Applied Ergonomics 82 (2020) 102916

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Some cues are more equal than others: Cue plausibility for false alarms in baggage screening

Alain Chavaillaz[a,*], Adrian Schwaninger[b], Stefan Michel[b], Juergen Sauer[a]

[a] Department of Psychology, University of Fribourg, Fribourg, Switzerland
[b] School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), Olten, Switzerland

ARTICLE INFO

Keywords:
Automation ability
Screening performance
Trust

ABSTRACT

This study investigated the effects of cue plausibility in a baggage screening task. 120 participants had to indicate whether a prohibited item was present in a series of grey-scaled X-ray images of baggage. They were assisted by a support system, which pointed at the location of a suspicious object. A 2 × 2 × 2 between-subjects design was used. Cue plausibility for false alarms (i.e. how the cued object was similar to a prohibited item) and support system reliability were manipulated at two levels (high/low). Furthermore, half of participants were provided with a rationale about automation failures (RAF) to reduce their negative impact on trust and performance. The results showed lower performance and more compliance with automation suggestions when cues were implausible than plausible. The RAF increased the response time and did not improve detection performance. Overall, this suggests that effective (computer-based) training is needed to reduce the negative effect of plausible cues.

1. Introduction

Screening baggage for prohibited items is a complex visual inspection task that consists of visual search and decision-making (Wolfe and van Wert, 2010). This can be challenging because prohibited items (also referred to as targets in this article) may be hidden among everyday objects and can be difficult to detect in an X-ray image for three reasons that have been described as image-based factors (Schwaninger et al., 2005). First, the potential target may be displayed from an unfamiliar point of view (e.g., non-canonical view of a gun). Second, a harmless object may be superimposed on the target making it more difficult to recognize. Third, the piece of baggage may be densely packed with many items so that the potential target is less easily spotted due to many distractors (i.e. high level of clutter). Whereas the number of main categories of prohibited items is limited (i.e. guns, knives, bombs, other), each category includes many different types of items.

Airport security officers (screeners) need to have mental representations of a large number of prohibited items as well as of harmless everyday objects, which they should be able to activate at any time (Jiang et al., 2004). These are not the only challenges in X-ray image inspection and visual search (for recent reviews see Biggs et al., 2018; Biggs and Mitroff, 2015). The inspection task therefore may be facilitated by providing screeners with an automated threat detection system that highlights areas in X-ray images that might contain a prohibited item (e.g. Chavaillaz et al., 2018; Rice and McCarley, 2011).

In the current study, we investigated the impact of such devices, in particular regarding certain (almost inevitable) shortcomings of the detection algorithm's performance (e.g. not fully reliable operation, erroneous detection of non-prohibited objects). We also examined whether it may reduce the negative impact of a not very powerful detection algorithm if a rationale was provided about how the automatic system functions and where it is likely to fail.

A central purpose of automation is to support operators in their work activities. Like humans, the automation can however sometimes fail and this may prevent operators from using it in an optimal way due to a lack of trust (e.g. Parasuraman and Riley, 1997; Wickens and Dixon, 2007). Lee and Moray (1992) suggested that automation performance represents an determining factor of how frequent automation is used and how much it is trusted. Automation performance refers to what the automation does to achieve the desired goals (Lee and See, 2004).

The main goal of automation in baggage screening is to detect prohibited items among harmless everyday objects. When the automation detects a potential prohibited item, it marks it with a frame, signaling to the human screener that the item might represent a threat (Hättenschwiler et al., 2018). Such a system can produce two types of

* Corresponding author. Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700, Fribourg, Switzerland.
E-mail addresses: alain.chavaillaz@unifr.ch (A. Chavaillaz), adrian.schwaninger@fhnw.ch (A. Schwaninger), stefan.michel@fhnw.ch (S. Michel), juergen.sauer@unifr.ch (J. Sauer).

https://doi.org/10.1016/j.apergo.2019.102916
Received 31 October 2018; Received in revised form 24 January 2019; Accepted 2 August 2019
Available online 08 August 2019

failure: misses and false alarms (FA). In case of a miss, automation indicates the absence of a target when there is one, which can lead to serious consequences (e.g. missing a bomb can cause a plane crash). In case of an FA, automation indicates the presence of a target when there is none, which can have an impact on the workload of screeners and may also impair operational efficiency of the security checkpoint, because more passenger bags are sent (unnecessarily) to secondary search that typically involves more time consuming explosive trace detection and manual search (Sterchi and Schwaninger, 2015). Automation performance encompasses different aspects, including automation reliability and ability (Lee and See, 2004). The number of correct recommendations provided by the automatic system can be described as *automation reliability*. This is in contrast to *automation ability*, which refers to how well the automatic system is able to identify an object as a prohibited item.

There is a plethora of studies that have investigated effects of automation reliability. The vast majority of them showed lower performance and lower trust ratings when automation reliability decreased. This was found in numerous work environments, such as waste processing (Wiegmann et al., 2001), process control (Chavaillaz et al., 2016), in-vehicle navigation (Ma and Kaber, 2007), flight simulation (Bailey and Scerbo, 2007), or baggage screening (e.g. Chavaillaz et al., 2018; Rice and McCarley, 2011). Low-reliability automation also resulted in lower compliance and reliance levels during automation usage (e.g. Rice and McCarley, 2011). Both are important concepts in automation research. Compliance refers to the propensity to adopt automation recommendations when it indicates the presence of a target, whereas reliance corresponds to the probability to accept automation suggestions when it indicates the absence of a target (Meyer, 2001).

In contrast to research on automation reliability, the literature on the effects of automation ability is scarce. The degree to which targets are easy or hard to detect by automated threat detection algorithms may be an important determinant of how automation ability is perceived by operators. In a series of studies on visual search, Madhavan and colleagues (2006; 2004) found that when automation had missed several easily detectable targets (easy-miss condition), participants trusted their own abilities rather than the one of the automatic system, even in complex situations in which automation provided the correct answer. The opposite result pattern was observed when misses were more difficult to notice (difficult-miss condition). Interestingly, it was not the detection performance, but the response bias that was affected by the easy-to-spot automation failures. Participants in the easy-miss condition more often indicated the presence of a target than in the difficult-miss condition. Similar results were obtained regarding trust and self-confidence when automation issued only easily detectable FA, i.e. the stimuli in the search display could not be easily mistaken as a target (Madhavan et al., 2006). Altogether, this suggests that participants are sensitive to how easily detectable a target is that is related to an automation failure.

In the context of baggage screening, an important aspect that is related to automation ability represents the idea behind 'cue plausibility'. This term refers to the degree to which a cued object is similar to a potential target (as in text comprehension literature, e.g. Isberner et al., 2013). Cue plausibility for FAs of the automatic system may influence screener performance, use of automation and trust. It may be easier for the screener to indicate the absence of a target when the wrongly cued object does *not* look like a target (e.g. pair of socks). At the same time, if the automatic system wrongly identified such an object as a target (in some ways, this corresponds to the 'easy-miss condition' described by Madhavan et al., 2006), trust would be expected to be more negatively affected than if cue plausibility was high. Cue plausibility may also affect performance since a screener may be misled by a cued object sharing features with a potential target or having a similar shape of a potential target (e.g. a pen may be mistaken for a rotated knife). Consequently, more time is needed to make a decision and there is a higher probability that screeners produce a false alarm. As mentioned above, high FA rates of the human-machine system have an impact on the efficiency of the security checkpoint (i.e. lower throughput, Sterchi and Schwaninger, 2015).

Since it is almost inevitable that automatic systems fail at some point, we need to think about ways of managing this problem, avoiding the known effects of automation failure (e.g. decrease in trust, insufficient use of automation). To avoid these negative consequences, operators could be provided with a rationale of how reliably the automatic system operates and where its weaknesses and limitations are. There is some work in the literature that examined the effects of providing automation training or instructions to the operator. It showed that participants gave higher trust ratings towards the automatic system when they had received a rationale for potential automation failures than when they had not (Dzindolet et al., 2003). Positive effects on trust by providing information or training on automation ability were observed across a range of domains, including military target detectors (Dzindolet et al., 2003), context-aware systems (Lim et al., 2009), in-car adaptive cruise control systems (Verberne et al., 2012), and a variety of classifiers that assign data into categories (Ribeiro et al., 2016).

Interestingly, the way information about automation performance was framed affected how the automation was used (Lacson et al., 2005). A positive framing (e.g. 'the system makes about 80% of correct decisions and maximize hits and correct rejections') reduced reliance compared to a negative framing ('the system failed 20% of the time and minimize misses and false alarms') and a neutral framing (i.e. both positive and negative framings). However, no such effect was observed for compliance. While the research literature indicated some effect of this form of automation training or instructions on trust and automation use, little is known about its influence on performance of the human-machine system. The only study that examined performance as an outcome variable found no effect of information framing (Lacson et al., 2005).

This study extends previous research by investigating effects of cue plausibility and automation reliability, and by providing a rationale on how the automatic system functions and where it is likely to fail. We were interested in effects on performance (d', response bias, response times), use of automation (compliance and reliance), and subjective measures (trust in automation, perceived automation reliability, workload). 120 student participants were asked to indicate the presence (or absence) of a prohibited item in a series of grey-scaled X-ray images of cabin baggage. They were supported in this task by an automated threat detection system, which provided cues indicating areas which might contain a prohibited item.

First, performance is expected to be better and trust ratings to be lower when implausible cues (i.e. low similarity between distractor and target) are presented than plausible ones (i.e. high similarity between distractor and target). The difference in trust ratings are expected to be larger when participants are instructed about automation failures than when they are not. Furthermore, a rationale of automation failures (RAF) will reduce the difference in performance and increase the difference in trust ratings usually observed among participants working with highly reliable and less reliable automated systems.

2. Method

2.1. Participants

93 female and 27 male psychology students of the University of Fribourg (N = 120) took part in the present experiment. They earned course credits in return for their participation. Participants were aged from 18 to 30 years (M = 21.90, SD = 2.04). Informed consent was obtained from participants prior to the beginning of the experimental protocol. The study was approved by the Ethics Committee of the Department of Psychology at the University of Fribourg (Switzerland).

2.2. Simulation

Participants worked with the Luggage Inspection Simulation (LIS; Chavaillaz et al., 2018, 2019) during experimental task completion. This simulation models the basic functions of a computer monitor of an X-ray machine at an airport security checkpoint for cabin baggage screening.

Grey-scaled X-ray images of cabin baggage served as stimuli. They were obtained by digitally merging a prohibited item into an X-ray image of real baggage. The verisimilitude of these stimuli was assessed by experts (former screeners at a large European airport) before being used in research or for screener training. In the present experiment, guns and knives were selected as prohibited objects since they can be recognized without prior experience (unlike other prohibited items like bombs).

To assess the plausibility of each cued object to be used in the main experiment, a pilot study was conducted. 38 participants from a previous study (Chavaillaz et al., 2018) were invited into the laboratory again to take part in the pilot study. They inspected the 256 X-ray images to be used in the main test. Each image contained a cued object. The object cued in 128 target-present images was always the target. In the 128 target-absent images, the cued object had a similar shape as a target (plausible cue) or had a totally different shape that could not be mistaken for a target (implausible cue). A plausible cue and an implausible one were determined for each target-absent image. Each participant inspected half of the target-absent images with a plausible cue and the other half with an implausible cue. The images containing plausible and implausible cues were counterbalanced across participants. Participants were asked to rate how plausible it was that the cued object was a target, using a slider (from 0 = 'implausible' to 1 = 'plausible') displayed beneath each image. The image remained on screen until the participants made a response. The item-based analysis on target-absent images showed that cued objects in the implausible condition had significantly lower ratings ($M = 0.25$, $SD = 0.17$) than in the plausible condition ($M = 0.40$, $SD = 0.26$), $t(127) = 15.081$, $p < .001$, Cohen's $D = 1.333$.

Furthermore, the number of objects in the baggage (bag complexity), their degree of superimposition (i.e. to what extent a harmless object is superimposed on a prohibited one), and the rotation of the target were counterbalanced across pictures.

Participants were provided with an automatic support system to assist them in X-ray image inspection. In each trial, the system indicated the presence or absence of a target in the search display by means of a direct cue (Goh et al., 2005). When it suggested the presence of a target, it surrounded the potential prohibited object with a rectangular red frame. The word 'Target' was also displayed in red on the top of the picture (see Fig. 1A, C and D). When no target was detected, the words 'No target' were written in green on the top of the picture (Fig. 1B).

2.3. Design

The current study used a 2x2x2-factorial between-subjects design, with the following independent variables: cue plausibility, system reliability and RAF. Cue plausibility of system false alarms (high vs. low) referred to the degree of similarity between the cued object and a potential target when the automation indicated the presence of a target (even though there was no target in the current picture). In the plausible condition the cued objects looked like a potential target (see Fig. 1C), whereas they did not in the implausible condition (see Fig. 1D). Different levels of cue plausibility were tested and validated in a pilot study (for a description, see below). The second factor was system reliability (high vs. low). In the low-reliability condition, 71.88% of system recommendations were correct (d' = 1.27, see below for the formula) whereas in the high-reliability condition the rate of correct responses was at 90.63% (d' = 2.75). In both conditions, the numbers of miss and false alarms were the same. The third factor was RAF (yes vs. no), with half of the participants being given a rationale of how automation works and why it sometimes fails.

2.4. Dependent measures

2.4.1. Performance

Three measurements were taken to capture participant performance. *Detection performance* corresponds to participant ability to identify the presence (or absence) of a target. It is measured by using d' [z(Hit)-z(FA)] (see Macmillan and Creelman, 2005). Participants' *response bias* (i.e. the tendency to respond positively or negatively) was assessed by using the criterion c [-0.5*(z(Hit) + z(FA))] (see Macmillan and Creelman, 2005). Finally, the *response time* [s] was measured for target-present and target-absent trials, respectively.

2.4.2. Use of automation

Two measurements of automation use were taken: compliance and reliance (Meyer, 2001). Using the same approach like Rice and McCarley (2011), *compliance* (%) measured participants' propensity to follow automation recommendation when it indicated the presence of a target (i.e. number of positive responses given by the participant when the automation provided a positive response, divided by the total number of positive responses provided by the automation). *Reliance* (%) estimated to what extent participants acknowledged automation recommendation when it indicated the absence of a target (i.e. number of negative responses given by the participant when the automation provided a negative response, divided by the total number of negative responses provided by the automation).

2.4.3. Subjective measures

The following subjective variables were measured: trust, accuracy of reliability estimate, and workload. (a) *Trust* in automation was measured by the Checklist of Trust between People and Automation (CTPA, Jian et al., 2000). The questionnaire was composed of 12 items, rated on a 7-point Likert scale. An item example was: 'I can trust the system'. (b) Perceived automation reliability was assessed by using the following question: 'How reliable was the support system during task completion (0–100%)?'. Due to the two levels of automation reliability used in the experimental design, we computed an *Accuracy of Reliability Estimate* (ARE). It corresponded to the difference between the actual and the perceived automation reliability. A positive value indicated that participants overestimated automation reliability, whereas a negative value reflected an underestimation. (c) Furthermore, participants completed the six items of the NASA-TLX (Hart and Staveland, 1988) on a 20-point Likert scale to assess their *workload*.

2.5. Procedure

The experimental session consisted of a pre-test and main test. The pre-test was conducted to ensure the homogeneity of participants between experimental groups with regard to their inherent ability to detect prohibited items. The impact of cue plausibility, system reliability and RAF was measured during the main test.

2.5.1. Pre-test

In this phase, participants were asked to inspect a set of 128 X-ray images of cabin baggage (64 target-present and 64 target-absent) from the X-Ray Object Recognition Test (Hardmeier et al., 2005; Schwaninger et al., 2005). They first read instructions about the task and response modalities, and then completed a practice block of eight trials (with feedback). Afterwards, they were asked to complete two blocks of 64 trials (without feedback) with a 5-min break in-between. The two sets of prohibited objects (i.e. guns and knives) were presented for 10 s each at the beginning of the practice block and again before the first experimental block to allow participants to become familiar with them. Participants gave their responses by clicking on the left or right

Fig. 1. Screenshots of the interface of LIS (Luggage Inspection Simulation) showing (A) a valid cue for a target-present image (i.e. knife is correctly identified), (B) a valid cue for a target-absent image (i.e. image contains no target and is recognized as such by the support system) (C) an invalid but plausible cue (i.e. support of the roller is mistaken for a gun), and (D) an implausible but invalid cue (i.e. plastic box is mistaken for a prohibited item). On the top right of the display, a count-down timer indicates the time left for taking a decision. *2-column fitting image.*

mouse button. The button-response mapping was counterbalanced across participants. For each trial, a fixation cross was presented for 500 ms followed by the presentation of the X-ray image. It stayed on screen for 4s and participants had 20s to respond. The next trial started after participants had responded or at the end of the time limit. If no response was made, it was scored as a target-absent response. A blank screen of 1.5 s was displayed between two trials.

2.5.2. Main test

At the beginning of the main test, participants read on-screen instructions about the automatic support system. The instruction explained how the direct cues worked and that they may sometimes be wrong (the exact reliability level was not mentioned). For the participants in the RAF group, there were additional instructions providing a rationale why automation might fail (e.g. Dzindolet et al., 2003). These instructions emphasized that the detection algorithm may mistake a non-target for a target because of its similarity with a target. The procedure of the main test was the same as for the pre-test with four exceptions. First, the practice block contained 32 trials with feedback about the correctness of the given answer. Second, there were four experimental blocks of 64 trials each with breaks of 2 min between blocks. Third, the X-ray image remained on screen for up to 20 s or until participants gave their response. Fourth, participants completed the fatigue scale at the end of each block. After the last block, participants filled in a series of online questionnaires (assessing trust in automation, perceived reliability level, and perceived workload).

2.6. Data analysis

To ensure that participants' inherent detection ability was similar across experimental groups, two analyses were carried out. The Levene test showed that performance was homogeneous across groups, $F(1,112) = 0.360$, $p = .924$. Furthermore, a one-way ANOVA showed no significant effect on detection performance (d') across experimental groups during the pre-test, $F(1,112) = 1.700$, $p = .116$, $\eta_p^2 = 0.096$, suggesting that participants' inherent detection ability did not differ significantly between groups. For the main test, a $2 \times 2 \times 2$ analysis of variance was carried out, with cue plausibility, system reliability, and RAF as between-subjects factors.

3. Results

3.1. Performance

3.1.1. Detection performance

The three-way ANOVA showed that detection performance (d') for participants receiving plausible cues ($M = 2.14$, $SD = 0.45$) was significantly lower than for participants being given implausible cues ($M = 2.57$, $SD = 0.38$), $F(1,112) = 35.700$, $p < .001$, $\eta_p^2 = 0.242$. Furthermore, participants working with a highly reliable system ($M = 2.48$, $SD = 0.41$) showed higher detection performance than participants working with a less reliable system ($M = 2.23$, $SD = 0.49$), $F(1,112) = 12.070$, $p < .001$, $\eta_p^2 = 0.097$. No effect was observed for RAF, $F(1,112) = 0.378$, $p = .540$, $\eta_p^2 = 0.003$. Finally, no interactions were significant, all Fs < 3.000.

3.1.2. Response bias

The three-way ANOVA revealed a main effect of cue plausibility on response bias (i.e. a shift in participants' response criterion according to cue plausibility), $F(1,112) = 41.439$, $p < .001$, $\eta_p^2 = 0.270$. Participants with plausible cues tended to respond more often that there was no target ($M = -0.21$, $SD = 0.28$) whereas participants with implausible cues overall responded more often that there was a target ($M = 0.18$, $SD = 0.29$). All other effects were not significant, all Fs < 1.2.

3.1.3. Response time

As expected, the response times were significantly higher for target absent ($M = 4.06$ s, $SD = 1.72$) than target-present images ($M = 1.48$ s, $SD = 0.50$), $t(119) = 19.773$, $p < .001$, Cohen's $D = 1.805$. For this reason, response times for target-present and target-absent images were analyzed separately.

3.1.3.1. Target-present. The $2 \times 2 \times 2$ ANOVA revealed no main effects of cue plausibility and system reliability, both Fs < 2.0. However, participants without RAF ($M = 1.37$ s, $SD = 0.34$) were significantly faster to indicate the presence of a target than participants with RAF ($M = 1.59$ s, $SD = 0.60$), $F(1,112) = 5.968$, $p = .016$, $\eta_p^2 = 0.051$. None of the interactions was significant, all Fs < 1.

3.1.3.2. Target-absent. Again, the $2 \times 2 \times 2$ ANOVA showed no main effects of cue plausibility, system reliability and instructions, all Fs < 1. There was an interaction between system reliability and RAF, $F(1,112) = 4.295$, $p = .041$, $\eta_p^2 = 0.037$ (see Fig. 2). Further analyses revealed that in the low-reliability condition participants without receiving a rationale ($M = 3.54$ s, $SD = 1.48$) responded faster than participants who received RAF ($M = 4.50$ s, $SD = 1.51$), $t(58) = 2.148$, $p = .034$, Cohen's $D = 0.642$. The reverse pattern was observed in the high-reliability condition, but the difference was not significant, $t(58) = 0.782$, $p = .436$, Cohen's $D = 0.209$. No other interactions were found to be significant, all Fs < 1.

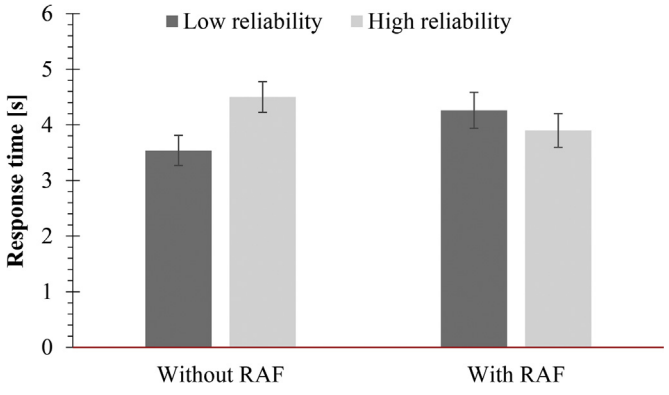

Fig. 2. Mean response time (and standard errors) for target-absent images as a function of system reliability (low vs. high) and rationale of automation failures (RAF; without vs. with). *1.5-column fitting image.*

3.2. Use of automation

3.2.1. Compliance

The 3-way ANOVA revealed that participants with plausible cues ($M = 79.45\%$, $SD = 9.93$) followed more often automation recommendations than participants with implausible cues ($M = 70.07\%$, $SD = 9.91$) when it indicated the presence of a target, $F(1,112) = 62.576$, $p < .001$, $\eta_p^2 = 0.358$. Furthermore, higher levels of compliance were observed in the high-reliability condition ($M = 82.23\%$, $SD = 6.50$) than in the low-reliability one ($M = 67.30\%$, $SD = 9.27$), $F(1,112) = 158.421$, $p < .001$, $\eta_p^2 = 0.586$. No other effects were significant, all Fs < 3.200.

3.2.2. Reliance

No significant effect was observed for cue plausibility, $F(1,112) = 2.692$, $p = .104$, $\eta_p^2 = 0.023$. However, participants working with a high-reliability system ($M = 85.43\%$, $SD = 8.58$) followed more often automation advice than participants under the low-reliability system ($M = 76.83\%$, $SD = 9.87$) when it reported the absence of a target, $F(1,112) = 25.764$, $p < .001$, $\eta_p^2 = 0.187$. No other significant effects were found, all Fs < 1.300.

3.3. Subjective measures

3.3.1. Trust

The overall level of trust was at $M = 3.96$ ($SD = 0.87$) of the 7-point Likert scale. The 3-way ANOVA revealed no significant main effects or interactions for trust ratings, all Fs < 1.

3.3.2. Accuracy of reliability estimate (ARE)

There was no significant effect of cue plausibility on ARE, $F(1,112) = 0.118$, $p = .731$, $\eta_p^2 = 0.001$. However, participants working with a high-reliability system ($M = -19.93$, $SD = 10.79$) underestimated system reliability more strongly than when operating a low-reliability system ($M = -10.29$, $SD = 12.67$), $F(1,112) = 19.537$, $p < .001$, $\eta_p^2 = 0.149$. All other main effects and interactions were not significant, all Fs < 1.300.

3.3.3. Workload

The overall ratings for perceived workload were at $M = 9.99$ ($SD = 1.93$), which in the middle of the NASA-TLX 20 point Likert scale (Hart and Staveland, 1988). No effects were significant, except for the 3-way interaction, $F(1,112) = 6.794$, $p = .010$, $\eta_p^2 = 0.057$ (see Fig. 3). Two further ANOVAs with system reliability and RAF were conducted separately for plausible and implausible cues. When automation provided plausible cues, neither the main effects, nor the interaction was significant, all Fs < 2.700. However, there was a significant interaction for implausible cues, $F(1,56) = 4.706$, $p = .034$, $\eta_p^2 = 0.078$. Post-hoc analyses showed that workload ratings were higher under low ($M = 11.27$, $SD = 1.35$) than high system reliability ($M = 9.66$, $SD = 2.05$) when no RAF was provided, $t(28) = 2.884$, $p = .007$, Cohen's $D = 0.929$. This difference between low ($M = 10.00$, $SD = 1.52$) and high system reliability ($M = 10.16$, $SD = 1.27$) was not observed for trained participants, $t(28) = -0.271$, $p = .788$, Cohen's $D = -0.114$. Finally, both main effects of system reliability and RAF were not significant, both Fs < 3.200.

4. Discussion

The aim of the current study was twofold. First, it investigated the effects of cue plausibility of false alarms of the automated threat

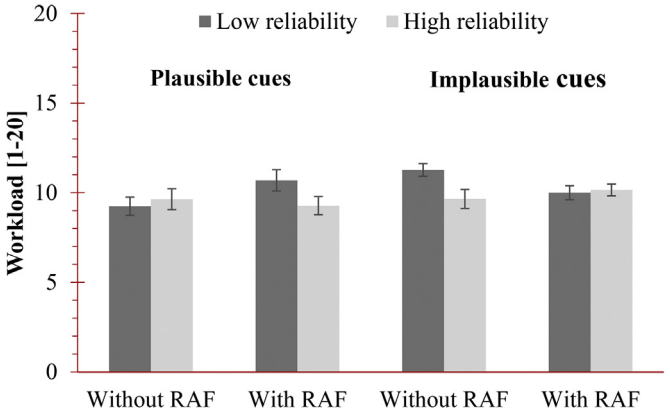

Fig. 3. Mean ratings of perceived workload (and standard errors) as a function of cue plausibility (plausible vs. implausible cues), system reliability (low vs. high), and rationale of automation failures (RAF; with vs. without). *1.5-column fitting image.*

detection system (i.e. how similar is a wrongly cued item to a potential target) under different operational conditions (i.e. low and high system reliability). Second, it examined the preventive effect of providing participants with a rationale for automation failures. Participants in the implausible cue condition (i.e. wrongly cued items were easy to classify as being prohibited or not) showed better detection performance than participants in the plausible cue condition. Furthermore, participants in the implausible cue condition showed a negative response bias (i.e. they were prone to report that there was a target rather than there was none), whereas participants in the plausible cue condition displayed a positive response bias. As expected, low system reliability impaired detection performance and reduced automation use (compliance and reliance) but, unexpectedly, trust ratings remained unaffected. Furthermore, participants systematically underestimated system reliability, with this effect being larger under high reliability than low reliability.

Participants displayed better performance (i.e. higher d'-score and faster response times) when wrong cues (false alarms of the automated threat detection system) were implausible than when they were plausible. Furthermore, participants tended to respond more often that a target was present when such cues were plausible (i.e. negative response bias) and that a target was absent when the cues were implausible (i.e. positive response bias). This is consistent with the findings of Madhavan et al. (2006), which showed that participants were sensitive to automation ability. At the same time, this represented a successful check of the experimental manipulation of cue plausibility.

Participants in the 'plausible cue'-condition showed higher compliance rates than in the one with implausible cues. This may be due to the following mechanism. If a cued object looks like a prohibited item (i.e. corresponding to a plausible cue), the human is more likely to accept an automation recommendation that turns out to be an FA, leading to higher compliance. It was interesting that cue plausibility had no effect on trust. The effects of cue plausibility on trust may depend on the mental model a screener has developed about the operation of the automatic system. For example, such effects would be expected if participants were surprised that an 'easy-to-detect' item (i.e. corresponding to implausible cues) was not detected by the automation. However, participants may have been aware that a machine algorithm operates very differently from human perceptual processes. Therefore, differences in detection abilities between machines and humans may not have come as a surprise.

This raises further questions about the complementary role of humans and machines in baggage screening. If the X-ray machine or the human screener carried out the task on their own (i.e. without any support from the other entity), the detection abilities would vary considerably as a function of the prohibited item being present in the baggage. For example, well trained and experienced screeners can detect fully functional improvised explosive devices containing explosive, detonators and a triggering device connected by wires rather well (e.g. Koller et al., 2009) whereas explosive material alone can be detected better by automated explosive detection systems (Hättenschwiler et al., 2018).

The poorer detection performance for plausible but wrong cues compared to implausible but wrong cues implies a need to look at ways of reducing the number of plausible wrong cues. This is probably achieved best by means of training. For example, a computer-based training system can be used, which is widely used so that screeners can learn what objects are prohibited and how they look like in X-ray images (e.g. Halbherr et al., 2013). In addition, everyday object training could further improve screeners' ability to distinguish harmless and prohibited items (Hättenschwiler et al., 2015).

With regard to the effects of automation reliability, our findings mostly confirm the results of previous research. For instance, Chavaillaz et al. (2018) and Rice and McCarly (2011) found that participants showed better detection performance under high reliability than low reliability. Similarly, as in previous work (e.g. Rice and McCarley, 2011), participants displayed lower rates of compliance and reliance under a low-reliability system than under a high-reliability system. However, in the present study, system reliability did not influence response time, which is in contrast to findings from previous research (e.g. Chavaillaz et al., 2018). Interestingly, trust was not affected by automation reliability. This is not in line with previous findings, which has typically found such an effect (e.g. Rice and McCarley, 2011). It was surprising that trust was unaffected given that the difference in automation reliability was noticed by participants, as demonstrated by the effect on perceived automation reliability. Remarkably, participants generally had difficulties in accurately assessing automation reliability. They generally underestimated automation reliability, which was observed even more so when system reliability was high. These results are in support of previous work, which found that the failure rate of automatic systems tended to be overrated (Chavaillaz et al., 2018; Wiegmann et al., 2001). This overrating seems to be even more pronounced for high-reliability systems, possibly due to the greater conspicuousness of automation failures when reliability is high rather than low (Chavaillaz et al., 2018).

Providing a short rationale about the working of automation may already influence participant behavior. However, in the present case there were mainly negative effects of RAF. It slowed down response time because participants became more cautious during visual inspection as they may have been more aware of the fact that the automation could fail. This was observed for target-present trials (in form of a main effect of RAF) as well as for target-absent trials (in form of an interaction). The slow-down in response time in itself would not have presented a problem if this negative effect had been compensated by positive ones such as better detection performance (or at least a more accurate automation reliability rating or improved trust ratings).

The only positive effect observed was that RAF had reduced workload under particularly difficult conditions (i.e. low reliability combined with implausible cues) as shown by an interaction between the independent variables. RAF offered in the present study appears to have been less effective than automation reliability training in previous studies. Previous work showed higher trust rating for trained participants than untrained ones (Dzindolet et al., 2003; Lim et al., 2009; Ribeiro et al., 2016; Verberne et al., 2012).

There may be a number of reasons why the present findings were not in line with previous work, mainly due to differences in the experimental procedures. In most of the studies (Lim et al., 2009; Ribeiro et al., 2016; Verberne et al., 2012), participants were only asked to rate their trust towards the system but without actually using it. In the study of Dzindolet et al. (2003), participants received performance feedback as part of training, which may have boosted the influence of training.

Although the present results suggested that this intervention was of little benefit in baggage screening and only slowed down participants in their decision-making process, we would advocate a further test of the effectiveness of automation reliability training by implementing a more extensive training session with interactive elements (e.g., providing examples of non-target cued objects). This modification may overcome two possible shortcomings of our RAF procedure (i.e. brevity and lack of interactivity).

The present study has a number of limitations. First, we used novices as participants instead of experts (e.g. real screeners). There is clearly a need to make use of novice participants due to the limited availability of screeners. Given that there are differences between screeners and novice participants (e.g., Chavaillaz et al., 2019; Clark et al., 2012), testing with screeners is strongly advised before recommendations from lab-based research involving novice participants are implemented. Second, we chose to use only guns and knives as prohibited items in the present study because novices are more familiar with these objects. Although reducing the range of potential prohibited items has made the task easier for the participants, we think that this does not represent a major problem because the match between expertise and task difficulty is similar to real screeners (i.e. screeners: more difficult targets/high expertise; student participants: easier targets/low expertise).

The present study holds several implications. First, given that cue plausibility had major effects on outcome measures (notably on detection performance), there is a need to reduce its negative effects. This can be achieved best by providing effective training to screeners to minimize the number of plausible objects. Since the current study only investigated cue plausibility for false alarms, future research should also look at misses and miscues (i.e. the cued item is harmless but somewhere else in the baggage there is a prohibited item). This helps address the question whether humans trust more a system that misses targets that are difficult to spot and recognize or a system that misses obvious targets (e.g. Madhavan et al., 2004). Second, providing a rationale about automation failures was not effective, i.e. it did not improve detection performance but slowed down decision time instead. We would suggest investigating the effectiveness of more intensive and interactive training (e.g. including performance feedback) to determine whether it improves the understanding of the strengths and weaknesses of the automatic system, which could increase detection performance. Third, automation designers should be aware of the paradoxical effect surrounding cue plausibility, which suggests that increasing the ability of detection systems may decrease screener performance. Screeners may take more time to take a decision, and mistake more often a non-target object for a target when the system provides plausible cues rather than implausible ones.

Overall, the current study has made some valuable contribution to the small body of research that examined the consequences of cue plausibility in human-machine interaction. There is clearly a need to extend this line of research given the complementary roles of screeners and increasingly automated X-ray machines in baggage screening.

Acknowledgements

This work was supported by the Swiss National Science Foundation [grant number 100019_149184]. The authors gratefully acknowledge the great help of Beatrice Macullo, Nina Leu and Flavia Pircher with data collection.

References

Bailey, N.R., Scerbo, M.W., 2007. Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. Theor. Issues Ergon. Sci. 8 (4), 321–348. https://doi.org/10.1080/14639220500535301.

Biggs, A.T., Kramer, M.R., Mitroff, S.R., 2018. Using cognitive psychology research to inform professional visual search operations. J. Appl. Res. Mem. Cogn. 7 (2), 189–198. https://doi.org/10.1016/j.jarmac.2018.04.001.

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. Appl. Cognit. Psychol. 29 (1), 142–148. https://doi.org/10.1002/acp.3083.

Chavaillaz, A., Schwaninger, A., Michel, S., Sauer, J., 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. Appl. Ergon. 61 (10), 1395–1408. https://doi.org/10.1080/00140139.2018.1481231.

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

Chavaillaz, A., Wastell, D., Sauer, J., 2016. System reliability, performance and trust in adaptable automation. Appl. Ergon. 52, 333–342. https://doi.org/10.1016/j.apergo.2015.07.012.

Clark, K., Cain, M.S., Adamo, S.H., Mitroff, S.R., 2012. Overcoming hurdles in translating visual search research between the lab and the field. In: Dodd, M.D., Flowers, J.H. (Eds.), The Influence of Attention, Learning, and Motivation on Visual Search. Springer, New York, NY, pp. 147–181.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance. Int. J. Hum. Comput. Stud. 58 (6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.

Goh, J., Wiegmann, D.A., Madhavan, P., 2005. Effects of automation failure in a luggage screening task: a comparison between direct and indirect cueing. In: Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society. 26-30 September, pp. 492–496 Orlando: FL.

Halbherr, T., Schwaninger, A., Budgell, G.R., Wales, A.W.J., 2013. Airport security screener competency: a cross-sectional and longitudinal analysis. Int. J. Aviat. Psychol. 23 (2), 113–129. https://doi.org/10.1080/10508414.2011.582455.

Hardmeier, D., Hofer, F., Schwaninger, A., 2005. The X-ray object recognition test (X-ray ORT)-a reliable and valid instrument for measuring visual abilities needed in X-ray screening. In: Proceedings of the 39th IEEE Carnahan Conference on Security Technology, pp. 189–192.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North-Holland, Amsterdam, pp. 139–183.

Hättenschwiler, N., Michel, S., Kuhn, M., Ritzmann, S., Schwaninger, A., 2015. A first exploratory study on the relevance of everyday object knowledge and training for increasing efficiency in airport security X-ray screening. In: 2012 International Carnahan Conference on Security Technology (ICCST). International Carnahan Conference on Security Technology (ICCST). IEEE, Taipei, Taiwan, pp. 25–30 September, 21-24.

Hättenschwiler, N., Sterchi, Y., Mendes, M., Schwaninger, A., 2018. Automation in airport security X-ray screening of cabin baggage: examining benefits and possible implementations of automated explosives detection. Appl. Ergon. 72, 58–68. https://doi.org/10.1016/j.apergo.2018.05.003.

Isberner, M.-B., Richter, T., Maier, J., Knuth-Herzig, K., Horz, H., Schnotz, W., 2013. Comprehending conflicting science-related texts: graphs as plausibility cues. Instr. Sci. 41 (5), 849–872. https://doi.org/10.1007/s11251-012-9261-2.

Jian, J.-Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. Int. J. Cogn. Ergon. 4 (1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.

Jiang, X., Gramopadhye, A.K., Melloy, B.J., 2004. Theoretical issues in the design of visual inspection systems. Theor. Issues Ergon. Sci. 5 (3), 232–247. https://doi.org/10.1080/1463922021000050005.

Koller, S.M., Drury, C.G., Schwaninger, A., 2009. Change of search time and non-search time in X-ray baggage screening due to training. Ergonomics 52 (6), 644–656. https://doi.org/10.1080/00140130802526935.

Lacson, F.C., Wiegmann, D.A., Madhavan, P., 2005. Effects of attribute and goal framing on automation reliance and compliance. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Sage CA. SAGE Publications, Los Angeles, CA, pp. 482–486.

Lee, J.D., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergon. 35 (10), 1243–1270. https://doi.org/10.1080/00140139208967392.

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46 (1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Lim, B., Dey, A.K., Avrahami, D., 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2119–2128.

Ma, R., Kaber, D.B., 2007. Effects of in-vehicle navigation assistance and performance on driver trust and vehicle control. Int. J. Ind. Ergon. 37 (8), 665–673. https://doi.org/10.1016/j.ergon.2007.04.005.

Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, first ed. Psychology Press, Mahwah, NJ, pp. 492.

Madhavan, P., Wiegmann, D.A., Lacson, F.C., 2004. Occasional automation failures on easy tasks undermines trust in automation. In: Proceedings of the 112th Annual Meeting of the American Psychological Association, pp. 1–6.

Madhavan, P., Wiegmann, D.A., Lacson, F.C., 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. Hum. Factors 48 (2), 241–256. https://doi.org/10.1518/001872006777724408.

Meyer, J., 2001. Effects of warning validity and proximity on responses to warnings. Hum. Factors 43 (4), 563–572. https://doi.org/10.1518/001872001775870395.

Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Hum. Factors 39 (2), 230–253. https://doi.org/10.1518/001872097778543886.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining - KDD '16. The 22nd ACM SIGKDD International Conference, San Francisco, California, USA. 13.08.2016 - 17.08.2016. ACM Press, New York, New York, USA, pp. 1135–1144.

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. J. Exp. Psychol. Appl. 17 (4), 320–331. https://doi.org/10.1037/a0024243.

Schwaninger, A., Hardmeier, D., Hofer, F., 2005. Aviation security screeners visual abilities & visual knowledge measurement. IEEE Aerosp. Electron. Syst. Mag. 20 (6), 29–35.

Sterchi, Y., Schwaninger, A., 2015. A first simulation on optimizing EDS for cabin baggage screening regarding throughput. In: 2012 International Carnahan Conference on Security Technology (ICCST). International Carnahan Conference on Security Technology (ICCST). IEEE, Taipei, Taiwan, pp. 55–60 September, 21-24.

Verberne, F.M.F., Ham, J., Midden, C.J.H., 2012. Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Hum. Factors 54 (5), 799–810. https://doi.org/10.1177/0018720812443825.

Wickens, C.D., Dixon, S.R., 2007. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. 8 (3), 201–212. https://doi.org/10.1080/14639220500370105.

Wiegmann, D.A., Rich, A., Zhang, H., 2001. Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. Theor. Issues Ergon. Sci. 2 (4), 352–367. https://doi.org/10.1080/14639220110110306.

Wolfe, J.M., van Wert, M.J., 2010. Varying target prevalence reveals two dissociable decision criteria in visual search. Curr. Biol. 20 (2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066.