Tobias Rieger, Dietrich Manzey

Human performance consequences of automated decision aids: The impact of time pressure

Citation details
Rieger, T., Manzey, D. (2020). Human Performance Consequences of Automated Decision Aids: The Impact of Time Pressure. Human Factors: The Journal of the Human Factors and Ergonomics Society, 001872082096501. https://doi.org/10.1177/0018720820965019.

Human performance consequences of automated decision aids: The impact of time pressure

Tobias Rieger

Technische Universität Berlin

Dietrich Manzey

Technische Universität Berlin

This is a post-peer-review, pre-copyedit version of an article published in Human Factors (accepted September 7th, 2020). The final authenticated version is available online at: http://dx.doi.org/10.1177/0018720820965019

Author Note

Address correspondence to: Tobias Rieger, Technische Universität Berlin, Department of Psychology and Ergonomics, Chair of Work, Engineering, and Organizational Psychhology, Marchstraße 12, 10587 Berlin, Germany, email: tobias.rieger@tu-berlin.de. The authors would like to thank Janina Dubberke for help in data collection of Experiment 1.

Abstract

Objective: The study addresses the impact of time pressure on human interactions with automated decision support systems (DSS) and related performance consequences.

Background: When humans interact with DSS, this often results in worse performance than could be expected from the automation alone. Previous research has suggested that time pressure might make a difference by leading humans to rely more on a DSS.

Method: In two laboratory experiments, participants performed a luggage screening task either manually, supported by a highly reliable DSS, or a low reliable DSS. Time provided for inspecting the X-rays was 4.5 vs 9 sec. Participants in the automation conditions were either shown the automation's advice prior (Experiment 1) or following (Experiment 2) their own inspection, before they made their final decision.

Results: In Experiment 1, time pressure compromised performance independent of whether the task was performed manually or with automation support. In Experiment 2, the negative impact of time pressure was only found in the manual, but not the two automation conditions. However, neither experiment revealed any positive impact of time pressure on overall performance, and the joint performance of human and automation was mostly worse than the performance of the automation alone.

Conclusion: Time pressure compromises the quality of decision making. Providing a DSS can reduce this effect, but only if the automation's advice follows the assessment of the human.

Application: The study provides suggestions for effective implementation of DSSs, but also supports concerns that highly reliable DSSs are not used optimally by human operators.

Keywords: Decision making, Human-automation interaction, Compliance and reliance, Stress, Trust in automation

Human performance consequences of automated decision aids: The impact of time pressure

Automation and specifically computer-based automated assistance systems are present in virtually every field today. The main ideas why automation is introduced in a large number of settings are that automation supposedly makes work safer, reduces workload, and improves overall performance (French, Duenser, & Heathcote, 2018; Sheridan & Parasuraman, 2005). One special case of automation are decision support systems (DSSs) which are intended to aid a human execute a certain task (Mosier & Fischer, 2010; Mosier & Manzey, 2020). DSSs can widely vary in their complexity but the general idea is that they provide users information on the true state of the world, based on automated processing and evaluation of information from the environment. They include a variety of systems ranging from binary alarm or target detection systems (e.g., smoke detection) to complex expert systems providing automatically generated diagnoses to radiologists or surgeons, based on artificial intelligence applications (e.g., Jiang et al., 2017). Regardless of the type of system, though, the final output to the user with DSSs is usually straightforward, i.e., includes a definite recommendation, diagnosis or just indication of a certain state, even though the processing (e.g., complex image processing) in the background might be highly complex. However, in a great deal of cases, humans do not use the DSS appropriately, which can result in automation misuse or disuse (Parasuraman & Riley, 1997). With respect to binary DSSs like alarm systems or target detection support, two different aspects of automation use which can be closely linked to potential misuse or disuse are compliance and reliance (Meyer, 2001, 2004). Specifically, operator compliance is defined by operators agreeing with the automation when it indicates that a target is present. Conversely, operator reliance is defined by operators agreeing with the automation when it indicates that no target is present. Thus, operator compliance can be considered to reflect the belief that a critical event is actually present when an alarm occurs, and operator reliance can be considered to reflect the belief that that the system will actually alert in case of a critical event (Meyer, 2001). However, there are cases where

the reliance or compliance of operators does not seem to be properly calibrated to the performance of a DSS.

Thus, interestingly, often when highly reliable DSSs are used, the joint performance of operator and automation is worse than that of the automation alone (see Bartlett & McCarley, 2017, for model-based accounts of that phenomenon; see also Alberdi, Povyakalo, Strigini, & Ayton, 2004; Meyer, 2001; Meyer, Wiczorek, & Günzler, 2014). Note that this refers to a hypothetical comparison of the performance of the automation with the operator possibly intervening (i.e., performance of the human-automation dyad) versus a situation where the automation alone would be in charge of the decision.

This shows that even with highly reliable systems, trust (and accordingly, compliance and reliance) are often mis-calibrated (Parasuraman & Riley, 1997), leading operators to alter outputs of the alarm system which were actually true.

Time Pressure and Automated DSSs

However, the degree of time pressure operators have in making their decisions while interacting with an automated DSS might make a difference in this respect. Time pressure is an everyday phenomenon in a plethora of workplaces and is usually considered a to-be-avoided workload factor in the human factors literature (e.g., Carayon & Gurses, 2008; Hendy, Liao, & Milgram, 1997; Moray, Dessouky, Kijowski, & Adapathya, 1991). Concerning trust in automation, we refer to the three-layered model of Hoff and Bashir (2015), where time pressure would be considered as an external situational factor. Note, however, that in the present research, our main focus is on behavioral consequences of trust rather than subjective measures of trust.

Specifically, there is evidence that time pressure can lead to more heuristic decision-making and thus not necessarily lead to concomitant performance decreases (see Gigerenzer & Gaissmaier, 2011; Shah & Oppenheimer, 2008, for reviews). In the context of human-automation-interaction, a possible heuristic might be an increased dependence on

the automation's suggestion. Using automation as a heuristic has been referred to as automation bias (Mosier, Skitka, Burdick, & Heers, 1996; Parasuraman & Manzey, 2010) and is usually considered as something that can negatively impact performance if critical events are overlooked due to automation misuse. If a DSS is highly reliable, however, automation bias can potentially even lead to improved performance—especially if the DSS performs better than the human alone. One factor which is known to increase the use of heuristics is time pressure (Payne, Bettman, & Johnson, 1988), and there are indeed findings which suggest that dependence on DSS increases under high time pressure (Rice, Hughes, McCarley, & Keller, 2008; Rice & Keller, 2009; Rice, Keller, Trafimow, & Sandry, 2010; Rice & Trafimow, 2012). For example, in their study, Rice and Keller (2009) found overall performance benefits of time pressure if the automation was highly reliable. To investigate this, they used a military context and presented their participants aerial view photographs with the participants having to decide whether there was a tank present or not. In order to vary time pressure, the time available for inspection of the photographs was varied across participants (8 vs. 2 sec). Moreover, participants were assigned a DSS to aid their decisions and DSS reliability was varied in five conditions (65%, 80%, 95%, 100%, manual). The key finding of this study was that participants who worked under high time pressure, depended more strongly on the DSS which led to concordant performance increases under high time pressure. However, in both this study (Rice & Keller, 2009) as well as later corroborating studies (Rice et al., 2010; Rice & Trafimow, 2012), the extremely high time pressure (i.e., 2 seconds for complex visual stimuli) might have left participants with no real choice but to follow the DSS. Thus, the main finding of a higher dependence on automation with time pressure might just be an artifact of this extreme time pressure condition. Moreover, time pressure was varied between-subjects—something that seems to be rather unlikely to happen in the real world where time pressure is expected to vary more frequently across different trials, depending on situational circumstances.

Consequently, we aim to revisit this issue using somewhat less extreme and more

realistic time pressure variation, giving participants an actual chance to inspect the stimulus themselves even under time pressure, as would be the case in most real-world scenarios.

One real-world context where both time pressure and automated DSSs can play an important role is luggage screening at airport security checkpoints (e.g., Chavaillaz, Schwaninger, Michel, & Sauer, 2018). Here, screeners must classify whether there is a prohibited article in the bag or not and time pressure while performing this screening task may vary frequently across the day depending on the number of passengers waiting in the line. Normally, airport security screeners have around 10 seconds per image (Buser, Sterchi, & Schwaninger, 2019), which decreases to around 4 seconds during busier periods (Schwaninger, Hardmeier, & Hofer, 2004). The basic task to be performed in luggage screening can be considered as a signal-detection task. As has formally been described in signal detection theory (SDT, Green & Swets, 1966), the performance in such tasks depends on two variables, i.e., the sensitivity (d') in terms of ability to distinguish between non-target (e.g., a pen) and targets (e.g., a knife), and the response threshold (C), i.e., on how much evidence decisions about the presence of targets are based (see Stanislaw & Todorov, 1999, for calculation of d' and C). DSSs are used in such tasks in order to help operators optimizing both aspects. However, depending on their capability, these systems might differ in their reliability, i.e., the probability that their advice is correct. Specifically, due to a safe-engineering approach, such target detection systems often have liberal response thresholds which make them more or less false-alarm prone. If time pressure does indeed increase DSS dependence, the related human performance consequences should be dependent on the reliability of the DSS. Only in the case that the automation reliability is considerably greater than the human alone performance, one would expect a visible benefit of the DSS in the joint human-automation performance.

The Present Experiments

The goal of the present experiments was to examine the influence of time pressure in a luggage screening task with and without an automated DSS available. To this end, we used a luggage screening task and participants performed the task either with no DSS support (i.e., manually), with a low reliability DSS (75% correct suggestions), or with a high reliability DSS (95% correct suggestions). Corresponding to most applications in safety-critical environments such as luggage screening, both conditions were essentially false-alarm prone. However, also some rare misses were simulated. 75% was chosen as the lower reliability level, in order to create a clear variation but still have a level which seems to be realistic for false-prone systems and which is still higher than the 70% threshold assumed to be the minimum reliability where automated systems are still considered to be useful (Wickens & Dixon, 2007). Contrasting to some earlier research (e.g., Rice & Keller, 2009), we varied time pressure blockwise within-subjects because real-world operators' work environments (and thus, time constraints) also change from time to time and do not remain constant. Furthermore, a time pressure condition was chosen which put the participants under considerable pressure to perform the task quickly, but also was long enough to get the task principally done even without automation.

In Experiment 1, we used a paradigm where participants in the automation conditions were first shown the DSS's advice and then made their own choice. Then, in Experiment 2, we reversed that order—that is, participants could first make their initial choice and were then shown the advice of the DSS with the possibility of either confirming or changing the own decision based on that advice.

Experiment 1

As was mentioned above, we used both a high reliability DSS condition and a low reliability DSS condition as well as an unsupported manual condition. These conditions were varied between-subjects because we wanted to avoid prior experiences with a highly

(or low) reliable system influence the use of the DSS. We varied time pressure within-subjects—that is, time pressure (i.e., low vs. high) was varied blockwise because it seems likely that in a lot of real-world contexts time pressure also varies from time to time.

We hypothesized that participants in the high reliability DSS condition would have better performance than those in the low reliability condition. The manual condition mainly served as a control condition to understand the effects of time pressure for the luggage-screening task without DSS support. Moreover, based on previous findings (e.g., Rice & Keller, 2009), we hypothesized that time pressure would decrease performance, but only if the automation were not highly reliable. This is based on the idea that operators are more dependent on the DSS under time pressure. Conversely, an increased dependence can possibly even lead to performance increases if the automation is highly reliable, although in this case, operators often show a high level of reliance and compliance, anyway. Thus, in the high reliability DSS condition, participants should have greater performance than in the manual condition, whereas this difference is not as clear a priori for the low reliability DSS condition. Beyond that, we also investigated whether time pressure would also make a difference for the level of reliance which has not been addressed before.

Method

This research complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board at the Technische Universität Berlin, Department of Psychology. Informed consent was obtained from each participant.

Participants. 60 participants (24 female) were recruited through the online recruiting system of the Department of Psychology at Technische Universität Berlin. Participants took part in the experiment for either course credits or monetary compensation of 9€. The mean age of participants was 28.62 (*SD*: 4.75), and the sample was predominantly right-handed (56 right-handed). Two additional participants in the manual condition were tested but excluded from any analyses due to accuracy not

deviating from chance in the experimental blocks.

Apparatus and Stimuli. We used images from the X-Ray Object Recognition Tests (X-Ray ORT) 1.3 and 2.0 (Hardmeier, Hofer, & Schwaninger, 2005; Schwaninger, Hardmeier, & Hofer, 2005) as stimuli. These stimuli can be classified as hard or easy based on three dimensions: point of view (i.e., canonical or not), bag complexity (i.e., high or low clutter in the bag), and object overlap (i.e., whether there is little or strong overlap of other objects on the target). To ensure similar difficulty for the stimuli—especially considering the time constraints for responding—we selected images with a similar difficulty according to these dimensions, using the images which were scored as difficult in any two of the three categories. These images could then either contain a target (i.e., gun or knife) or not, and we mirrored the images both horizontally and vertically to double the number of available stimuli, resulting in a total stimulus set of 320 images.

One to six participants were simultaneously tested at independent PC workstations, separated by opaque screens. The experiment was run with E-Prime 3.0. We used 24 inch screens with a 1920x1200 resolution, and the stimuli presented were sized 700x550 pixels, centered vertically and horizontally on screen. Responses were made using the 'q' and 'p' keys on a German standard keyboard, with the left and right index fingers, respectively, and it was counterbalanced across participants which response side was associated with a target present/absent response.

Procedure. Participants were randomly assigned to one of the three between-subjects conditions: One third of the participants was tested in the manual condition, one third in the high reliability (i.e., 95%) automation condition, and one third in the low reliability (i.e., 75%) condition. The experiment consisted of two training blocks and six experimental blocks, with 40 trials each. In every block, half of the trials were target present trials, and half of the trials were target absent trials. The full experiment took around 60 minutes.

At the beginning of the experiment, participants were shown four overview images

with all target stimuli (i.e., two gun and two knife overview images) for 10 seconds to familiarize participants with the target items. All participants had the same first training block, with easier stimuli than in the remainder of the experiment (i.e., stimuli with only one of the three dimensions scored as difficult). Because the first training block was mainly used to familiarize participants with the basic screening task, no automation support was given in this block. Participants were instructed to respond as fast and as accurately as possible.

In the automation conditions, the second training block introduced the automated DSS, which was visualized by showing two circles with one of them being filled with either red or green color, to indicate a dangerous item or to indicate that no target was present, prior to stimulus onset. The two automation conditions differed in terms of their false alarm rate, but not in their hit rate, and the differences between the automation conditions are shown in Table 1. In addition to the false alarms, one miss each appeared in the last and penultimate block of both conditions. Misses were included only toward the end of the blocks in order to avoid any strong performance consequences of first-failure effects of misses skewing the whole experiment (Wickens & Xu, 2002). We made sure that there were no automation failures of any kind in consecutive trials. Moreover, there were no cases of more than three consecutive trials requiring the same response and also no cases of the same image appearing on consecutive trials. All of these constraints were embedded in a sequential trial procedure.

After the DSS was introduced in the second block, participants filled out two questionnaires. First, they were asked to fill out a 2x2 contingency table (i.e., Hits, FAs, Misses, CRs) on how they perceived the automation in the past 40 trials. Second, they were asked to fill a questionnaire on trust in technical systems (Wiczorek, 2011). Subsequent to filling out the questionnaires, participants were shown the true 2x2 contingency table for 100 experimental trials, to bridge the description-experience gap (e.g., Hertwig & Erev, 2009).

Condition	Reliability	Hit Rate	FA Rate	d'	C
High Reliability	95%	98.33%	8.33%	3.40	-0.34
Low Reliability	75%	98.33%	48.33%	2.08	-1.00

Table 1

Key characteristics of the automated decision support systems in the experimental blocks. Note that the signal detection theory (SDT) measures are loglinear-corrected (Hautus, 1995) because some cells of participant x condition produced perfect hit rates, requiring this correction for the empirical data. Thus, we have corrected d' and the criterion C in this table analogously. FA: false alarm. d': detection performance in SDT, C: criterion to measure response bias.

After the first two training blocks, the time pressure manipulation was introduced and the six experimental blocks started. That is, in half of the blocks, participants had 4.5 seconds to make their response, and 9 seconds in the other half. These cutoffs were chosen based on a pilot study to allow participants to still process the stimulus but to set them under moderate time pressure. Time pressure alternated blockwise and it was counterbalanced across participants whether time pressure was high in odd/even blocks.

The trial procedure was as follows. In the manual condition, trials started with a fixation cross for 2000 ms. In the automated DSS conditions, trials started with the automation advice by presenting either a red (target present) or green (target absent) circle as the automation advice for 1500 ms, followed by a 500 ms fixation cross. The side where red/green circles were presented corresponding with the response key assignment. Then the stimulus remained on screen for a maximum of 4.5/9 seconds, depending on the time pressure condition in the respective block, or until a response was made. This was indicated by a countdown on the upper left above the image. Each trial was followed by a 500 ms inter-trial-interval with a blank screen. In the two practice blocks, instead of the

inter-trial-interval, participants received feedback on their response after every trial for 1000 ms. The trial procedure for the automation conditions' experimental blocks is visualized in Figure 1.

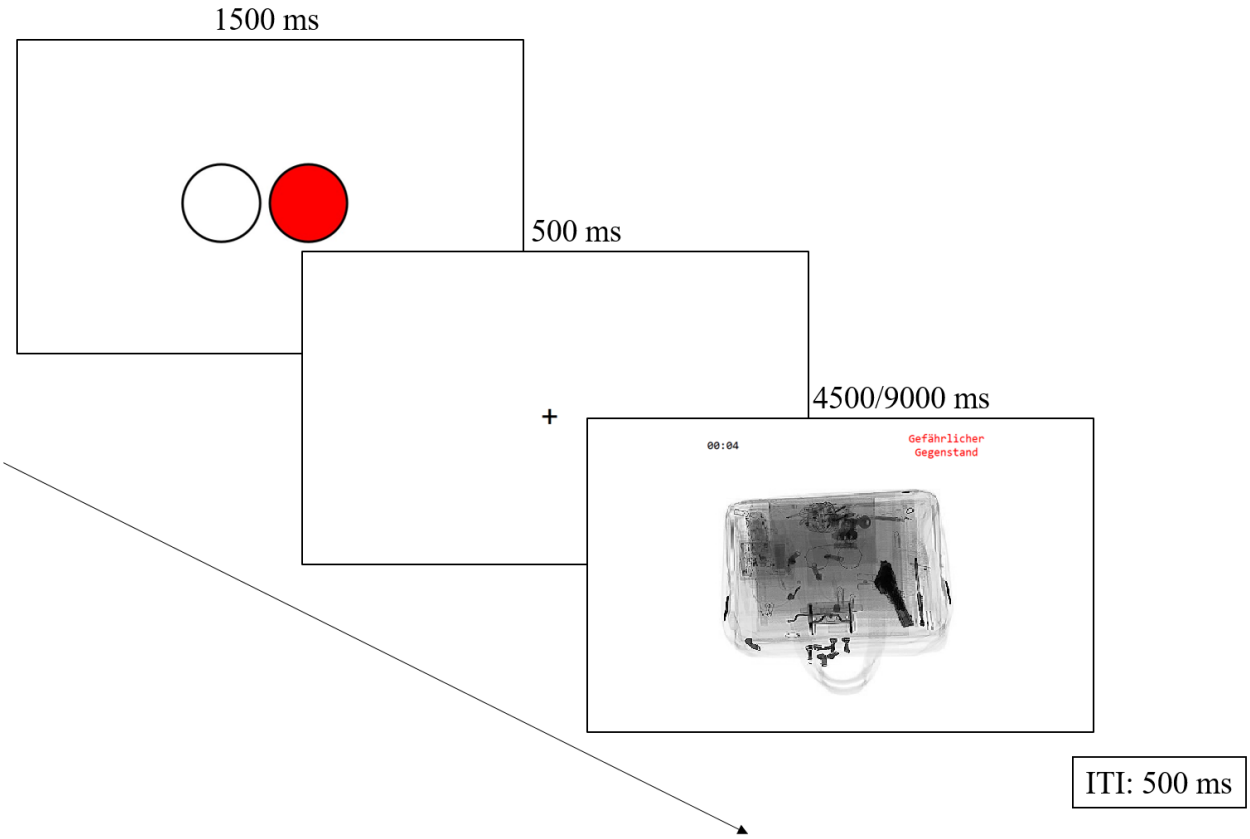

Figure 1. Typical trial sequence in the automation condition of Experiment 1. ITI: inter-trial-interval (blank screen).

Design. Automation condition was varied between subjects (i.e., high reliability, low reliability, manual), and time pressure (i.e., high, low) alternated blockwise within subjects. Thus, the present study used a 3 (automation condition) x 2 (time pressure) repeated-measures mixed design.

Results

Performance. We conducted a 3 (automation condition) x 2 (time pressure) ANOVA on the percentage of correct responses (PC) as our primary analysis. The

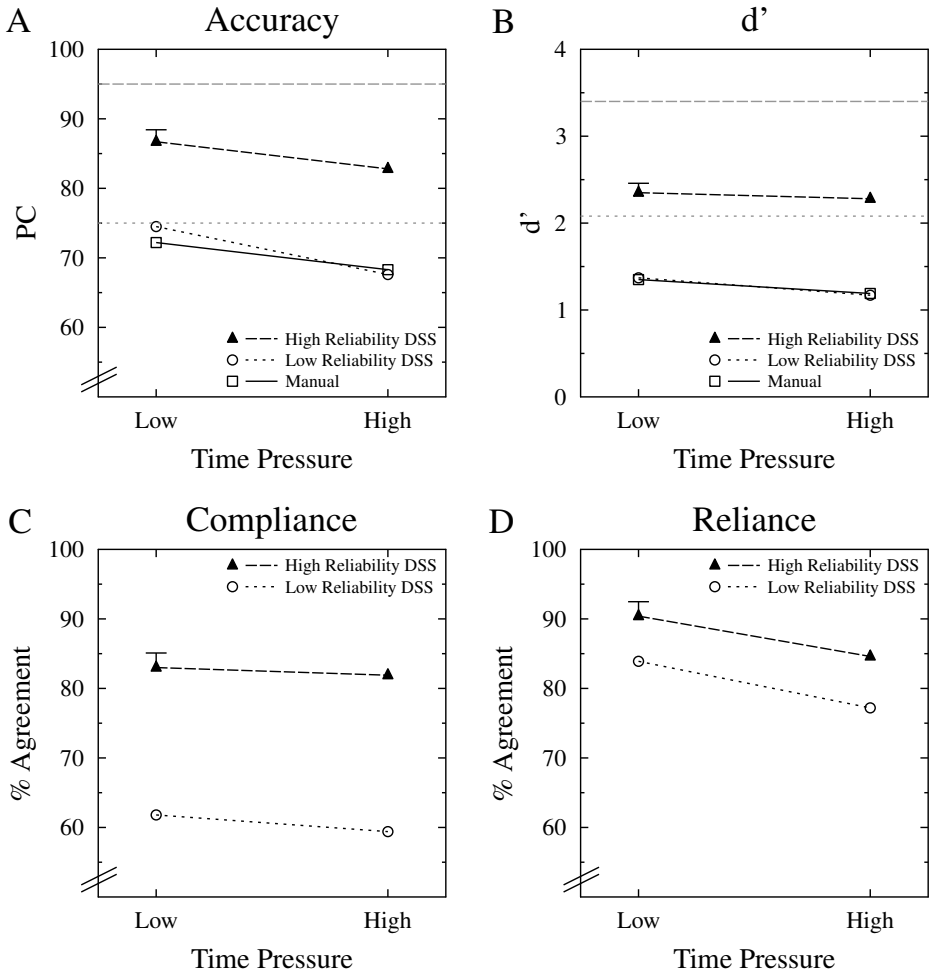

Figure 2. A: Percent correct (PC), B: signal detection theory's d' (loglinear corrected), C: compliance, and D: reliance as a function of time pressure, separately for each condition. In subplots A and B, the dotted horizontal lines represent the values (PC and d', respectively) for the performance of the DSS alone. Error bar represents the pooled standard error. DSS = decision support system.

alpha-level for this and all subsequent analyses was set to the conventional .05 level. The corresponding means of this ANOVA are displayed in Figure 2A. This ANOVA revealed a significant main effect of automation condition, $F(2, 57) = 37.774$, $p < .001$, $\eta_p^2 = 0.57$. That is, pairwise comparisons with a Bonferroni-corrected alpha level revealed that responses in the high reliability DSS condition were significantly more accurate (84.7%)

than in the low reliability DSS condition (71.1%, $p < .001$) and than in the manual condition (70.3%, $p < .001$). The difference between the manual and the low reliability DSS condition was non-significant ($p = .678$). Moreover, the main effect of time pressure was also significant, $F(1, 57) = 23.803$, $p < .001$, $\eta_p^2 = 0.295$, indicating a higher PC under low (77.8%) than under high (72.9%) time pressure. Interestingly, the interaction was non-significant, $F(2, 57) = 1.022$, $p = .367$, $\eta_p^2 = 0.035$.

Additionally, in the high reliability automation condition, the average performance of the human-automation-dyad was significantly worse than the automation alone, even without time pressure (86.7%), $t(19) = 7.454$, $p < .001$, clearly showing a sub-optimal DSS use. In the low reliability automation condition, joint performance under low time pressure (74.5%) did not differ from the automation reliability, $t(19) = 0.378$, $p = .710$.

We conducted a parallel analysis on the signal detection measure d' and the corresponding means can be found in Figure 2B. Because some cells of participant x condition produced perfect hit rates, we loglinear corrected all signal detection analyses according to Hautus (1995). Moreover, because it is not that clear for target absent trials with no response whether this trial should count as a correct rejection or false alarm, we excluded these trials from the SDT analyses. Obviously, target present trials with no response were counted as a miss. This ANOVA again revealed a significant main effect of automation condition, $F(2, 57) = 53.692$, $p < .001$, $\eta_p^2 = 0.653$. Pairwise comparisons with a Bonferroni-corrected alpha level revealed significant differences between the high reliability DSS (2.31) and the low reliability DSS (1.27, $p < .001$) as well as the manual (1.27, $p < .001$) condition. The difference between the low reliability DSS condition and the manual condition was non-significant ($p = .974$). Moreover, there was again a significant main effect of time pressure, $F(1, 57) = 5.212$, $p = .026$, $\eta_p^2 = 0.084$, with a higher sensitivity under low (1.69) than under high (1.55) time pressure. Again, the interaction was non-significant, $F(2, 57) = 0.389$, $p = .680$, $\eta_p^2 = 0.013$.

Compliance and Reliance. The finding that time pressure led to impairments of performance independent of automation reliability was surprising. We analyzed the effects of time pressure separately for compliance and reliance to better understand to what extent the effects on overall performance were related to time pressure induced changes in compliance and reliance in the two automation support conditions. We defined compliance as the agreement rate with the automation when the automation indicated that there was a target present, and reliance as the agreement rate with the automation when the automation indicated that there was no target present.

We conducted a 2 (automation condition) x 2 (time pressure) ANOVA on compliance. The corresponding means can be found in Figure 2C. This ANOVA revealed a significant main effect of automation condition, $F(1, 38) = 48.526$, $p < .001$, $\eta_p^2 = 0.561$, with higher compliance in the high reliability (82.4%) than in the low reliability (60.6%) condition. Neither the main effect of time pressure was significant, $F(1, 38) = 1.334$, $p = .255$, $\eta_p^2 = 0.034$, nor the interaction of condition and time pressure, $F(1, 38) = 0.174$, $p = .679$, $\eta_p^2 = 0.005$.

Results of a parallel analysis for reliance are visualized in Figure 2D. The ANOVA revealed a main effect of condition, $F(1, 38) = 5.619$, $p = .023$, $\eta_p^2 = 0.129$, with higher reliance in the high reliability (87.5%) than in the low reliability (80.5%) condition. Moreover, interestingly, the main effect of time pressure was significant, $F(1, 38) = 17.947$, $p < .001$, $\eta_p^2 = 0.321$, with less reliance under high (80.9%) than under low (87.1%) time pressure. The interaction was non-significant, $F(1, 38) = 0.117$, $p = .734$, $\eta_p^2 = 0.003$.

Response Times. Finally, we also analyzed response times (RTs) as a dependent measure. We excluded incorrect responses from the RT analyses (24.7%), and after visual inspection we also excluded RTs shorter than 500 ms (0.2% of the remaining trials). Because a luggage screening task is essentially a visual search task with a self-terminating search, with usually longer RTs for target absent trials, we added target present/absent as an additional factor in the ANOVA, resulting in a 3(condition) x 2(time pressure) x

2(target present/absent) ANOVA with repeated measures for the last two factors, and the results are displayed in Figure 3A-C. Unsurprisingly, RTs were faster in target present (2593 ms) than in target absent (4023) trials, $F(1, 57) = 392.93$, $p < .001$, $\eta_p^2 = 0.873$. Moreover, responses were faster under high (2583 ms) than under low (4043 ms) time pressure, $F(1, 57) = 360.72$, $p < .001$, $\eta_p^2 = 0.864$. The main effect of condition was non-significant, $F(2, 57) = 3.0525$, $p = .055$, $\eta_p^2 = 0.097$, with unreliably smaller RTs in the high reliability DSS (3086 ms) than in the low reliability DSS (3390 ms) and the manual (3448 ms) conditions. Interestingly, the interaction of target presence and time pressure was significant, $F(1, 57) = 126.43$, $p < .001$, $\eta_p^2 = 0.689$, with a much larger effect of time pressure in target absent (Δ: 1946 ms) than in target present (Δ: 913 ms) trials. The interaction of target presence and condition was also significant, $F(2, 57) = 10.264$, $p < .001$, $\eta_p^2 = 0.265$. Here, the effect of target presence was largest in the manual condition (1603 ms), and smaller in the low reliability DSS (1482 ms) and high reliability DSS (1269 ms) conditions. The three-way interaction was also significant, $F(2, 57) = 5.2916$, $p = .008$, $\eta_p^2 = 0.157$. That is, as becomes evident from Figure 3 the difference between target present/absent trials under high time pressure was particularly small in the high reliability DSS condition (Δ: 671 ms) compared to the manual (Δ: 1088 ms) and low reliability DSS (Δ: 983 ms) conditions. The interaction of time pressure and condition was non-significant ($p = .203$).

In an exploratory manner, we extended our RT analyses to distributional analyses, using the deciles of the RT distribution to better understand the effects of time pressure. To that end, we calculated the deciles for each participant, separately for each time pressure condition, and averaged these deciles for each condition across participants. These deciles are visualized in Figure 3D-F. As is evident from the figure, participants in the high time pressure condition usually responded faster than they had to—as there is already a difference in the 20th percentile of trials.

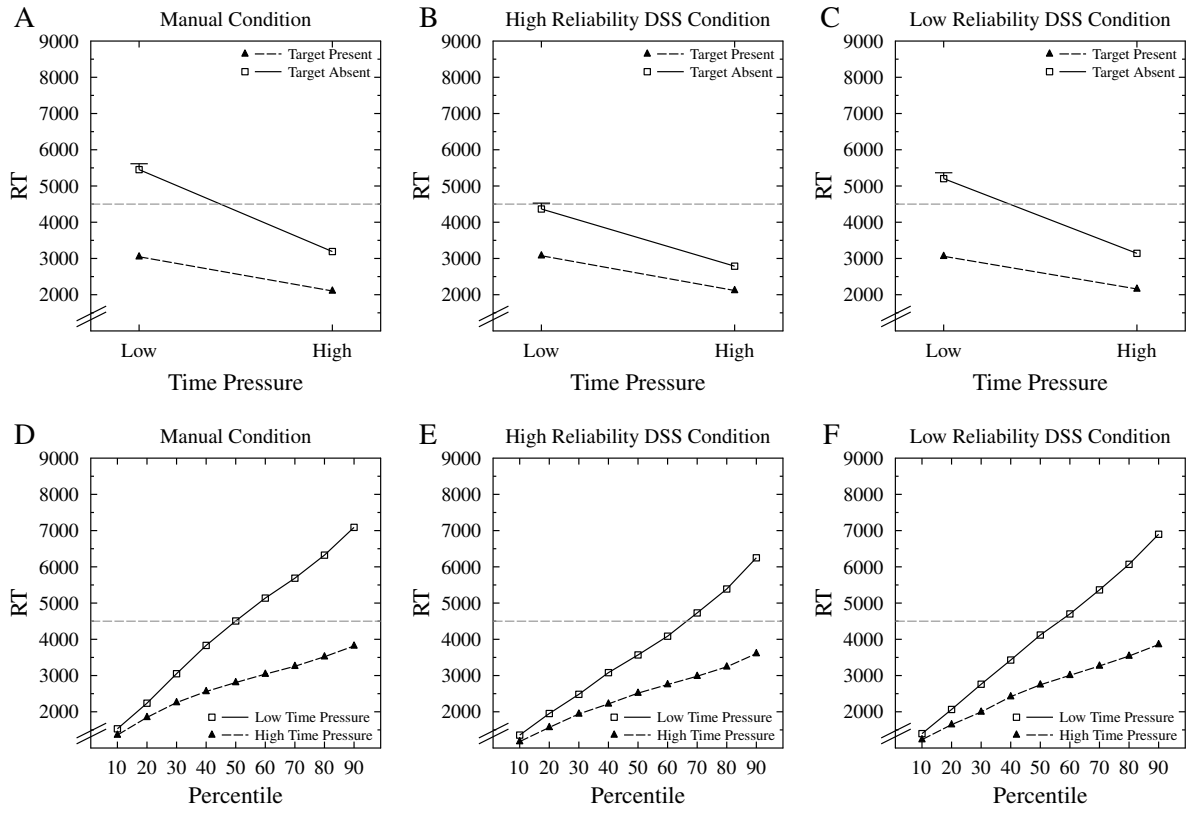

Figure 3. Response time (RT) data separately for each condition of Experiment 1. Panels A-C display the mean RTs separately for target present and absent trials, as a function of time pressure. Panels D-F display the deciles of the RT distribution separately for low and high time pressure trials. Error bar in the upper panels represents the pooled standard error. DSS = decision support system.

Questionnaire Data. To check how participants perceived the automation, we also analyzed subjective data. That is, we analyzed questionnaire on trust in technical systems (Wiczorek, 2011, scale from 1-4). Participants in the high reliability (3.20) group had higher trust on the trust in technical systems questionnaire than participants in the low reliability (2.89) group, $t(38) = 2.807$, $p = .008$, $d = 0.88$. In the contingency table, participants on average estimated the high reliability DSS to be 89.3% accurate, and the low reliability DSS to be 72.4% accurate. Because participants were informed about the true reliability of their DSS after filling out the table, we are confident that participants

generally had a realistic perception of the DSS's reliability.

Discussion

The key findings of Experiment 1 are that (a) time pressure led to worse overall performance, (b) this negative effect of time pressure was not attenuated (or even led to positive effects) when a highly reliable automation was available, and (c) that performance increased with a highly reliable automation but was still worse than the automation alone. Moreover, participants' overall compliance did not differ between the high and low time pressure blocks but interestingly, participants were more reliant on the automation under low than under high time pressure. The overall performance decreases thus seem to be connected to a decreased reliance under high time pressure, in both automation conditions. Besides the general performance consequences, in the RT data, we found that participants sped up their responses a lot more than necessary, and could have taken more time in a lot of high time pressure trials to inform their decision.

The finding of a general negative effect of time pressure, even with a highly reliable DSS available, clearly contrasts some earlier findings (Rice & Keller, 2009; Rice et al., 2010; Rice & Trafimow, 2012). In this earlier research, time pressure was found to increase the dependence on the automation which clearly improved overall performance in cases of a highly reliable system. The results of the present experiment suggest that these earlier findings could have mainly been due to the extreme level of time pressure used in that study (only two seconds to inspect complex aerial photographs) which probably left participants no other choice than to follow the advice of the automation. In contrast, the present experiment did not use a time pressure manipulation that extreme, leaving the participants the possibility to also at least to some degree manually inspect the visual stimuli. In this experiment, it is quite clear that participants made use of this possibility which is most clearly reflected in the reliance measure which showed that reliance even decreased somewhat with increasing time pressure. Also, the elevated time pressure led

them to speed up their responses more than would have been needed. This suggests that time pressure in our study did not seduce participants to delegate their responsibility to the automation but, instead, induced a sort of self-pressure to inspect the X-ray as quickly as possible. Moreover, the data clearly suggests that in our study participants generally did not use the highly reliable automation adequately, regardless of whether they were under time pressure or not. That is, independent of the level of time pressure, participants would have generally been better served to just follow the DSS advice rather than interfering with the automation. This lack of adequate use of the DSS also reiterates the problem that joint performance of human and automation is often worse than that of the automation alone (Bartlett & McCarley, 2017).

Overall, this pattern of results suggests that the participants were generally reluctant to just follow the automated decision aid. Perhaps, they wanted to make sense of their role as responsible operator by checking and correcting the automation. Intervening with the DSS's recommendations then led to more false than proper corrections, particularly with the high reliability DSS. One reason for this might be because participants always inspected the stimulus after seeing the advice. This could have prompted them in a particular way to become active and not just accepting the automated advice. Thus, we designed Experiment 2 in order to give participants the possibility to first make their own decision and then give them the DSS advice—allowing for the possibility to specifically use the automation in cases of own uncertainty and to still contribute properly to the decision-making process in cases where one is confident in the own decision. One would then assume that under time pressure, participants should feel less sure about their own decisions and depend more on the automation. This setup could then possibly reduce the negative effects of time pressure—or even lead to positive effects if participants realize they should depend on the automation as much as possible under time pressure.

Experiment 2

The second experiment was designed to check whether changing the order of image display (and the concomitant response) and automation advice makes a difference for joint human-automation decision-making. That is, in the automation conditions, participants were always first shown the image and were asked to make their initial response to the stimulus while the stimulus was displayed. Then, they were shown the automation advice and had the opportunity to stay with their choice or to make a change based on the automation's advice. Note that the automated DSSs were the same as in Experiment 1, and the manual condition was exactly the same as in Experiment 1.

Our hypotheses were largely the same as in Experiment 1. We expected an increased dependence on the automated DSS under high time pressure compared to low time pressure. That is, as was argued above, we expected that because participants can now first make their own decision and then get the advice, they can use it more specifically in trials where they are less secure about their own judgment. Thus, they would still feel in the loop (rather than just following a cue's advice) but could still benefit from the automation. Again, we hypothesized that time pressure would decrease performance, but only if the automation is not highly reliable. Conversely, high time pressure could even lead to performance increases with a highly reliable automation.

Method

Participants. A fresh sample of 60 participants (34 female, 1 diverse) was tested in Experiment 2. They ranged in age from 18 to 37 ($M = 26.35$) and were predominantly right-handed (52 right-handed). Participants took part in the experiment for either course credits or monetary compensation of 9€. Two additional participants in the low reliability condition were also tested but excluded from any analyses due to issues with understanding the instructions in one case and the experiment environment crashing in the other case.

Apparatus, Stimuli, Procedure, and Design. No changes were made in the manual condition. In the automation support conditions, the apparatus, stimuli, procedure, and design were the same as in Experiment 1 except for the following changes. First, in the automation conditions, the fixation cross was always present for 2 seconds, regardless of automation condition. Second, participants always first made their target absent/present choice and only subsequently received automation support for their decision. The time pressure manipulation remained the same, that is, it changed blockwise whether participants had 4.5 or 9 seconds to make their initial decision. Then, after seeing the automation decision, they had the opportunity to either confirm their initial judgment by pressing the same key again or to change their decision by pressing the other response key. The stimulus disappeared from the screen contingently after the initial response if one was given or after a maximum of 4.5/9 seconds. Participants were always shown their own response for 500 ms after they made their response, indicated by a green/red circle for target absent/present responses, or an indication that they did not respond in time, respectively. Subsequently, the automation advice appeared on the screen below their own initial decision, also displayed by a green/red circle, for a maximum of 4.5 or 9 seconds in the high time pressure blocks and low time pressure blocks, respectively. Participants were instructed that if they decided to not press any response key when the automation was shown, that this was interpreted as staying with the initial judgment.

Results

Performance. For the automation conditions, we always used the last response given as the final decision to inform the performance measures. As in Experiment 1, we conducted a 3 (automation condition) x 2 (time pressure) ANOVA on PC, and the corresponding means can be found in Figure 4A. This ANOVA again revealed a significant main effect of condition, $F(2, 57) = 28.349$, $p < .001$, $\eta_p^2 = 0.499$. Pairwise comparisons with a Bonferroni-corrected alpha level revealed more accurate responses in the high

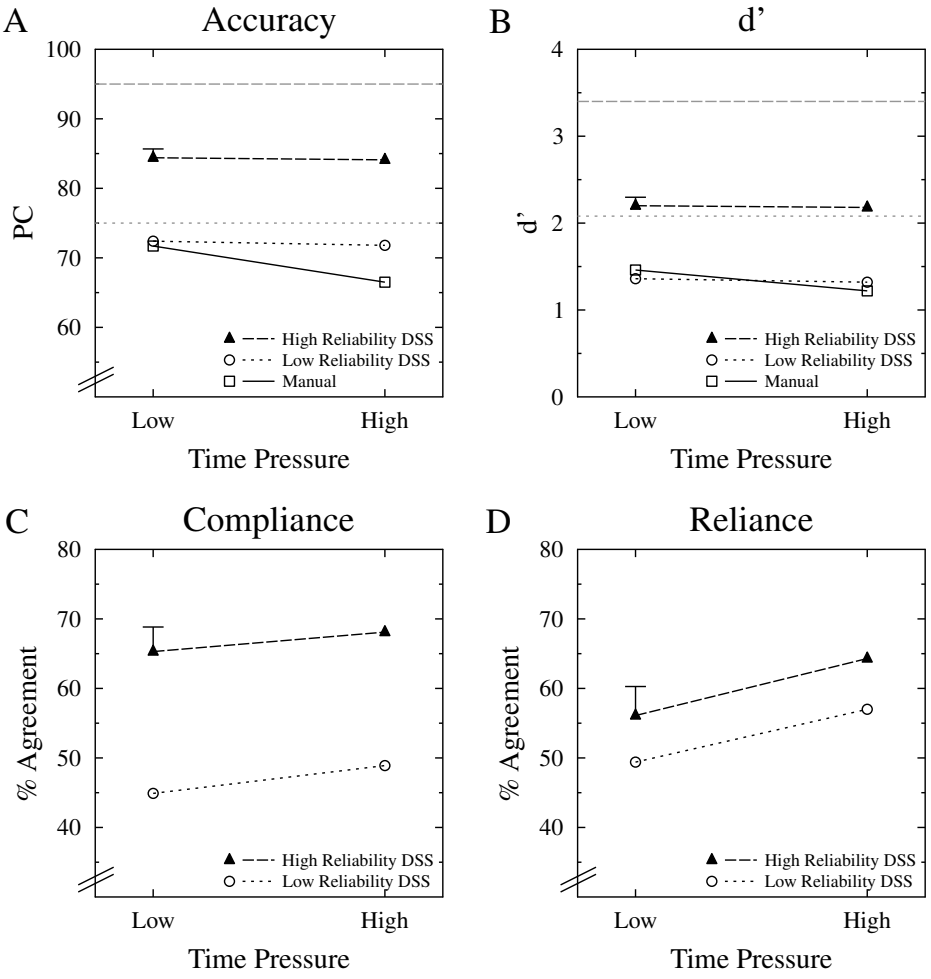

Figure 4. A: Percent correct (PC), B: signal detection theory's d' (loglinear corrected), C: compliance, and D: reliance as a function of time pressure, separately for each condition. In subplots A and B, the dotted horizontal lines represent the values (PC and d', respectively) for the performance of the DSS alone. Error bar represents the pooled standard error. DSS = decision support system.

reliability DSS condition (84.3%) than in the low reliability DSS (72.1%, $p < .001$) and than in the manual (69.1%, $p < .001$) conditions. The difference between the manual and the low reliability DSS condition was non-significant ($p = .083$). There was also a significant main effect of time pressure, $F(1, 57) = 7.583$, $p = .008$, $\eta_p^2 = 0.117$, with less accurate responses under high (74.2%) than under low (76.2%) time pressure. Interestingly,

and contrasting to Experiment 1, there was now a significant interaction of automation condition and time pressure, $F(2, 57) = 4.660$, $p = .013$, $\eta_p^2 = 0.141$. As is evident in Figure 4A, the time pressure effect seemed to vanish in both the high reliability DSS (Δ: 0.3%) and the low reliability DSS (Δ: 0.6%) conditions, but not in the manual condition (Δ: 5.2%).

As in Experiment 1, we additionally checked whether the performance of the human-automation dyad differed from the performance alone under low time pressure. Again, in the high reliability automation condition, the average performance of the human-automation-dyad was significantly worse than the automation alone, even without time pressure (84.4%), $t(19) = 5.249$, $p < .001$. In contrast to Experiment 1, the performance of the dyad in the low reliability condition (72.4%) also differed significantly from the automation alone, $t(19) = 2.328$, $p = .031$.

We again conducted a parallel analysis for d' and the results are displayed in Figure 4B. This ANOVA revealed a significant main effect of automation condition, $F(2, 57) = 18.452$, $p < .001$, $\eta_p^2 = 0.393$. Pairwise comparisons with a Bonferroni-corrected alpha level revealed greater sensitivity in the high reliability DSS condition (2.19) than in both the low reliability DSS (1.34, $p < .001$) and the manual (1.34, $p < .001$) conditions. There was no difference between the manual and the low reliability DSS conditions ($p = .991$). For d', the main effect of time pressure was non-significant, $F(1, 57) = 3.091$, $p = .084$, $\eta_p^2 = 0.051$. The interaction was non-significant, $F(2, 57) = 1.686$, $p = .194$, $\eta_p^2 = 0.056$, however, a very similar visual pattern is evident from Figure 4B.

Compliance and Reliance. Due to the experimental set-up of Experiment 2, we had to take a slightly different approach to measure compliance and reliance. That is, we restricted compliance and reliance analyses to all trials where the initial response was not the same as the automation's suggestion which was shown to participants. Then, we measured compliance and reliance as the proportion of trials where the participant's response equaled the automations' decision after they had seen the automation's advice.

Thus, the agreement rates of both experiments are not directly comparable because the underlying trial base was not the same.

We conducted a 2 (automation condition) x 2 (time pressure) ANOVA for compliance, and the results are visualized in Figure 4C. This ANOVA revealed a significant main effect of automation condition, $F(1, 38) = 8.239$, $p = .007$, $\eta_p^2 = 0.178$, with higher compliance in the high reliability DSS condition (66.7%) than in the low reliability DSS condition (46.9%). The main effect of time pressure was non-significant, $F(1, 38) = 1.866$, $p = .180$, $\eta_p^2 = 0.047$. The interaction was also non-significant, $F(1, 38) = 0.063$, $p = .803$, $\eta_p^2 = 0.002$.

A parallel ANOVA was conducted for reliance with the results shown in Figure 4D. For reliance, there was no main effect of automation condition, $F(1, 38) = 1.030$, $p = .317$, $\eta_p^2 = 0.026$. Interestingly, there was a main effect of time pressure, $F(1, 38) = 7.087$, $p = .011$, $\eta_p^2 = 0.157$, with higher reliance under high (60.6%) than under low (52.8%) time pressure. The interaction was non-significant, $F(1, 38) = 0.0096$, $p = .923$, $\eta_p^2 = 0$.

Response Times. As in Experiment 1, we also analyzed RTs. We used only the RTs during the initial target display, before the automation was shown because that most likely represents the actual target search. Again, we restricted our analyses to correct RTs only (exclusion of 34.4% of all trials) with a minimum RT of 500 ms (0.2% of all remaining trials). We again included target presence as an additional factor in the ANOVA. The corresponding means of this ANOVA are displayed in Figure 5A-C. This ANOVA revealed faster RTs in target present (2252 ms) than in target absent (3597 ms) trials, $F(1, 57) = 250.84$, $p < .001$, $\eta_p^2 = 0.815$. Moreover, responses were faster under high (2472 ms) than under low (3377 ms) time pressure, $F(1, 57) = 84.218$, $p < .001$, $\eta_p^2 = 0.596$. As in Experiment 1, the interaction of target presence and time pressure was again significant, $F(1, 57) = 74.539$, $p < .001$, $\eta_p^2 = 0.567$, with a larger effect of time pressure in target absent (547 ms) than in target present (1265 ms) trials. No effect including condition was significant (p-values $> .427$) which makes sense because the initial decision in Experiment 2 was basically a manual decision.

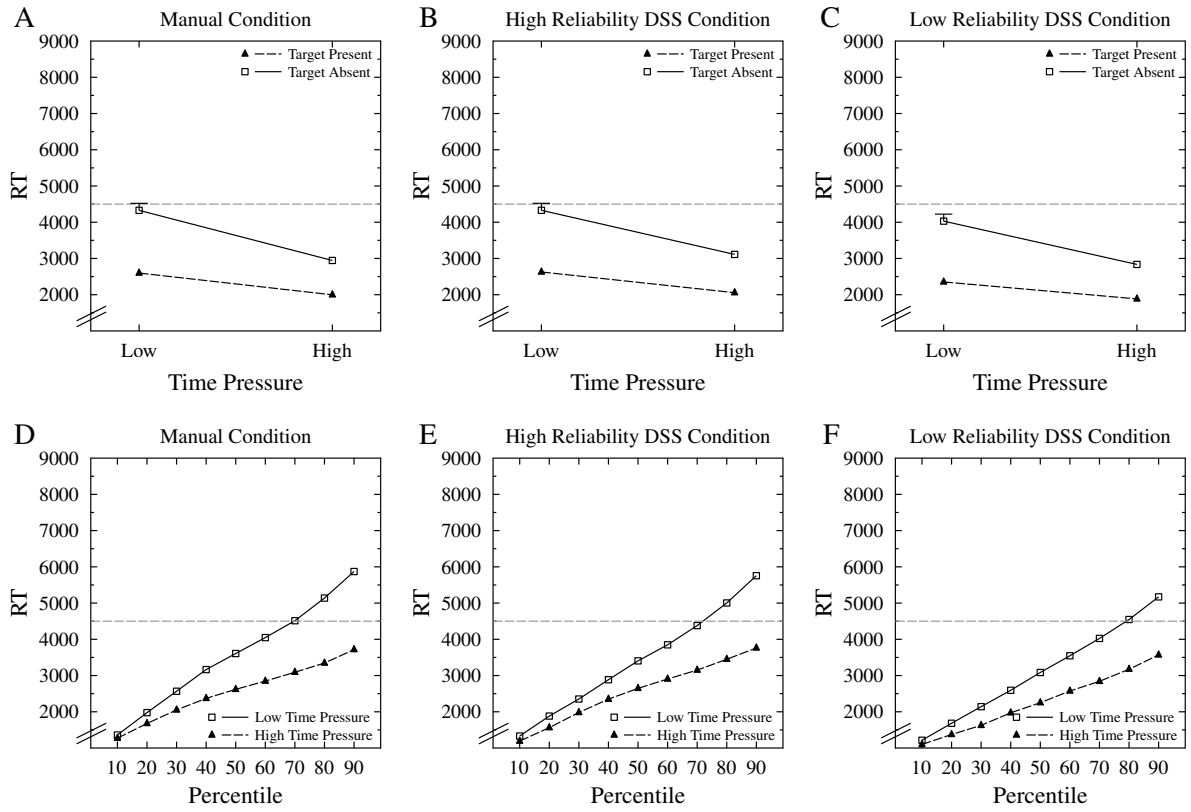

Figure 5. Response time (RT) data separately for each condition of Experiment 2. Panels A-C display the mean RTs separately for target present and absent trials, as a function of time pressure. Panels D-F display the deciles of the RT distribution separately for low and high time pressure trials. Error bar in the upper panels represents the pooled standard error. DSS = decision support system.

We again conducted an exploratory analysis on the RT distribution which is visualized in Figure 5D-F. As in Experiment 1, it seems like participants responded much faster than necessary under time pressure, again with differences in the faster part of the RT distribution. Thus, without the real need to speed up the fastest responses, it seems like participants still did so under time pressure.

Questionnaire Data. In Experiment 2, there was only a slight descriptive trend in the subjective data, with unreliably higher trust in the high reliability DSS (3.09) than in the low reliability DSS (2.88) on the scale of Wiczorek (2011), $t(38) = 1.656$, $p = .106$, $d =$

0.523. In the contingency table, participants on average estimated the high reliability DSS to be 90.8% accurate, and the low reliability DSS to be 77.6% accurate, again close to the actual reliabilities.

Discussion

To summarize the key finding of Experiment 2, the initial-decision then automation-advice setup apparently reduced the negative impact of time pressure on performance to little or no effect, as indicated by the PC data. Moreover, we also found interesting results for compliance and reliance, with an increase for reliance under high time pressure and higher compliance in the high reliability DSS condition. Thus, it seems like the experimental setup of Experiment 2, with participants receiving the automation advice only after having made their own decision, made quite a difference for operator reliance under time pressure. That is, in this scenario, participants reliance was greater under high than under low time pressure, probably also playing a role in the significant interaction found for PC. It also makes sense that there was only a significant effect of condition on compliance, as the alarm systems only differed in terms of their false alarm rate. The results of Experiment 2 thus seem to confirm the hypothesis that giving participants the DSS's advice only after they had made their initial choice could reduce negative effects of time pressure by increasing the use of the DSS in those high time pressure trials, particularly for reliance. The present findings align well with the findings by Ho, Pavlovic, Myers, and Arrabito (2013) who found that giving participants the possibility to only use the automation when they felt it was useful increased overall performance compared to higher levels of automation that always gave recommendations.

General Discussion

The goal of the present research was to investigate performance consequences of time pressure with different automated DSSs (or none) available. To this end, we used a luggage-screening task and had participants carry out this task, under both low and high

time pressure. Moreover, participants were randomly assigned to one of three conditions using either no automated DSS (manual), a highly reliable DSS (95% reliability) or a low reliable DSS (75% reliability). In Experiment 1, participants in the automation conditions were first shown a cue and then the stimulus, and in Experiment 2, we reversed that order and gave participants the possibility to change their initial decision based on the DSS's advice. The main findings were that time pressure largely led to negative effects on performance—however, reversing the order of decision-making in Experiment 2 strongly reduced these negative effects. Moreover, regarding the dependence on the automation, our results were rather mixed—with less reliance on the DSS under high time pressure in Experiment 1 and an increased reliance under high time pressure in Experiment 2. Compliance was largely unaffected by introducing time pressure. Perhaps most interestingly, in both experiments and in all four automation conditions, the joint mean performance (in both PC and d') of the human-automation dyad was descriptively worse than the automation alone.

As was mentioned above, contrary to some earlier findings (e.g., Rice & Keller, 2009), we did not find time pressure induced benefits for performance even with a highly reliable DSS. Thus, it seems as if time pressure—at least in the present study—did not lead to more heuristic or optimized decision-making, as was suggested by earlier studies (e.g., Payne et al., 1988; Rice & Keller, 2009). At best, our results indicate that having a DSS available after having made the initial own choice can reduce the negative effects of time pressure, as was shown in Experiment 2. That is, changing the order of decision-making reduced the negative effects of time pressure, and increased reliance on the DSS, providing some evidence for an increased automation dependence under high time pressure. Thus, it seems like participants used the DSS more selectively in those trials where they were less secure about their own decisions—and it makes sense that those trials were mostly time-pressured trials. However, there were still no positive effects of time pressure as was reported by earlier studies (e.g., Rice & Keller, 2009). From a practical standpoint,

changing the order of decision-making could be recommendable in contexts where time pressure cannot be avoided.

The present findings also reinforce earlier concerns (e.g., Bartlett & McCarley, 2017; Meyer, 2001; Meyer et al., 2014) about the joint performance of a human-automation dyad. It seems clear that participants did not use the automation adequately, regardless of whether they were under time pressure or not. That is, if the automation would not have been interfered with by a human, the overall performance would have been higher, especially with the high reliability DSS. Even under time pressure there was no more adequate automation use, despite the fact that time pressure decreased manual performance—and one would assume that decreased personal performance capability would lead to stronger dependence on the automation (Rice & Keller, 2009).

The combined findings of a negative impact of time pressure on performance, and the fact that the automation alone mostly showed greater performance than the human-automation dyad, raises the question of whether a higher level of automation (e.g., Kaber, 2018; Sheridan & Verplank, 1978) might be adequate under high time pressure, as previously suggested by Moray, Inagaki, and Itoh (2000). That is, Moray et al. (2000) argued that under high time pressure, one can truly benefit from automation, and particularly if the automation keeps the human out of the loop. This proposal seems particularly promising considering the fact that the joint performance was below that of the automation in the present experiments. Moreover, Johnson, Ren, Kuchar, and Oman (2002) also suggested that "subjects were reticent to deviate from highly automated ... suggestions even when significant improvements were still possible" (p.132)—thus, keeping the human out of the loop might be best for overall performance. Obviously, in a lot of work settings, it is not possible to remove the human from the loop, be it for legal reasons or availability in emergency situations—and the present results implicate that time pressure should be avoided in such situations.

Besides the main performance consequences of time pressure and automation, we also

analyzed some additional measures, with some interesting implications. Specifically, in both experiments, participants under high time pressure took less time than they could have had to make their decision. This not only shows that the manipulation was successful, but also that in real-world contexts, one could try to use recommendations to workers that they should always take the time they have at hand to complete a task, avoiding rushed decisions.

No study comes without limitations and the present experiments are no exception to that. First, the present study used quite a high base rate (50%). Obviously, the base rate at airport security checkpoints is much lower, but this gives nonetheless even more importance to the low reliability DSS condition with a high false alarm rate. That is, it is necessarily true that with decreasing base rates and a constant automation reliability, more false alarms are produced (Parasuraman & Riley, 1997). Nevertheless, future research might investigate the influence of base rate. Second, the DSS we used was of rather simple nature, and evidence has been provided (Chavaillaz et al., 2018) that direct cues (i.e., cues marking the target) can improve performance compared to a simple cue (such as the one we used). Third, we must acknowledge that giving participants a *second chance* for their decision in the automation conditions in Experiment 2 might have decreased time pressure, especially because participants responded rather quickly after seeing the DSS's advice, as this was a rather simple agree/disagree statement. Note also that even though the negative effect of time pressure was ameliorated in Experiment 2, the overall performance in the automation conditions did not really differ between Experiment 1 and 2.

One question which our present research cannot address is what role self-confidence with the task plays for automation dependence, and future research should investigate this. Moreover, establishing a proper mental model of the automations' capabilities might also change automation use and should be investigated in future research.

To conclude, we argue that time pressure should in fact be avoided—especially in safety-critical environments such as security checkpoints at airports. Moreover, giving

automation advice after the initial choice might make it possible to reduce such negative effects of high time pressure. However, our findings also reinforce earlier concerns whether to keep the human in the loop at all if operators worsen the overall performance compared to the automation alone (e.g., Bartlett & McCarley, 2017; Meyer, 2001). Thus, it seems fair to conclude that performance was not optimized ideally in the automation conditions and the human-automation dyad was worse than the high reliability automated DSS alone would have been—and considerably worse than an ideal human-automation partnership would possibly allow for (Sorkin & Woods, 1985).

Key Points

- Time pressure largely leads to negative effects on performance
- Joint human-automation performance falls below automation-alone performance with a highly reliable system, even under high time pressure
- Presenting the automation advice after participants have made their initial choice reduced negative effects of time pressure but joint performance was still worse than the isolated automation performance

References

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (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

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human Factors*, *59*(6), 881–900. doi: 10.1177/0018720817700258

Buser, D., Sterchi, Y., & Schwaninger, A. (2019). Effects of time on task, breaks, and target prevalence on screener performance in an X-ray image inspection task. In *2019 International Carnahan Conference on Security Technology (ICCST).* IEEE. doi: 10.1109/ccst.2019.8888408

Carayon, P., & Gurses, A. P. (2008). Nursing workload and patient safety—a human factors engineering perspective. In *Patient safety and quality: An evidence-based handbook for nurses.* Agency for Healthcare Research and Quality (US).

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. *Ergonomics*, 1–14. doi: 10.1080/00140139.2018.1481231

French, B., Duenser, A., & Heathcote, A. (2018). *Trust in automation* (Tech. Rep. No. CSIRO Report EP184082). CSIRO, Australia.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. doi: 10.1146/annurev-psych-120709-145346

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.*

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 39th Annual 2005 International Carnahan Conference on Security Technology.* IEEE. doi: 10.1109/ccst.2005.1594876

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. doi: 10.3758/bf03203619

Hendy, K. C., Liao, J., & Milgram, P. (1997). Combining time and intensity effects in assessing operator information-processing load. *Human Factors*, *39*(1), 30–47. doi: 10.1518/001872097778940597

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. doi: 10.1016/j.tics.2009.09.004

Ho, G., Pavlovic, N., Myers, V., & Arrabito, R. (2013). Reducing false alarms in automated target recognition by lowering the level of automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 374–378. doi: 10.1177/1541931213571081

Hoff, K. A., & Bashir, M. (2015). Trust in automation. *Human Factors*, *57*(3), 407–434. doi: 10.1177/0018720814547570

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. doi: 10.1136/svn-2017-000101

Johnson, K., Ren, L., Kuchar, J., & Oman, C. (2002). Interaction of automation and time pressure in a route replanning task. In *International Conference on Human-Computer Interaction in Aeronautics* (pp. 132–137).

Kaber, D. B. (2018). Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 7–24. doi: 10.1177/1555343417737203

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., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors*, *56*(5), 840–849. doi: 10.1177/0018720813512865

Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors*, *33*(6), 607–629. doi: 10.1177/001872089103300602

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44–58. doi: 10.1037/1076-898x.6.1.44

Mosier, K. L., & Fischer, U. M. (2010). Judgment and decision making by individuals and teams: Issues, models, and applications. *Reviews of Human Factors and Ergonomics*, *6*(1), 198–256. doi: 10.1518/155723410x12849346788822

Mosier, K. L., & Manzey, D. (2020). Humans and automated decision aids: A match made in heaven? In M. Mouloua & P. A. Hancock (Eds.), *Human performance in automated and autonomous systems: Current theory and methods* (pp. 19–42). Boca Raton: CRC Press.

Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation bias, accountability, and verification behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*(4), 204–208. doi: 10.1177/154193129604000413

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi: 10.1518/001872097778543886

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *14*(3), 534–552. doi: 10.1037/0278-7393.14.3.534

Rice, S., Hughes, J., McCarley, J. S., & Keller, D. (2008). Automation dependency and performance gains under time pressure. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, pp. 1326–1329). doi: 10.1177/154193120805201905

Rice, S., & Keller, D. (2009). Automation reliance under time pressure. *Cognitive Technology*, *14*(1), 36-44.

Rice, S., Keller, D., Trafimow, D., & Sandry, J. (2010). Retention of a time pressure heuristic in a target identification task. *The Journal of General Psychology*, *137*(3), 239–255. doi: 10.1080/00221309.2010.484447

Rice, S., & Trafimow, D. (2012). Time pressure heuristics can improve performance due to increased consistency. *The Journal of General Psychology*, *139*(4), 273–288. doi: 10.1080/00221309.2012.705187

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. In *38th Annual 2004 International Carnahan Conference on Security Technology, 2004.* IEEE. doi: 10.1109/ccst.2004.1405402

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems Magazine*.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. doi: 10.1037/0033-2909.134.2.207

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1).

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Tech. Rep.). Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction*, *1*, 49-75. doi: 10.1207/s15327051hci0101_2

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi: 10.3758/bf03207704

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. doi: 10.1080/14639220500370105

Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention* (Tech. Rep.). Institute of Aviation Tech. Report AHFD-02-14/MAAD-02-2. Savoy: University of Illinois, Aviation Research Lab.

Wiczorek, R. (2011). Entwicklung und Evaluation eines mehrdimensionalen Fragebogens zur Messung von Vertrauen in technische Systeme. In *Reflexionen und Visionen der Mensch-Maschine-Interaktion–Aus der Vergangenheit lernen, Zukunft gestalten* (Vol. 9, pp. 621–626).

Short Biographies

Tobias Rieger is a researcher and lecturer at the Department of Psychology and Ergonomics, Technische Universtität Berlin, Germany. He earned a master in psychology at the University of Freiburg in 2018 and is currently working on a PhD addressing issues of human performance consequences of automation.

Dietrich Manzey is a university professor of work, engineering and organizational psychology in the Department of Psychology and Ergonomics, Technische Universtität Berlin, Germany. He earned his PhD in experimental psychology at the University of Kiel, Germany, in 1988 and his habilitation in psychology at the University of Marburg, Germany, in 1999.