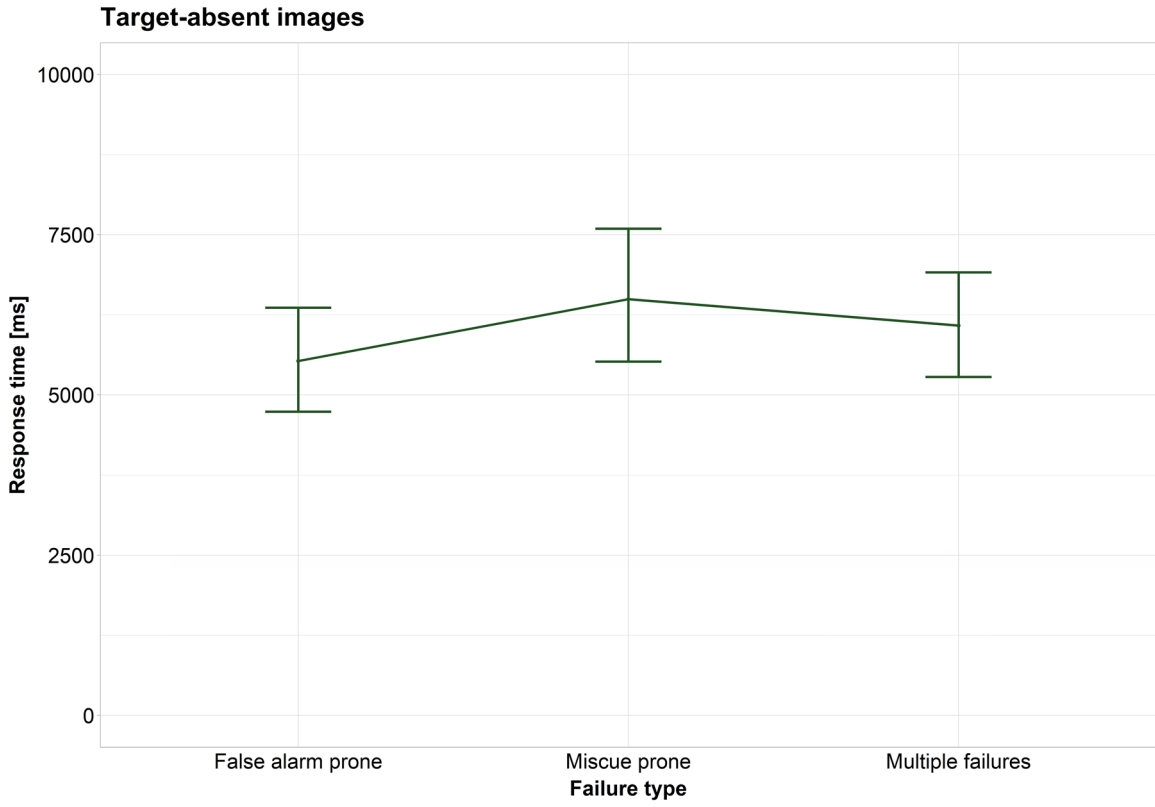

Figure 6. Target-absent RT (mean of medians) by failure type condition and prohibited article category. Error bars show bootstrapped 95% confidence intervals (1000 iterations).

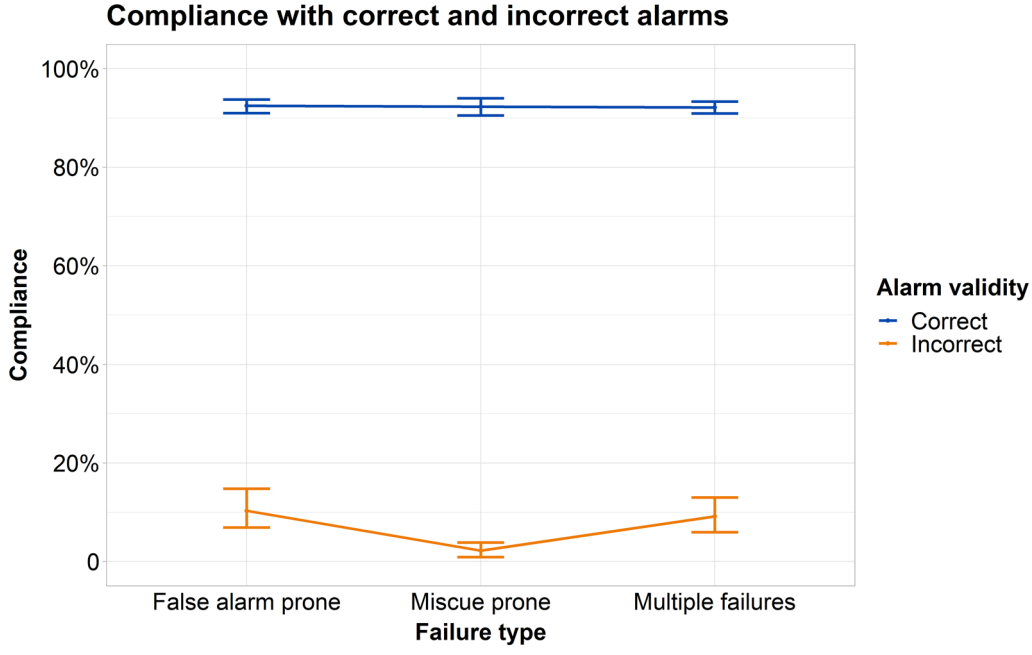

Figure 7. Mean operator compliance with correct and incorrect EDSCB alarms by failure type condition. Error bars show bootstrapped 95% confidence intervals (1000 iterations).

with correct EDSCB alarms did not differ between failure type conditions (all *ps* > .99).

Discussion

The goal of the present study was to investigate how the different failure types of a decision support system affect operator performance during a visual inspection task and whether operators recognise different automation failures. Specifically, the experiment aimed to examine the spill-over effects from a cueing system for one type of target (EDSCB for explosive) on human detection of different types of targets (guns and knives). Results showed that the failure proneness of EDSCB affected the detection performance of screeners, but only for items that are more challenging to identify: Screeners' detection performance for knives was worse when EDSCB was miscue prone or made multiple failures than when EDSCB was false alarm prone. Moreover, screeners could recognise many miscues and distinguish between correct and incorrect EDSCB alarms.

Detection of nonexplosive prohibited articles (guns and knives)

Detection of nonexplosive targets depended on the failure proneness of the EDSCB. Specifically, miscues reduced the correct detection of knives more than that of guns. This could be due to differences in target difficulty as previous studies found that guns are easier to detect than knives because they are larger and vary less in shape than knives (Halbherr et al. 2013; Hättenschwiler et al. 2018; Huegli et al. 2020; Koller, Drury, and Schwaninger 2009).

It is noteworthy that the EDSCB in the current study targeted only explosives and EDSCB miscues occurred on images containing guns or knives representing a realistic scenario. The miscues wrongly indicating the presence of an explosive impaired the detection of nonexplosive targets (guns and knives). These are spill-over effects from a cueing system for one type of target (EDSCB for explosive) on detection of different types of targets (guns and knives). One possible explanation is that miscues cause operators to stop searching for prohibited articles. This would be consistent with previous research on subsequent search errors providing evidence that search errors become more frequent when the actual target is dissimilar to the one that was searched for first (for a review, see Adamo et al. 2021), and when the distractors (in this case the miscue) appear in a salient manner (Lawrence and Pratt 2022; Moher 2020). Further, our results did not reveal an effect of the failure proneness of the system on response times, but guns were detected better and faster than knives. These results speak against a trade-off between speed and accuracy, that is, better detection at the expense of a slower response time (Heitz, 2014). Even greater effects of miscuing than in the present study have been found in previous studies of experts performing medical screening tasks (Alberdi et al. 2004; Kunar et al. 2017) and students performing a baggage X-ray screening task (Goh, Wiegmann, and Madhavan 2005). However, both studies used black-and-white images. The colouring of the X-ray images in our study might have reduced the negative effect of miscues, because colour facilitates the detection of prohibited articles (von Bastian, Schwaninger, and Michel 2008) due to pop-out effects (Wolfe 2021; Wolfe et al. 2011). Interestingly, screeners performed worse in the realistic multiple failures condition than in the miscue prone condition. One reason for this finding could be that in the multiple failures condition both miscues and regular false alarms occurred, but in the miscue prone condition, incorrect alarms were always miscues. Screeners perhaps adapted and recognised that when an incorrect alarm appeared, they must search the image for another prohibited article. In the multiple failures condition, both false alarms and miscues occurred, and screeners made more errors. This is consistent with previous findings showing an adverse effect of failure inconsistency on human–machine system performance (Bahner, Hüper, and Manzey 2008; Sauer, Chavaillaz, and Wastell 2016).

Detection of explosives

The EDSCB had the same automation hit rate in all three test conditions when alarming on explosives. As expected, the failure proneness of the EDSCB did not affect the human–machine system hit rate and response times when inspecting images with IEDs. This is good news, because it suggests that the failure proneness of the system does not affect the usage of the EDSCB when its cues are correct. Moreover, the human–machine system hit rate was at least 15% or more, higher than the EDSCB false alarm rate of 7%. That is, screeners assessed and used the information provided by the EDSCB. When EDSCB missed an IED, screeners detected most of them. IEDs are composed of a power source, a triggering device, a detonator, and an explosive charge usually connected by wires (Turner 1994; Wells and Bradley 2012) that trained screeners can detect well (Halbherr et al. 2013; Koller

et al. 2008, Koller, Drury, and Schwaninger 2009; Schuster et al. 2013; Schwaninger and Hofer 2004).

Target-absent trials

False alarm rates by airport security screeners are critical for checkpoint efficiency because further examination of bags at the checkpoint involves more time-consuming explosives trace detection and the manual opening of baggage (Sterchi and Schwaninger 2015). Further, target-absent response times are more relevant for efficiency in baggage screening, because most bags do not contain prohibited articles (in both our experiment and at airports). Our results show that the failure proneness of the EDSCB did not affect the human–machine false alarm rate and response times when inspecting target-absent images. This result was not expected, because the three test conditions did differ in the number of false alarms on target-absent images. Moreover, the human–machine system false alarm rate did not exceed the EDSCB false alarm rate. From a practical point of view, however, this is good news, because screeners could keep the human–machine system false alarm rate within an acceptable range (Sterchi and Schwaninger 2015) regardless of the amount of EDSCB false alarms on target-absent images.

Operator compliance and subjective trust perception

In our study, trust perception did not differ between the three different failure proneness conditions, suggesting that the nature of false alarms did not affect trust in automation. However, besides trust, other factors such as recognising automation failures affect the use of a decision support system (Spain 2009). Trust scores of our screeners were relatively low (mean of about 3 on a 7-point Likert scale) and lower than in a previous study using the same questionnaire to evaluate trust in a decision support system that made only false alarms and miscues but no misses (Chavaillaz et al. 2019). We argue that screeners built their trust based on their experience with EDSCB detecting (or somewhat missing) IEDs, and that miscues did not affect screeners' trust perception. Compliance has been described as behavioural expressions of trust in automation (Hoff and Bashir 2015; Lee and See 2004). However, the relationship between compliance and subjective trust in automation is vague (Chavaillaz et al. 2019).

When EDSCB correctly alarmed IEDs, screeners showed more compliance with EDSCB than when EDSCB provided incorrect alarms. When EDSCB triggered miscues, operator compliance was lower than when EDSCB triggered normal false alarms. These results are important because they show that screeners can differentiate between correct and incorrect EDSCB alarms. Moreover, compliance with incorrect alarms was lower in the miscue prone condition (i.e. miscues only) than in the multiple failures condition in which both normal false alarms and miscues occurred. When EDSCB yielded a miscue in the miscue prone condition, screeners could adjust and inspect the image more closely than in the more realistic multiple failures condition. Both results provide evidence that screeners actively process and adapt to automation alarms, which speaks against complacency (Alberdi et al. 2004, 2008; Meyer, Wiczorek, and Günzler 2014; Onnasch et al. 2014; Rice and McCarley 2011). Screeners cannot become complacent when only 48 out of 80 EDSCB alarms are correct.

Limitations and future research

This study has some limitations that could be addressed in future research. First, it was conducted with screeners from only one airport who were trained and certified according to the standards of the European Regulation (European Parliament 2015). It would be interesting to examine whether these results are also found in other airports, and whether specific training on EDSCB on-screen alarm resolution could reduce the problem of EDSCB miscues. Furthermore, future studies should compare how experienced screeners react to different kinds of errors compared to inexperienced ones. Second, many effects we found on performance were only small. However, the negative effects of miscues are still relevant. For example, about 8 percent difference across conditions for knives (see Figure 3) is substantial in terms of security risks. Third, in X-ray screening at airports, the extremely low target prevalence of prohibited articles is increased artificially to about 2%–4% using threat image projection and covert tests (Hofer and Schwaninger 2005; Meuter and Lacherez 2016; Schwaninger 2009; Skorupski and Uchroński 2016; Wetter, Hardmeier, and Hofer 2008). Target prevalence in our study was higher than in real-world scenarios to achieve enough trials to run statistical tests. A lower target prevalence could have negatively affected the detection of prohibited articles (Biggs and Mitroff 2015; Buser, Sterchi, and Schwaninger 2020; Wolfe et al. 2007; Wolfe and Van Wert 2010) and increased the potentially dangerous misguidance effects of miscues (Kunar et al. 2017; Schwarz and Miller 2016). Therefore, we expect that the negative effects of miscues would increase even more in real

working conditions. Fourth, the quality of EDSCB is still improving. For example, EDSCB based on computer tomography, which is becoming more common at airports, could soon achieve EDSCB false alarm rates below 5% (Huegli, Merks, and Schwaninger 2020). This technical progress needs to be considered when discussing the practical implications of our study. Fifth, eye-tracking could be useful to investigate whether miscues impede further scanning and/or detection (for some targets). Finally, it could be interesting to investigate the impact of other distractors, such as interruptions, compared to the effects of miscues.

Practical implications

We have shown that EDSCB miscues can impair the detection of nonexplosive targets, especially knives when they are located elsewhere in the X-ray image. Furthermore, this effect was prominent when miscues and multiple failures occurred during the test. Screeners show more compliance with correct than with incorrect EDSCB alarms, and they recognise most miscues. Nevertheless, we consider miscues to be a problem that needs to be addressed when conducting EDSCB on-screen alarm resolution. In our study, the EDSCB had a hit rate of 75% and a false alarm rate of 7%. Using EDSCB based on computer tomography that achieve even higher hit rates (80% or more) and very low false alarm rates (5% or less) would improve the detection of explosive material in bags. Additionally, automated prohibited item detection systems that alarm on nonexplosive targets and also achieve very high hit rates and low false alarm rates (Liang et al. 2019) would increase human–machine system performance even more. Resolving all these alarms in secondary screening with explosives trace detection and manual bag opening can be considered reasonable and should not require additional staff resources (Sterchi and Schwaninger 2015). We recommend giving screeners clear instructions to send every alarmed bag to a secondary search, as suggested by Hättenschwiler et al. (2018) and Huegli, Merks, and Schwaninger (2020). However, screeners should still visually inspect X-ray images for guns, knives, and IEDs. The latter is crucial, because EDSCB still misses some IEDs (Howell 2017), and screeners can detect them well by visual inspection, as demonstrated in our study.

Conclusion

The present study investigated whether the failure proneness of a decision support system (in our case, EDSCB) affects human–machine system performance during a visual inspection task and whether operators recognise automation failures. Our results show that EDSCB miscues impaired the detection of knives located elsewhere in the image. These are spill-over effects from a cueing system for one type of target (EDSCB for explosive) on detection of a different type of target (knives). Screeners showed more compliance with correct than with incorrect EDSCB alarms indicating that screeners could differentiate between correct and incorrect automation alarms. Although screeners recognised many miscues, we recommend that when EDSCB indicates that the bag might contain explosive material, the baggage should always be further examined in a secondary screening.

Notes

1. We used the median RT as our estimate because it provides a less biased estimate of the underlying RT (Brenner and Smeets 2019; Rousselet and Wilcox 2020) and because it has been used in several other studies when distributions of RT were skewed (Gordon et al. 2020; Horowitz et al. 2003; Horowitz and Wolfe 2003; Thornton and Zdravković 2020; Wolfe 2022).
2. Note that only 95 participants completed the trust perception questionnaire.

Acknowledgments

The authors thank Tobias Rieger for his insightful thoughts on the manuscript and Jonathan Harrow for proofreading the manuscript regarding English grammar and style.

Authors contributions

All authors have made substantial contributions to the conception and design of the work. DH performed the acquisition and analysis of the data. All authors helped with the interpretation of data. DH drafted the manuscript, and all authors substantively revised it. All authors have approved the submitted version. All authors have agreed both to be personally accountable for their own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated and resolved, and that the resolution will be documented in the literature.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Ethics approval

The study complied with the American Psychological Association Code of Ethics, and it was also approved by the Internal Review Board of the Department of Psychology, University of Fribourg. Participants were informed about the study procedures and goals and gave written informed consent.

Funding

This study was funded by the Swiss National Science Foundation (project number 100019_149184).

ORCID

David Huegli http://orcid.org/0000-0002-7176-7765
Alain Chavaillaz http://orcid.org/0000-0001-7191-1360
Juergen Sauer http://orcid.org/0000-0003-2105-1694
Adrian Schwaninger http://orcid.org/0000-0001-7753-106X

Data availability statement

The datasets generated and analysed during the current study are not publicly available because they contain security-sensitive information. On reasonable request, data in an aggregated form are available from the corresponding author.

References

Adamo, S. H., B. J. Gereke, S. Shomstein, and J. Schmidt. 2021. "From "Satisfaction of Search" to "Subsequent Search Misses": A Review of Multiple-Target Search Errors across Radiology and Cognitive Science." *Cognitive Research* 6: 1–19. Springer Science and Business Media Deutschland GmbH. doi:10.1186/s41235-021-00318-w.

Alberdi, E., A. A. Povyakalo, L. Strigini, P. Ayton, and R. Given-Wilson. 2008. "CAD in Mammography: Lesion-Level versus Case-Level Analysis of the Effects of Prompts on Human Decisions." *International Journal of Computer Assisted Radiology and Surgery* 3 (1-2): 115–122. doi:10.1007/s11548-008-0213-x.

Alberdi, E., A. Povykalo, L. Strigini, and P. Ayton. 2004. "Effects of Incorrect Computer-Aided Detection (CAD) Output on Human Decision-Making in Mammography." *Academic Radiology* 11 (8): 909–918. doi:10.1016/j.acra.2004.05.012.

Avril, E., J. Cegarra, L. Wioland, and J. Navarro. 2022. "Automation Type and Reliability Impact on Visual Automation Monitoring and Human Performance." *International Journal of Human–Computer Interaction* 38 (1): 64–77. doi:10.1080/10447318.2021.1925435.

Bahner, J. E., A. D. Hüper, and D. Manzey. 2008. "Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience." *International Journal of Human-Computer Studies* 66 (9): 688–699. doi:10.1016/j.ijhcs.2008.06.001.

Bartlett, M. L., and J. S. McCarley. 2017. "Benchmarking Aided Decision Making in a Signal Detection Task." *Human Factors* 59 (6): 881–900. doi:10.1177/0018720817700258.

Bartlett, M. L., and J. S. McCarley. 2019. "No Effect of Cue Format on Automation Dependence in an Aided Signal Detection Task." *Human Factors* 61 (2): 169–190. doi:10.1177/0018720818802961.

Biggs, A. T., and S. R. Mitroff. 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. doi:10.1002/acp.3083.

Bliss, J. P., R. D. Gilson, and J. E. Deaton. 1995. "Human Probability Matching Behaviour in Response to Alarms of Varying Reliability." *Ergonomics* 38 (11): 2300–2312. doi:10.1080/00140139508925269.

Boskemper, M. M., M. L. Bartlett, and J. S. McCarley. 2022. "Measuring the Efficiency of Automation-Aided Performance in a Simulated Baggage Screening Task." *Human Factors* 64 (6): 945–961. doi:10.1177/0018720820983632.

Brenner, E., and J. B. J. Smeets. 2019. "How Can You Best Measure Reaction Times?" *Journal of Motor Behavior* 51 (5): 486–495. doi:10.1080/00222895.2018.1518311.

Buser, D., Y. Sterchi, and A. Schwaninger. 2020. "Why Stop after 20 Minutes? Breaks and Target Prevalence in a 60-Minute X-Ray Baggage Screening Task." *International Journal of Industrial Ergonomics* 76: 102897. doi:10.1016/j.ergon.2019.102897.

Chancey, E. T., J. P. Bliss, Y. Yamani, and H. A. H. Handley. 2017. "Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence." *Human Factors* 59 (3): 333–345. doi:10.1177/0018720816682648.

Chavaillaz, A., and J. Sauer. 2017. "Operator Adaptation to Changes in System Reliability under Adaptable Automation." *Ergonomics* 60 (9): 1261–1272. doi:10.1080/00140139.2016.1261187.

Chavaillaz, A., A. Schwaninger, S. Michel, and J. Sauer. 2018. "Automation in Visual Inspection Tasks: X-Ray Luggage Screening Supported by a System of Direct, Indirect or Adaptable Cueing with Low and High System Reliability." *Ergonomics* 61 (10): 1395–1408. doi:10.1080/00140139.2018.1481231.

Chavaillaz, A., A. Schwaninger, S. Michel, and J. Sauer. 2019. "Expertise, Automation and Trust in X-Ray Screening of Cabin Baggage." *Frontiers in Psychology* 10: 256. doi:10.3389/fpsyg.2019.00256.

Chavaillaz, A., A. Schwaninger, S. Michel, and J. Sauer. 2020. "Some Cues Are More Equal than Others: Cue Plausibility for False Alarms in Baggage Screening." *Applied Ergonomics* 82: 102916. doi:10.1016/j.apergo.2019.102916.

Chavaillaz, A., D. Wastell, and J. Sauer. 2016. "System Reliability, Performance and Trust in Adaptable Automation." *Applied Ergonomics* 52: 333–342. doi:10.1016/j.apergo.2015.07.012.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York, NY: Lawrence Erlbaum Associates.

Darnell, M., and D. Lamy. 2022. "Spatial Cueing Effects Do Not Always Index Attentional Capture: evidence for a Priority Accumulation Framework." *Psychological Research* 86 (5): 1547–1564. doi:10.1007/s00426-021-01597-0.

Dixon, S. R., and C. D. Wickens. 2006. "Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload." *Human Factors* 48 (3): 474–486. doi:10.1518/001872006778606822.

Drew, T., C. Cunningham, and J. M. Wolfe. 2012. "When and Why Might a Computer-Aided Detection (CAD) System Interfere with Visual Search? An Eye-Tracking Study." *Academic Radiology* 19 (10): 1260–1267. doi:10.1016/j.acra.2012.05.013.

European Parliament. 2015. "Commission Implementing Regulation (EU)2015/1998 of 5 November 2015 Laying down Detailed Measures for the Implementation of the

Common Basic Standards on Aviation Security." *Official Journal of the European Union* 1 (58): 1–146.

Gaspelin, N., E. Ruthruff, and M.-C. Lien. 2016. "The Problem of Latent Attentional Capture: Easy Visual Search Conceals Capture by Task-Irrelevant Abrupt Onsets." *Journal of Experimental Psychology* 42 (8): 1104–1120. doi:10.1037/xhp0000214.

Goh, J., D. A. Wiegmann, and P. Madhavan. 2005. "Effects of Automation Failure in a Luggage Screening Task: A Comparison between Direct and Indirect Cueing." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49 (3): 492–496. doi:10.1177/154193120504900359.

Gordon, A., R. Geddert, J. Hogeveen, M. K. Krug, S. Obhi, and M. Solomon. 2020. "Not So Automatic Imitation: Expectation of Incongruence Reduces Interference in Both Autism Spectrum Disorder and Typical Development." *Journal of Autism and Developmental Disorders* 50 (4): 1310–1323. doi:10.1007/s10803-019-04355-9.

Halbherr, T., A. Schwaninger, G. R. Budgell, and A. Wales. 2013. "Airport Security Screener Competency: A Cross-Sectional and Longitudinal Analysis." *The International Journal of Aviation Psychology* 23 (2): 113–129. doi:10.1080/10508414.2011.582455.

Harris, D. H. 2002. "How to Really Improve Airport Security." *Ergonomics in Design: The Quarterly of Human Factors Applications* 10 (1): 17–22. doi:10.1177/106480460201000104.

Hättenschwiler, N., Y. Sterchi, M. Mendes, and A. Schwaninger. 2018. "Automation in Airport Security X-Ray Screening of Cabin Baggage: Examining Benefits and Possible Implementations of Automated Explosives Detection." *Applied Ergonomics* 72: 58–68. doi:10.1016/j.apergo.2018.05.003.

Heitz, R. P. 2014. "The Speed-Accuracy Tradeoff: History, Physiology, Methodology, and Behavior. Frontiers in Neuroscience." *Frontiers in Neuroscience* 8: 150–119. 10.3389/fnins.2014.00150.

Hofer, F., and A. Schwaninger. 2005. "Using Threat Image Projection Data for Assessing Individual Screener Performance." *WIT Transactions on the Built Environment* 82: 417–426. doi:10.2495/SAFE050411.

Hoff, K. A., and M. Bashir. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57 (3): 407–434. doi:10.1177/0018720814547570.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70. doi:10.1142/S0218195905001683.

Horowitz, T. S., B. E. Cade, J. M. Wolfe, and C. A. Czeisler. 2003. "Searching Night and Day: A Dissociation of Effects of Circadian Phase and Time Awake on Visual Selective Attention and Vigilance." *Psychological Science* 14 (6): 549–557. doi:10.1046/j.0956-7976.2003.psci_1464.x.

Horowitz, T. S., and J. M. Wolfe. 2003. "Memory for Rejected Distractors in Visual Search? In." *Visual Cognition* 10 (3): 257–298. doi:10.1080/13506280143000005.

Howell, J. 2017. "The Modern IED: Design and Trends." *Aviation Security International Magazine*, 1–15.

Huegli, D., S. Merks, and A. Schwaninger. 2020. "Automation Reliability, Human–Machine System Performance, and Operator Compliance: A Study with Airport Security Screeners Supported by Automated Explosives Detection Systems for Cabin Baggage Screening." *Applied Ergonomics* 86: 103094. doi:10.1016/j.apergo.2020.103094.

Huegli, D., S. Merks, and A. Schwaninger. 2023. "Benefits of Decision Support Systems in Relation to Task Difficulty in Airport Security X-Ray Screening." *International Journal of Human–Computer Interaction* 39 (19): 3830–3845. doi:10.1080/10447318.2022.2107775.

Jian, J.-Y., A. M. Bisantz, and C. G. Drury. 2000. "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics* 4 (1): 53–71. doi:10.1207/S15327566IJCE0401_04.

Koller, S. M., C. G. Drury, and A. Schwaninger. 2009. "Change of Search Time and Non-Search Time in X-Ray Baggage Screening Due to Training." *Ergonomics* 52 (6): 644–656. doi:10.1080/00140130802526935.

Koller, S. M., D. Hardmeier, S. Michel, and A. Schwaninger. 2008. "Investigating Training, Transfer and Viewpoint Effects Resulting from Recurrent CBT of X-Ray Image Interpretation." *Journal of Transportation Security* 1 (2): 81–106. doi:10.1007/s12198-007-0006-4.

Kunar, M. A., D. G. Watson, S. Taylor-Phillips, and J. Wolska. 2017. "Low Prevalence Search for Cancers in Mammograms: Evidence Using Laboratory Experiments and Computer Aided Detection." *Journal of Experimental Psychology. Applied* 23 (4): 369–385. doi:10.1037/xap0000132.

Lawrence, R. K., and J. Pratt. 2022. "Salience Matters: Distractors May, or May Not, Speed Target-Absent Searches." *Attention, Perception & Psychophysics* 84 (1): 89–100. doi:10.3758/s13414-021-02406-x.

Lee, J., and N. Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human-Machine Systems." *Ergonomics* 35 (10): 1243–1270. doi:10.1080/00140139208967392.

Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. doi:10.1518/hfes.46.1.50_30392.

Liang, K. J., J. B. Sigman, G. P. Spell, D. Strellis, W. Chang, F. Liu, T. Mehta, and L. Carin. 2019. "Toward Automatic Threat Recognition for Airport X-Ray Baggage Screening with Deep Convolutional Object Detection." *arXiv Prep*: 1–11.

Luck, S. J., N. Gaspelin, C. L. Folk, R. W. Remington, and J. Theeuwes. 2021. "Progress toward Resolving the Attentional Capture Debate." *Visual Cognition* 29 (1): 1–21. doi:10.1080/13506285.2020.1848949.

Manzey, D., N. Gérard, and R. Wiczorek. 2014. "Decision-Making and Response Strategies in Interaction with Alarms: The Impact of Alarm Reliability, Availability of Alarm Validity Information and Workload." *Ergonomics* 57 (12): 1833–1855. doi:10.1080/00140139.2014.957732.

Mendes, M., A. Schwaninger, and S. Michel. 2011. "Does the application of virtually merged images influence the effectiveness of computer-based training in x-ray screening?" Proceedings of the 45th IEEE International Carnahan Conference on Security Technology, Mataro Spain, October 18–21, 2011.

Meuter, R. F., and P. F. Lacherez. 2016. "When and Why Threats Go Undetected: Impacts of Event Rate and Shift Length on Threat Detection Accuracy during Airport Baggage Screening." *Human Factors* 58 (2): 218–228. doi:10.1177/0018720815616306.

Meyer, J. 2001. "Effects of Warning Validity and Proximity on Responses to Warnings." *Human Factors* 43 (4): 563–572. doi:10.1518/001872001775870395.

Meyer, J. 2004. "Conceptual Issues in the Study of Dynamic Hazard Warnings." *Human Factors* 46 (2): 196–204. doi:10.1518/hfes.46.2.196.37335.

Meyer, J., R. Wiczorek, and T. Günzler. 2014. "Measures of Reliance and Compliance in Aided Visual Scanning." *Human Factors* 56 (5): 840–849. doi:10.1177/0018720813512865.

Meyer, J., and J. K. Kuchar. 2021. "Maximal benefits and possible detrimental effects of binary decision aids." 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 1–6. doi:10.1109/ICHMS53169.2021.9582632.

Moher, J. 2020. "Distracting Objects Induce Early Quitting in Visual Search." *Psychological Science* 31 (1): 31–42. doi:10.1177/0956797619886809.

Mosier, K. L., and D. Manzey. 2020. "Humans and Automated Decision Aids: A Match Made in Heaven." In *Human Performance in Automated and Autonomous Systems. Current Theory and Methods*, edited by P. A. Hancock & M. Mouloua, 19–41. Boca Raton, FL: Taylor & Francis Group.

Onnasch, L., C. D. Wickens, H. Li, and D. Manzey. 2014. "Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis." *Human Factors* 56 (3): 476–488. doi:10.1177/0018720813501549.

Parasuraman, R., T. B. Sheridan, and C. D. Wickens. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics* 30 (3): 286–297. doi:10.1109/3468.844354.

Parasuraman, R., and C. D. Wickens. 2008. "Humans: Still Vital after All These Years of Automation." *Human Factors* 50 (3): 511–520. doi:10.1518/001872008X312198.

Petrozziello, A., and I. Jordanov. 2019. "Automated Deep Learning for Threat Detection in Luggage from X-Ray Images.". In *Analysis of Experimental Algorithms*, edited by I. Kotseiras, E. Parsopoulos, Konstatinos, and A. Tsokas, 505–512. Springer International Publishing.

Pharmer, R. L., C. D. Wickens, B. A. Clegg, and C. A. P. Smith. 2021. "Effect of Procedural Elements on Trust and Compliance with an Imperfect Decision Aid." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65 (1): 633–637. doi:10.1177/1071181321651191.

Posner, M. I., C. R. Snyder, and B. J. Davidson. 1980. "Attention and the Detection of Signals." *Journal of Experimental Psychology* 109 (2): 160–174. doi:10.1037/0096-3445.109.2.160.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rice, S., and J. S. McCarley. 2011. "Effects of Response Bias and Judgment Framing on Operator Use of an Automated Aid in a Target Detection Task." *Journal of Experimental Psychology* 17 (4): 320–331. doi:10.1037/a0024243.

Rieger, T., L. Heilmann, and D. Manzey. 2021. "Visual Search Behavior and Performance in Luggage Screening: effects of Time Pressure, Automation Aid, and Target Expectancy." *Cognitive Research: Principles and Implications* 6 (1): 12. doi:10.1186/s41235-021-00280-7.

Rousselet, G. A., and R. R. Wilcox. 2020. "Reaction Times and Other Skewed Distributions." *Meta-Psychology* 4: 1630. doi:10.15626/MP.2019.1630.

Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors* 49 (1): 76–87. doi:10.1518/001872007779598082.

Ruthruff, E., M. Faulks, J. W. Maxwell, and N. Gaspelin. 2020. "Attentional Dwelling and Capture by Color Singletons." *Attention, Perception & Psychophysics* 82 (6): 3048–3064. doi:10.3758/s13414-020-02054-7.

Sauer, J., A. Chavaillaz, and D. Wastell. 2016. "Experience of Automation Failures in Training: effects on Trust, Automation Bias, Complacency and Performance." *Ergonomics* 59 (6): 767–780. doi:10.1080/00140139.2015.1094577.

Schuster, D., J. Rivera, B. C. Sellers, S. M. Fiore, and F. Jentsch. 2013. "Perceptual Training for Visual Search." *Ergonomics* 56 (7): 1101–1115. doi:10.1080/00140139.2013.790481.

Schwaninger, A., and F. Hofer. 2004. "Evaluation of CBT for Increasing Threat Detection Performance in X-Ray Screening.". In *The Internet Society 2004, Advances in Learning, Commerce and Security*, edited by K. Morgan & M. J. Spector , 147–156. Wessex: WIT Press.

Schwaninger, A., and S. Merks. 2019. "Single-View, Multi-View and 3D Imaging for Baggage Screening: what Should be Considered for Effective Training?" *Aviation Security International Magazine*.

Schwaninger, A. 2009. "Why Do Airport Security Screeners Sometimes Fail in Covert Tests?" Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology, October 5-8, 41–45. doi:10.1109/CCST.2009.5335568.

Schwarz, W., and J. Miller. 2016. "GSDT: An Integrative Model of Visual Search." *Journal of Experimental Psychology*. 42 (10): 1654–1675. doi:10.1037/xhp0000247.

Skorupski, J., and P. Uchroński. 2016. "A Human Being as a Part of the Security Control System at the Airport." *Procedia Engineering* 134: 291–300. doi:10.1016/j.proeng.2016.01.010.

Sorkin, R. D., and D. D. Woods. 1985. "Systems with Human Monitors: A Signal Detection Analysis." *Human–Computer Interaction* 1 (1): 49–75. doi:10.1207/s15327051hci0101.

Spain, R. D. 2009. "The Effects of Automation Expertise, System Confidence, and Image Quality on Trust, Compliance, and Performance." Doctoral diss., Old Dominion Uniersity. doi:10.25777/q87j-mr24.

Sterchi, Y., and A. Schwaninger. 2015. "A First Simulation on Optimizing EDS for Cabin Baggage Screening Regarding Throughput." Proceedings of the 49th IEEE International Carnahan Conference on Security Technology, Taipei Taiwan, September 21–24, 55–60. doi:10.1109/CCST.2015.7389657.

Thornton, I. M., and S. Zdravković. 2020. "Searching for Illusory Motion." *Attention, Perception & Psychophysics* 82 (1): 44–62. doi:10.3758/s13414-019-01750-3.

Turner, S. 1994. *Terrorist Explosive Sourcebook: Countering Terrorist Use of Improvised Explosive Devices*. Boulder, CO: Paladin Press.

von Bastian, C. C., A. Schwaninger, and S. Michel. 2008. "Do Multi-View X-Ray Systems Improve X-Ray Image Interpretation in Airport Security Screening?" *Zeitschrift Für Arbeitswissenschaft 3* 62: 165–173. doi:10.3239/9783640684991.

Wells, K., and D. A. Bradley. 2012. "A Review of X-Ray Explosives Detection Techniques for Checked Baggage." *Applied Radiation and Isotopes* 70 (8): 1729–1746. doi:10.1016/j.apradiso.2012.01.011.

Wetter, O., D. Hardmeier, and F. Hofer. 2008. "Covert Testing at Airports: Exploring Methodology and Results." Proceedings of the 43rd IEEE International Carnahan Conference on Security Technology, Zurich Switzerland, October 5–8. doi:10.1109/CCST.2008.4751328.

Wickens, C. D., A. Science, B. A. Clegg, A. Z. Vieane, F. Collins, and A. L. Sebok. 2015. "Complacency and Automation Bias in the Use of Imperfect Automation." *Human Factors* 57 (5): 728–739. doi:10.1177/0018720815581940.

Wiegmann, D. A., J. S. McCarley, A. F. Kramer, and C. D. Wickens. 2006. "Age and Automation Interact to Influence Performance of a Simulated Luggage Screening Task." *Aviation, Space, and Environmental Medicine* 77 (8): 825–831.

Wolfe, J. M. 2021. "Guided Search 6.0: An Updated Model of Visual Search." *Psychonomic Bulletin & Review* 28 (4): 1060–1092. doi:10.3758/s13423-020-01859-9.

Wolfe, J. M. 2022. "How One Block of Trials Influences the Next: persistent Effects of Disease Prevalence and Feedback on Decisions about Images of Skin Lesions in a Large Online Study." *Cognitive Research: Principles and Implications* 7 (1): 10. doi:10.1186/s41235-022-00362-0.

Wolfe, J. M., G. A. Alvarez, R. Rosenholtz, Y. I. Kuzmova, and A. M. Sherman. 2011. "Visual Search for Arbitrary Objects in Real Scenes." *Attention, Perception & Psychophysics* 73 (6): 1650–1671. doi:10.3758/s13414-011-0153-3.

Wolfe, J. M., T. S. Horowitz, M. J. Van Wert, N. M. Kenner, S. S. Place, and N. Kibbi. 2007. "Low Target Prevalence is a Stubborn Source of Errors in Visual Search Tasks." *Journal of Experimental Psychology. General* 136 (4): 623–638. doi:10.1037/0096-3445.136.4.623.Low.

Wolfe, J. M., and M. J. Van Wert. 2010. "Varying Target Prevalence Reveals Two Dissociable Decision Criteria in Visual Search." *Current Biology: CB* 20 (2): 121–124. doi:10.1016/j.cub.2009.11.066.

Xiao, H., S. Nazir, H. Li, H. U. Khan, and C. Li. 2021. "Decision Support System to Risk Stratification in the Acute Coronary Syndrome Using Fuzzy Logic." *Scientific Programming* 2021: 1–9. doi:10.1155/2021/6571905.

Zirk, A., R. Wiczorek, and D. Manzey. 2020. "Do we Really Need More Stages? Comparing the Effects of Likelihood Alarm Systems and Binary Alarm Systems." *Human Factors* 62 (4): 540–552. doi:10.1177/0018720819852023.