Tobias Rieger, Dietrich Manzey

Understanding the impact of time pressure and automation support in a visual search task

Understanding the Impact of Time Pressure and Automation Support in a Visual Search Task

Tobias Rieger and Dietrich Manzey

Technische Universität Berlin, Department of Psychology and Ergonomics, Chair of Work, Engineering, and Organizational Psychology

This is a post-peer-review, pre-copyedit version of an article published in Human Factors (accepted June 12th, 2022). The final authenticated version is available online at: https://doi.org/10.1177/00187208221111236

Author Note

Address correspondence to: Tobias Rieger, Technische Universität Berlin, Department of Psychology and Ergonomics, Chair of Work, Engineering, and Organizational Psychology, Marchstraße 12, 10587 Berlin, Germany. email: tobias.rieger@tu-berlin.de The authors would like to thank Yuhong Vanessa Jiang and Li Sha for providing the code to generate the noise images. Moreover, the authors would also like to thank Xintao Xing and Judith Mengel for helpful discussions on study design and thoroughly testing the experiment.

Abstract

Objective: To understand the impact of time pressure and automated decision support systems (DSS) in a simulated medical visual search task.

Background: Time pressure usually impairs manual performance in visual search tasks, but DSS support might neutralize this negative effect. Moreover, understanding the impact of time pressure and DSS support seems relevant for many real-world applications of visual search.

Method: We used a visual search paradigm where participants had to search for target letters in a simulated medical image. Participants performed the task either manually or with support of a highly reliable DSS. Time pressure was varied within-subjects by either a trialwise time-pressure manipulation (Experiment 1) or a blockwise manipulation (Experiment 2). Performance was assessed based on signal detection measures. To further analyze visual search behavior, a mouse-over approach was used.

Results: In both experiments, results showed impaired sensitivity under high compared to low time pressure in the manual condition, but no negative effect of time pressure when working with a highly reliable DSS. Moreover, participants searched less under time pressure and when receiving DSS support, indicating participants followed the automation without thoroughly checking recommendations. However, the human-DSS team's sensitivity was always worse than that of the DSS alone, independent of the strength of time pressure.

Conclusion: Negative effects of time pressure can be ameliorated when receiving support by a DSS, but joint overall performance remains below DSS-alone performance.

Application: Highly reliable DSS seem capable of ameliorating the negative impact of time pressure in complex detection tasks.

Keywords: Human-automation interaction, Visual Search, Decision Making, Trust in Automation, Compliance and Reliance, Stress

Précis: Using a simulated medical visual search task, the present experiments were conducted to study human performance consequences of automated decision support systems (DSS) and time pressure. Time pressure negatively impacted performance, but only when the task was done manually but not with DSS. However, the overall joint performance of human and DSS was consistently worse than that of the DSS alone.

Understanding the Impact of Time Pressure and Automation Support in a Visual Search Task

Automation is making inroads into more and more application domains and one special kind of automation which is widely used is the decision support system (DSS). Specifically, DSS are usually meant to aid a human in performing a certain task and they aim to provide users automatically generated information on the state of the world based on information from the environment (Mosier & Manzey, 2020). Despite the wide variety in underlying complexity and application domains of DSS, the final output to the user is usually a quite straightforward indication of a certain state—be it a beeping sound of a smoke detector or the marking of malignant tissue in a mammogram. However, unfortunately, automation is often inappropriately used, resulting in disuse (the automation is not used enough given its capabilities) or misuse (the automation is used too much given its capabilities) (Parasuraman & Riley, 1997). With highly reliable DSS, particularly disuse is often a problem, resulting in overall human-automation performance often being below that of the automation alone (Bartlett & McCarley, 2017; Boskemper et al., 2021; Meyer & Kuchar, 2021; Rieger & Manzey, 2022). This shows that—even when a system is highly reliable—the use of the system is often not adequate, leading to suboptimal overall performance.

One particular domain in which automation is increasingly finding application is the medical field in general (e.g., Jiang et al., 2017; Luz et al., 2016) and medical image evaluation such as radiology in particular (e.g., Alberdi et al., 2004; Drew et al., 2012), where even novel AI-based technologies have been introduced recently (Bejnordi et al., 2017; McKinney et al., 2020). Medical image evaluation is basically an applied visual search task, be it in deciding whether a mammogram contains a malignant mass or in checking a sonography for an aortic perforation. In visual search, the goal is to differentiate signals (e.g., a tumor) from noise (e.g., non-malignant tissue) and the sensitivity of an operator or a system to discriminate between these two states can be described in terms of

signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2004). Specifically, in SDT, performance can formally be described using the sensitivity (d', the ability to discriminate between signal and noise) and the response criterion (C, the tendency to respond either "signal" or "noise"), with a more liberal criterion ($C < 0$) when one tends to respond "signal" and a more conservative criterion ($C > 0$) when one tends to respond "noise". These two formal measures can be calculated from hit and false alarm rates (see Stanislaw and Todorov, 1999, for calculation of the two measures).

In any field, DSS are usually introduced because it can presumably make work safer and increase overall performance (French et al., 2018; Mosier & Manzey, 2020; Sheridan & Parasuraman, 2005). In the specific case of medical imaging, the idea is that DSS can help identify potential abnormalities and therefore support radiologists or other medical professionals (Dorrius et al., 2011; Drew et al., 2012). Specifically, automated DSS can aid distinguishing between signals and noise in this context. Note that in safety-critical approaches to system design, DSS tend to be designed in a way that it tends to produce false alarms (i.e., liberal criterion) in order to avoid misses of critical signals as well as possible. However, given technological advancements, the sensitivity of DSS keeps improving. In the specific field of radiology, for instance, some algorithms have been designed which outperform radiologists in their typical workflow (Bejnordi et al., 2017) Nevertheless, primarily for legal reasons, humans are often the final decision-maker and ultimately responsible, even in scenarios where the automation alone is better than the human alone.

DSS support in an applied visual search task is also interesting from a theoretical perspective. Specifically, serial, self-terminating (SST) models of visual search (e.g., Bricolo et al., 2002; Pashler, 1987; Snodgrass, 1972; Treisman & Gelade, 1980; van Zandt & Townsend, 1993) assume that visual search proceeds serially (i.e., one search item at a time) and terminates as soon as an item has been identified as a target (i.e., self-termination). Thus, target present decisions are usually based on a shorter search than

target absent decisions. When being supported by an automated DSS, based on SST models, one would expect the difference in search amount between target present and absent to be decreased because the advice that no target is present helps inform an "absent" decision with a shorter search being done (Rieger et al., 2021).

In previous research, it has been shown that automated DSS can aid performance in applied visual search tasks such as luggage screening (e.g., Huegli et al., 2020; Madhavan & Wiegmann, 2007; Rieger & Manzey, 2022) but only if the performance of the automation alone is greater than that of the operator working manually (Rieger et al., 2021). However, as was mentioned above, even in this case the human usually remains in charge of the final judgment and decision-making, and, as a consequence, the use (i.e., when to rely on and to comply with a DSS, see Meyer, 2001) of automated systems is often less-than-ideal. Specifically, the use of the information provided by the DSS is often not properly calibrated (Parasuraman & Riley, 1997), leading to operators interfering with system recommendations which are actually correct. The other side of this lack of proper calibration is that operators also do not always notice when the DSS makes an error which has been described as the downside of automation bias (Parasuraman & Manzey, 2010). In summary, even though automated DSS can be quite helpful to improve critical performance, they are often not used appropriately.

In assessing possible consequences of DSS on human visual search performance, it further has to be considered that in a plethora of real-world applications, time pressure is a ubiquitous workload factor. For instance, consider a luggage screener at the airport who might see a long line of passengers waiting, or a physician in the emergency room needing to quickly perform an urgent sonography. It seems obvious that in those cases, support by an automated DSS might be particularly helpful as time pressure leads to decreased manual performance in these kinds of tasks (e.g., Rice & Keller, 2009; Rieger & Manzey, 2022). Moreover, some earlier research has even shown that by increasing dependence under very high time pressure, overall performance could even increase under time pressure

with a highly reliable DSS (Rice et al., 2008; Rice & Keller, 2009; Rice et al., 2010; Rice & Trafimow, 2012). Specifically, if operators followed the recommendations of a highly reliable DSS under high time pressure more strictly, this might lead to improved overall performance. So far, however, this effect has mainly been demonstrated under conditions of extremely high time pressure where individuals might not even have had a choice but to depend on the automation (e.g., 2 s for complex visual stimuli in Rice and Keller, 2009). Moreover, it remains unclear whether an increased dependence under time pressure is capable of closing the gap of joint human-DSS performance vs. DSS-alone performance, as even then, dependence might not be high enough. Finally, even though some previous research (McCarley, 2009; Rieger et al., 2021) has examined visual search behavior under time pressure, the specific interplay of time pressure and support by an automated DSS and its consequences on visual search behavior remains under-researched.

Based on this background, the present research aims to gain a better understanding of the consequences of time pressure and the availability of DSS support on visual search performance as well as a possible interaction of both factors. To study this, we conducted two online experiments using an uncovering visual search paradigm where participants used their mouse to uncover parts of the stimuli to search for specific targets. As the medical field is one domain in which automation is being introduced more and more, we used a task somewhat similar to mammography. However, to also make the task feasible for novices, the task was to search for specific letters which were embedded in noise which resembles that typically present in mammograms (Burgess et al., 2001). This paradigm allowed us to not only measure the responses (i.e., decisions about the presence of a target) but also to get a better understanding of the search which was done prior to each decision. Note, though, that the use of this mouse-uncovering technique in this kind of paradigm makes the search also somewhat artificial as visual search in the real world is not restricted to a fixed-size square to uncover an image. However, by using this paradigm, we were able to directly check how much of each stimulus participants searched through and what impact

time pressure and automation support had on this. We always briefly displayed the full stimulus before the mouse-over search started. This was done to allow participants some attentional guidance for the stimuli (e.g., Wolfe, 2021) as would be the case in any real-world application without sequential stimulus uncovering. This kind of paradigm has been previously used (Matzen et al., 2021; Rieger et al., 2021) and a similar paradigm has been shown to be able to capture comparable effects as eye-tracking studies (Matzen et al., 2021).

In the first experiment, two groups of participants performed the simulated x-ray screening task with or without support by a highly reliable DSS, respectively. Time pressure was varied within-subjects by using a trialwise countdown for each stimulus, using either a rather strict or loose time restriction. We hypothesized that performance (i.e., sensitivity) is generally better with a highly reliable DSS than when working manually (e.g., Rieger & Manzey, 2022). Moreover, we hypothesized that there was a negative effect of time pressure on performance, but only in the manual group with no negative effect (or even a positive effect) in the automation group (e.g., Rice & Keller, 2009). As we used a DSS which was highly reliable but not perfect, we decided to implement a liberal response threshold, i.e., to have more false alarms than misses to mimic the safety-first approach typically found in the real-world. Assuming that participants would follow the DSS in the majority of cases, we hypothesized that the overall response bias were more liberal in the automation than in the manual group. Further, in line with Rieger et al., 2021, we hypothesized that the criterion became more conservative under time pressure, regardless of automation group.

We also had hypotheses with respect to search amount which we defined as the percentage of the image uncovered during the mouse-over search. In line with predictions from SST models of visual search (e.g., Bricolo et al., 2002; Pashler, 1987), we hypothesized a main effect of target presence with a smaller search amount for target present than for target absent trials. However, crucially, we also hypothesized that target

presence interacted with time pressure. Specifically, we expected that under high time pressure, the absent search is cut shorter and therefore, the difference between present and absent trials is reduced (though not zero) (Rieger et al., 2021). Further, we hypothesized a similar interaction between target presence and group. That is, as the automated DSS would likely give some additional evidence to base decisions on, we expected that this would cut short absent searches more strongly than present searches (Rieger et al., 2021).

In Experiment 2, we wanted to extend our findings to another time-pressure manipulation. The experiment involved a replication of all factors as the first one, but used a different approach to vary the time pressure. Specifically, we aimed to induce time pressure through opportunity costs (Rieskamp & Hoffrage, 2008), where we impose a time limit per block instead of a fixed time per trial. In the real world, time pressure sometimes also comes by having a total time available for a number of tasks instead of a fixed limit for each task. Therefore, this approach of manipulating time pressure should extend the external validity of our research.

Experiment 1

As previously mentioned, the goal of Experiment 1 was to investigate the impact of time pressure and automated DSS support on performance and search behavior in a simulated x-ray screening task. To this end, we varied between-subjects whether they worked on the task manually (i.e., manual group) or whether they received support by a highly reliable automated DSS (i.e., automation group). Moreover, as time pressure likely also varies from time to time in the real world, we varied time pressure within-subjects (i.e., low vs. high) by inducing a trialwise time limit. The time-pressure condition changed after the first half of the experiment with the order of time-pressure conditions counterbalanced across participants.

Method

This research complied with the tenets of the Declaration of Helsinki, and both experiments were approved by the local ethics committee at the Department of Psychology, Technische Universität Berlin, Germany. Informed consent was obtained from each participant.

Participants

Participants were recruited via the platform Prolific, as well as a web portal of Technische Universität Berlin. They were either paid £4 (Prolific) or received course credit (TU Berlin participants). 60 participants (25 female, 35 male) were randomly assigned to the automation/manual groups and counterbalancing block order groups in equal numbers. Note that there was no bias of recruiting platform in the assignment of participants to the different (counterbalancing) groups ($X^2(3, N = 60) = 1.89$, $p = 0.596$). Participants ranged in age from 18 to 40 ($M_{age} = 25.02$, $SD = 4.72$) and reported normal or corrected-to-normal vision (including potential color deficiency) and fluency in English. The desired sample size was pre-determined in a power analysis using GPower (Faul et al., 2007) for a 2×2 mixed-ANOVA. That is, to achieve at least .85 power for relatively small effects of $f = 0.2$, at least 60 participants were required and therefore recruited.

To ensure good data quality despite conducting the experiment online and avoid including participants with arbitrary performance, we excluded a total of 40 additional participants due to a) poor accuracy (i.e., below 60%, 25 participants), b) participants pressing the same response key more than 15 times in a row which seems highly unlikely if they paid attention to the task given a random sequence of the two response alternatives (4 participants), c) not responding in time more than eight times (10 participants), and d) missing data for one block (1 participant). We established these criteria prior to data collection to be able to continue recruiting participants until having a full sample of 60 participants. This was done to ensure that participants who were not fully engaged with

the task were not included in further analyses.

Apparatus and Stimuli

The experiment was programmed using jspsych (de Leeuw, 2015) and was run on a JATOS (Lange et al., 2015) server so that participants could individually run the experiment in their browser. The experiment worked only on laptop/desktop computers. Thus, participants were required to use only these devices in order to take part.

The task involved a simulated medical x-ray screening task where participants searched for a target letter "E" among distractor letters "F" which were embedded in simulated noise with the power spectrum of $1/f^3$. This kind of noise resembles the power spectrum of mammograms (Burgess et al., 2001) and was the same as used in previous research (Hebert et al., 2020; Sha et al., 2018). In each image, there were a total of 12 letters which could be rotated by 0, 90, 180, or 270 degrees. If there was no target present, there were just 12 "F"'s and if there was a target present, 11 "F"s were shown along with one "E". Responses were given using the "q" key for target absent, and using the "w" key for target present. Letters were placed in randomly selected locations in an invisible 10x10 matrix to avoid overlap. The letters were white letters in font size 20 on a black background and then blended with the noise images at an opacity of 94%. Half of the stimuli contained a target. Images were sized 512x512 px but resized to 500x500 px upon presentation. The stimulus order for the experiment was randomly generated, but the same for each participant.

Procedure

Participants were randomly assigned to either the manual or the automation group, with the order of time-pressure conditions (i.e., high or low time pressure in the first half of the main part of the experiment) being counterbalanced across participants. The experiment consisted of six blocks and took approximately 30 minutes. In each block, half the trials were target present trials and half the trials target absent trials. The first two

blocks were considered as practice and were manual blocks, regardless of group assignment.

The first practice block consisted of just 10 trials and stimuli were shown in full without the mouse-over procedure to familiarize participants with the stimulus material. In this practice block, each trial started with a 1000 ms fixation cross, followed by the stimulus which was displayed for a maximum of 16 s. Participants received feedback after each trial for 1000 ms.

With the second practice block (20 trials), participants were introduced to the mouse-over search and the trial countdown. Specifically, each trial again started with a fixation cross of 1000 ms, but now this was followed by a 200 ms full preview of the stimulus, after which again a fixation cross appeared. With the onset of the preview, a trial countdown (16 s in the practice block) started which was displayed to the participants in the upper left corner above the stimulus. To start the mouse-over search, the participants now had to click on the second fixation cross. This feature was implemented to ensure that, on each trial, the search had to begin from the same position. Upon clicking on the fixation cross, a gray foreground appeared, and upon mouse movement, an area of 100x100 px around the mouse was uncovered. As soon as the mouse was moved, previously uncovered areas were grayed out again. Upon registering the response on the "q" or "w" key, participants received feedback on their accuracy for 2000 ms. From this block on, the remaining number of trials in the respective block was always continuously displayed above the image stimulus.

After completing the training blocks, participants were then instructed that now the main part of the experiment would begin, and, that they would have severe or loose time restriction for each trial (depending of the respective counterbalancing group), indicated again top left above the stimulus. After two blocks in the respective time-pressure condition, the time-pressure condition was changed to the condition not yet worked on. In the high-time-pressure blocks, participants had a total of 6 s for each trial (from preview onset). In the low-time-pressure blocks, participants now had a total of 11 s for each trial.

Each main block consisted of 28 trials, resulting in a total of 112 experimental trials. Participants were instructed to avoid not responding in time and that not responding in time would be counted as a target absent response. Instead of a trial-by-trial feedback, participants now received feedback on their performance after every block (i.e., overall feedback on the number of correct decisions in the block just finished). Only in the case that a trial was not answered in time, participants were reminded directly after the trial to respond in time and that no response counts as target absent. Other than replacing the trial feedback with a blank-screen inter-trial-interval (ITI) of 1000 ms, the trial procedure in the main blocks was the same as in the second practice block.

In the automation group, prior to the main blocks, participants were additionally informed that they would now be supported by an automation. Specifically, they were instructed that if the automation detected a target, the full image would be framed by a red rectangle, and if no target was detected by the automation (i.e., target absent), the image would be framed by a green rectangle. The rectangle appeared with the preview onset. In addition, on the top right above the stimulus, an information text was shown in red/green, respectively with onset of the clickable fixation cross (i.e., "Present – ALARM" in red or "Absent – All good" in green). An exemplary high-time-pressure trial with automation support is shown in Figure 1. They were also instructed that the automation is not perfect but had a reliability of over 90%. The true reliability of the automation was 92.86% (hit rate: 96.42%, FA rate: 10.17%), with perfect reliability in the first block of each condition and three false alarms and one miss in the the second block of each condition. This was done in order to allow participants to build up some trust towards the DSS by not having them experience failures right away in the respective condition. This resulted in a loglinear-corrected (Hautus, 1995) d_a of 2.80 and a liberal response criterion C of -0.23. Automation failures were always associated with the same image stimuli to ensure that any time-pressure effect in these trials was not related to image difficulty.

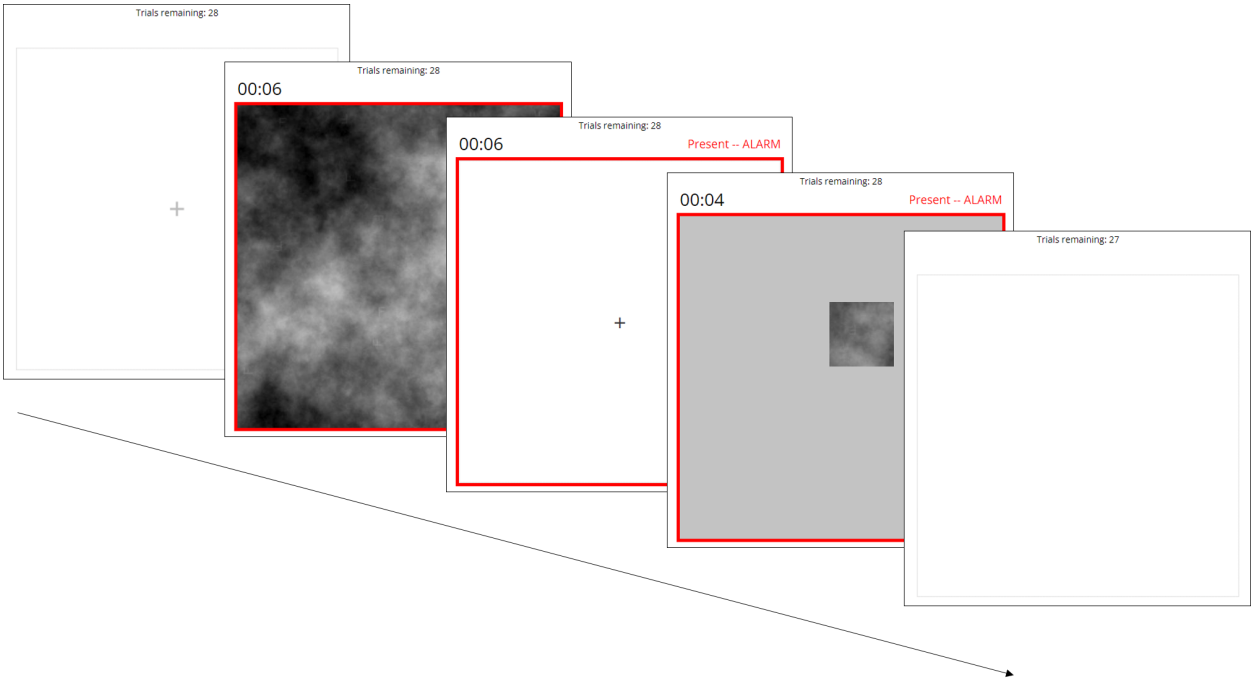

Figure 1

Trial procedure example of a high-time-pressure block with automation support. Note that in the manual group, there was no colored surrounding rectangle. In Experiment 2, rather than a trialwise countdown, the countdown ran continuously throughout the block while on average providing the same time per stimulus as in Experiment 1.

Design and Dependent Variables

Automation support was varied between-subjects (i.e., manual vs. automation) and time pressure was varied within-subjects (i.e., high vs. low). We used the signal detection measures d_a (as specified in Sterchi et al., 2019) and the criterion C (see Stanislaw and Todorov, 1999 for computation) as our main dependent variables. In case of the manual group, d_a and C directly indicate the visual search performance of the individual participant. In the group with automation support, they indicate the overall sensitivity and response bias of decisions made knowing the assessment provided by the DSS. In both groups, the corrected d_a was used because it has been found to be more appropriate for luggage screening data than the normal d'—and as luggage screening and medical image

perception share many commonalities (Gale et al., 2000), d_a seems like the appropriate measure here, too. Note, though, that the results are basically the same when using the standard d' instead of d_a. As some participants had perfect hit or false alarm rates, we applied the loglinear correction (Hautus, 1995) to all hit/false alarm rates. Moreover, we also analyzed the search amount, that is, the percentage of the image which was uncovered using the mouse-over. Search amount was calculated from the mouse-movements (i.e., how much of the image was uncovered by the search square in relation to the overall image size). Finally, for the automation group, we also analyzed dependence, using agreement rates with the automated DSS (i.e., agreement with target recommendations (i.e., alarms) as a measure of compliance and agreement with target absent recommendations (i.e., non-alarms) as a measure of reliance). Including these additional measures potentially allows us to gain a better understanding of the automation use.

Assumptions for the parametric tests were checked and we found some violations. However, we are confident that our ANOVA results were not affected by these violations for three reasons. First, the F-test is relatively robust towards deviations of variables from normal distribution and inhomogeneity of variances (Glass et al., 1972), particularly with equal group sizes (Blanca et al., 2018) as was the case in our research. Second, we re-ran the two-way ANOVAs using robust ANOVAs (Mair & Wilcox, 2020) which returned the same results. Third, most of our effects are relatively large (Cohen, 1988) and it seems highly unlikely that they would disappear without the few violations we detected.

Results

Note that in general, main effects with interactions for the same factor are only interpreted if the main effect is present at all levels of the respective other factor(s). Otherwise, we only interpret the interaction.

Performance

For d_a (see Figure 2A), the main effect of group was significant, $F(1, 58) = 74.161$, $p < 0.001$, $\eta_p^2 = 0.561$, with better sensitivity in the automation (2.10) than in the manual (1.12) group. Moreover, the main effect of time pressure was also significant, $F(1, 58) = 23.06$, $p < 0.001$, $\eta_p^2 = 0.284$, however, this was modulated by a significant group $\times$ time pressure interaction. That is, the interaction of group and time pressure was also significant, $F(1, 58) = 10.948$, $p = 0.002$, $\eta_p^2 = 0.159$. As becomes clear from Figure 2A, the effect of time pressure was only present in the manual group (1.49 vs. 0.75, $p < 0.001$) but not in the automation group (2.17 vs. 2.04, $p = 0.295$). Moreover, it also becomes evident from Figure 2A that participants in the automation group did not properly depend on the automation, as their performance was much worse than that of the automation alone. We followed up on this analysis by using a simple t-test to check whether even under low time pressure, the joint performance of the DSS-human dyad was worse than that of the automation alone, which was the case, $t(29) = 5.63$, $p < 0.001$.

For the response criterion C (see Figure 2B), only the main effect of group was significant, $F(1, 58) = 105.255$, $p < 0.001$, $\eta_p^2 = 0.645$, with a more liberal response criterion (-0.001) in the automation group and a conservative response criterion in the manual group (0.64). This main effect makes sense as the automation was more false alarm than miss prone, however, this also means that participants not supported by the DSS failed to detect many targets. Neither the main effect of time pressure ($p = 0.781$) nor the interaction of the two factors ($p = 0.103$) were significant.

Search Amount

Our paradigm allowed us to further analyze the search amount, i.e., how much of the stimulus participants uncovered during their search prior to their decision. We conducted an ANOVA with group as the between-subjects and time pressure and target presence as within-subjects factors, as previous research has shown that search amount

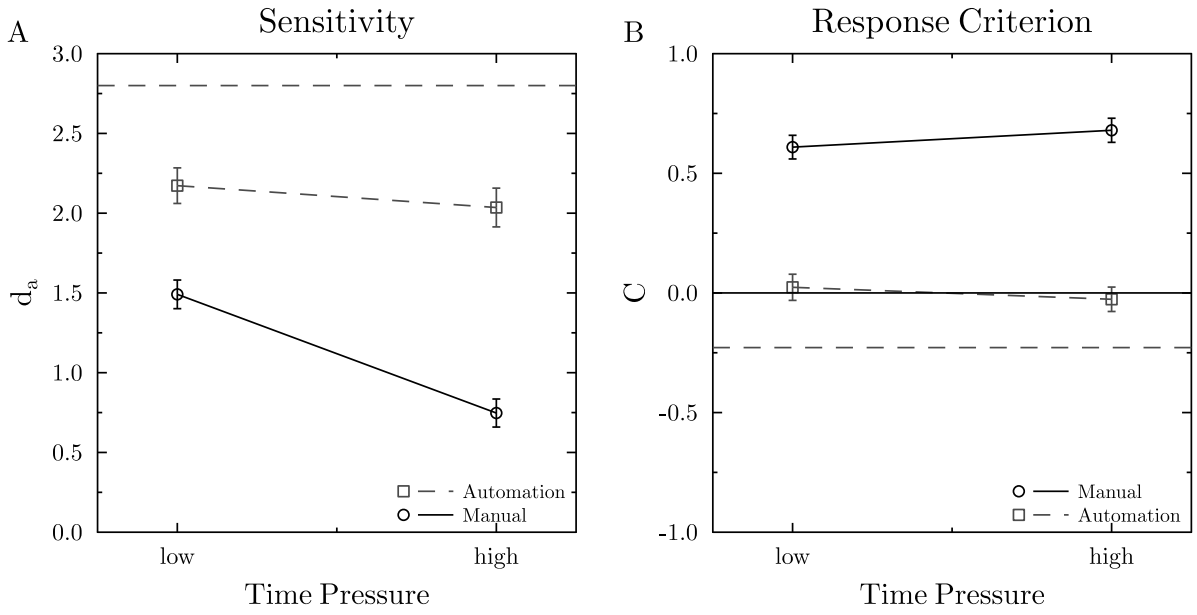

Figure 2

Results for the means of the signal detection theory measures d_a (A) and C (B) in Experiment 1 as a function of time pressure (low vs. high) and group (automation vs. manual). Dashed grey lines represent the automation alone, that is the system characteristics. Error bars represent the standard error of the mean.

differs between target absent and present trials (Rieger et al., 2021). This ANOVA (see Figure 3) revealed a significant main effect of group, $F(1, 58) = 14.835$, $p < 0.001$, $\eta_p^2 = 0.204$, with participants in the automation group searching through much less of the stimulus (39.3%) than participants in the manual (59.3%) group. The main effect of time pressure was also significant, $F(1, 58) = 75.865$, $p < 0.001$, $\eta_p^2 = 0.567$, with a smaller search amount under high (43.4%) than under low (55.2%) time pressure. Moreover, as expected in line with visual search models with a self-terminating component, the main effect of target presence was also significant, $F(1, 58) = 177.774$, $p < 0.001$, $\eta_p^2 = 0.754$, with much more being uncovered in target absent (55.8%) than in target present (42.8%) trials. There were also several significant interactions except the group × time pressure interaction ($p = 0.205$). Interestingly, the interaction of group and target presence was

significant, $F(1, 58) = 23.112$, $p < 0.001$, $\eta_p^2 = 0.285$, with a larger effect of target presence in the manual (difference: 17.6%) than in the automation (difference: 8.2%) group, indicating that participants terminated their search earlier with automation support. Moreover, the interaction of time pressure and target presence was also significant, $F(1, 58) = 71.494$, $p < 0.001$, $\eta_p^2 = 0.552$, with a much larger difference between absent and present trials under low (difference: 17.3%) than under high (difference: 8.6%) time pressure, as under high time pressure, participants were likely forced to terminate their search earlier. Finally, the three-way interaction was also significant and is best understood by Figure 3, $F(1, 58) = 4.221$, $p = 0.044$, $\eta_p^2 = 0.068$. That is, in the manual group, the difference between the time-pressure effects on target absent vs. present trials was larger (10.9% difference between time-pressure effects in absent and present) than in the automation group (6.6% difference), as seen in the differences of the absent vs. present slopes in Figure 3's panels A and B.

Automation Use

Finally, for overall dependence, compliance, and reliance, there was no effect of time pressure (all $ps > 0.490$). However, when taking a closer look at the automation failures, it became clear that when under time pressure, participants were more likely to follow the erroneous recommendations of the automation. Specifically, there was a main effect of time pressure, $F(1, 29) = 10.545$, $p = 0.003$, $\eta_p^2 = 0.267$, with higher agreement with automation errors under high (84.2%) than under low (67.5%) time pressure. Moreover, in an exploratory manner, we separated the data into trials where the automation indicated an alarm (and the stimuli for which the automation would have indicated an alarm in the manual group) and trials where the automation indicated no alarm (and the stimuli for which the automation would have said absent in the manual group). We found that participants with the DSS responded present on 86.8% of the trials where the automation indicated an alarm, while only 52.1% of the participants in the manual group responded

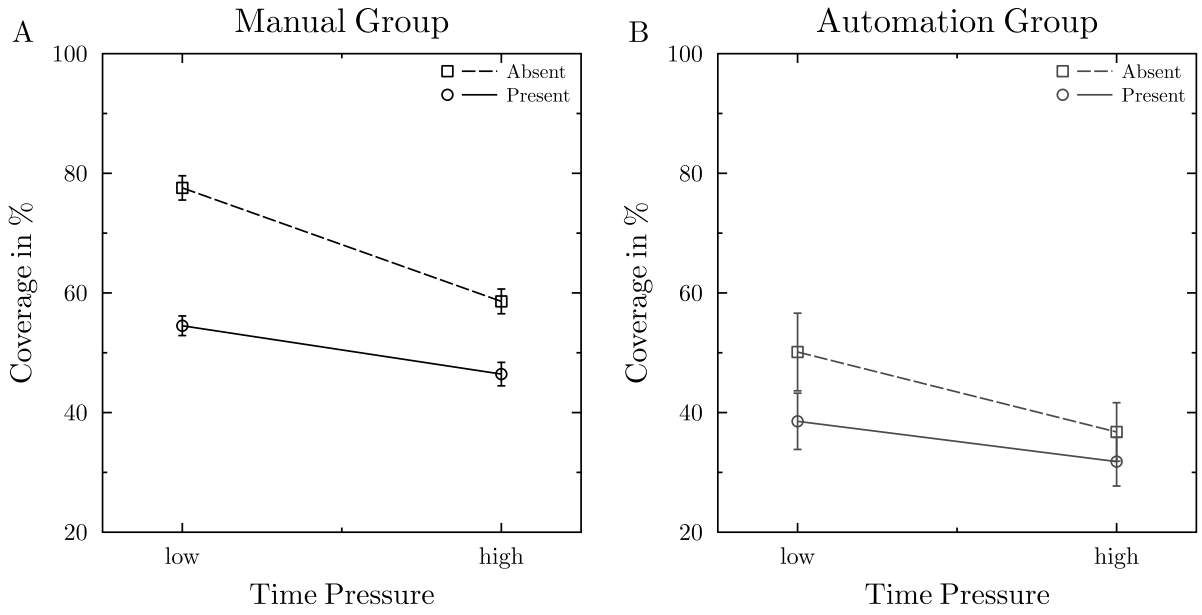

Figure 3

Mean search amount in % separately for each group (A: manual, B: automation), as a function of time pressure and target presence. Error bars represent the standard error of the mean.

present for the same stimuli. Vice versa, for the trials in which the automation recommended absent, participants in the automation group responded absent on 93.0% of these trials, while 91.2% of the manual group participants responded absent for the same stimuli.

Discussion

In line with our hypotheses, we found a negative effect of time pressure on sensitivity in the manual but not in the automation group. However, in line with some (Rieger et al., 2021; Rieger & Manzey, 2022) findings but in contrast to other previous studies using very severe time-pressure manipulations (Rice et al., 2008; Rice & Keller, 2009; Rice et al., 2010; Rice & Trafimow, 2012), we did not find any performance benefits of time pressure with a highly reliable DSS at hand. Crucially, though, the joint

performance of participants working with the DSS was still worse than that of the automation alone, reinforcing earlier concerns (e.g., Bartlett & McCarley, 2017; Rieger & Manzey, 2022) about sub-optimal dependence with highly reliable DSS. Specifically, the joint sensitivity (d_a) of the human-DSS dyad was below that of the DSS alone. Moreover, the joint response criterion (C) was also more conservative than that of the DSS, indicating that participants changed some DSS hits into joint misses.

The findings in search amount allowed us to get a better understanding of the results obtained in the SDT performance measures. It was evident that with an automated DSS, participants searched through less of the stimuli than when working manually, indicating some kind of automation bias (Parasuraman & Manzey, 2010). Some participants even chose to fully depend on the automation (i.e., agreed with 100% of the recommendations: five cases out of 30 under high and four cases out of 30 under low time pressure). Further, there were some participants in the automation group who only searched very little (i.e., eight cases out of 30 of search amounts smaller than 10% in the automation group for both high and low time pressure). Thus, in line with our predictions based on SST models of visual search, the interaction of target presence and automation group revealed that participants based their target absent responses on less evidence when being supported by a DSS than when working manually. Moreover, also in line with SST model predictions, when working under high time pressure, the difference in search amount between target absent and present trials was smaller than when having more time available, also indicating that high time pressure forces absent decisions to be made based on a less thorough search.

In evaluating these results, it should be noted that not every real-world case of time pressure comes on the basis of single trials. Specifically, sometimes, time pressure also comes from a limited overall time to conduct a certain number of tasks involving several decisions (Rieskamp & Hoffrage, 2008). For instance, consider a radiologist who might need to fill in for a colleague, thus leaving them with an increased workload for a certain duration, having to work on more images during the same time span than normal. This

would likely result in some kind of time pressure which could be different from having a specific time limit for each case. In order to corroborate the findings of Experiment 1 and also to extend the findings to another variant of time pressure, Experiment 2 was conducted using a blockwise manipulation of time pressure.

Experiment 2

The second experiment was designed to check whether the results of Experiment 1 can be extended to another manipulation of time pressure, that is, a blockwise countdown which allows participants to manage the time available more freely across trials rather than enforcing a strict limit per trial. Note though, that we provided participants with some feedback (for details, see below) about their current speed during the block. This kind of progress feedback is unlikely to be available in any real-world application with a similar time restriction for a certain number of tasks. However, in this experiment, we did this to allow participants some information about their progress and time management, as in contrast to the real world, the task and its time allocation were of course relatively new to the participants. Specifically, we expected to replicate our findings of Experiment 1 and therefore had the same hypotheses. Regarding the most central hypothesis of performance, we again hypothesized that sensitivity would be higher with automation support than without, a negative effect of time pressure and that the automation can ameliorate this negative effect. Note though, that our hypotheses regarding the criterion and search amount were also the same as in Experiment 1.

Method

Participants

A fresh sample of 60 participants (40 male, 19 female, 1 non-binary) was recruited via Prolific and a web portal of Technische Universität Berlin. They ranged in age from 18 to 35 ($M_{age} = 24.22$, $SD = 4.34$) and were predominantly right-handed (54 right-handed, 5

left-handed, 1 ambidextrous), again reporting normal or corrected-to-normal vision. Again, there was no systematic bias in (counterbalancing) group assignment of participants recruited via the different platforms ($X^2(3, N = 60) = 6.34$, $p = 0.096$).

We applied the same exclusion criteria as in Experiment 1 to ensure high quality data despite collecting data online. Overall, 49 additional participants were excluded because of a) accuracy below 60% (6 participants), b) because of pressing the same response key more than 15 times in a row (6 participants), c) because they did not complete the experiment within 60 minutes (5 participants), or d) because they did not respond in time more than eight times (32 participants). Note that while the latter number might seem high, in this experiment it was theoretically possible to start each block and just wait for the experiment to finish, which unfortunately some participants did.

Apparatus, stimuli, design, and procedure

Apparatus, stimuli, design, and dependent variables and their assessment were the same as in Experiment 1. The only change in the procedure was that in Experiment 2, the time restrictions were not given on a trial-by-trial basis, but induced through opportunity costs via a blockwise countdown. However, to ensure consistency with Experiment 1, the mean total time available for each stimulus was the same. With respect to the procedure, the initial practice block with no mouse-over was the same as in Experiment 1. Then, in the second practice block with mouse-over, participants were familiarized with the blockwise provision of time available for the task. Specifically, they had a total time of 6 min and 20 s for performing the 20 trials, resulting from adding the time of the fixation and feedback to the time given per trial and multiplying this sum by the number of trials in this block. Applying the same rationale, this resulted in blockwise times of 3 min 44 s and 6 min and 4 s for the following experimental blocks with high and low time pressure, respectively. During these blocks, the experiment automatically checked whether the average response time of participants would lead them to finish the block in time. To make

participants aware that they need to speed up their responses to finish all trials of the block in time, every five trials and on the final five trials of each block, they were shown an orange warning "you need to speed up" along with a 500 ms 500 Hz beeping sound but only if they were too slow on average up to that point. This was done in order to allow participants some information about their time management. Participants were made aware of this in-block feedback prior to the mouse-over blocks. Again, participants were instructed to try to respond to all trials in time and that not responding in time would count as target absent, but as in this experiment it was theoretically possible to not see all trials of a block, we only included the trials which the participant had actually seen in the analyses.

Results

Performance

The effects in the performance results (see Figure 4) were quite similar to those obtained in Experiment 1. For d_a, there were again significant main effects for both group, $F(1, 58) = 14.484$, $p < 0.001$, $\eta_p^2 = 0.2$, as well as time pressure, $F(1, 58) = 17.843$, $p < 0.001$, $\eta_p^2 = 0.235$. However, these main effects were modulated by a significant interaction of the two factors. Specifically, the interaction of time pressure and group was again significant, $F(1, 58) = 11.905$, $p = 0.001$, $\eta_p^2 = 0.17$, with a difference between low and high time pressure only present in the manual (1.80 vs. 1.13, $p < 0.001$) but not in the automation group (2.04 vs. 1.97, $p = 0.586$), again showing the potential of automation to reduce the negative effects of time pressure. Further, the difference between the two groups was only significant under high ($p < 0.001$) but not under low ($p = 0.135$) time pressure. Again, the joint performance of human and DSS was significantly worse than that of the DSS alone, even under low time pressure, $t(29) = 6.55$, $p < 0.001$, indicating that participants deteriorated the high performance of the automation alone.

For the response criterion C (see Figure 4B), there was again a main effect of group, $F(1, 58) = 36.435$, $p < 0.001$, $\eta_p^2 = 0.386$, with a much more liberal criterion with

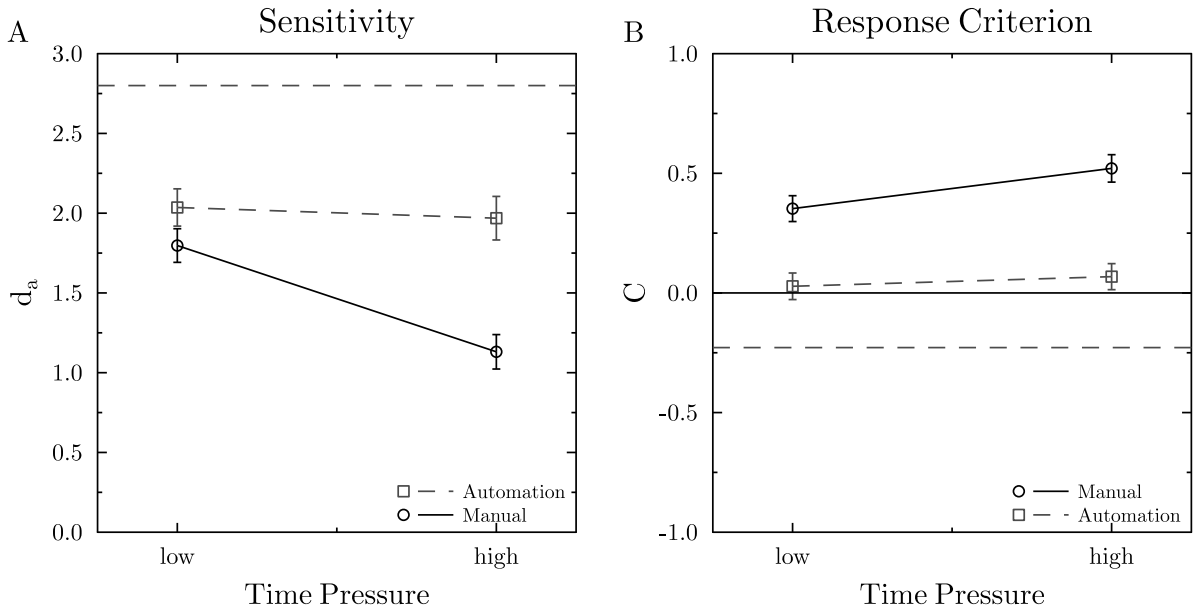

Figure 4

Results for the means of the signal detection theory measures d_a (A) and C (B) in Experiment 2 as a function of time pressure (low vs. high) and group (automation vs. manual). Dashed grey lines represent the automation alone, that is the system characteristics. Error bars represent the standard error of the mean.

automation (0.05) than when working manually (0.44) which makes sense against the background of the DSS's liberal criterion. In this experiment and in contrast to Experiment 1, the main effect of time pressure was also significant, $F(1, 58) = 5.486$, $p = 0.023$, $\eta_p^2 = 0.086$, with a more conservative criterion under high (0.29) than under low (0.19) time pressure which is in line with the findings of Rieger et al. (2021). The interaction was not significant ($p = 0.156$).

Search Amount

We again also analyzed the search amount to get a better understanding of the strategies involved in the search (see Figure 5) using an ANOVA with the additional factor target presence. The result pattern was quite similar to that in Experiment 1. That is,

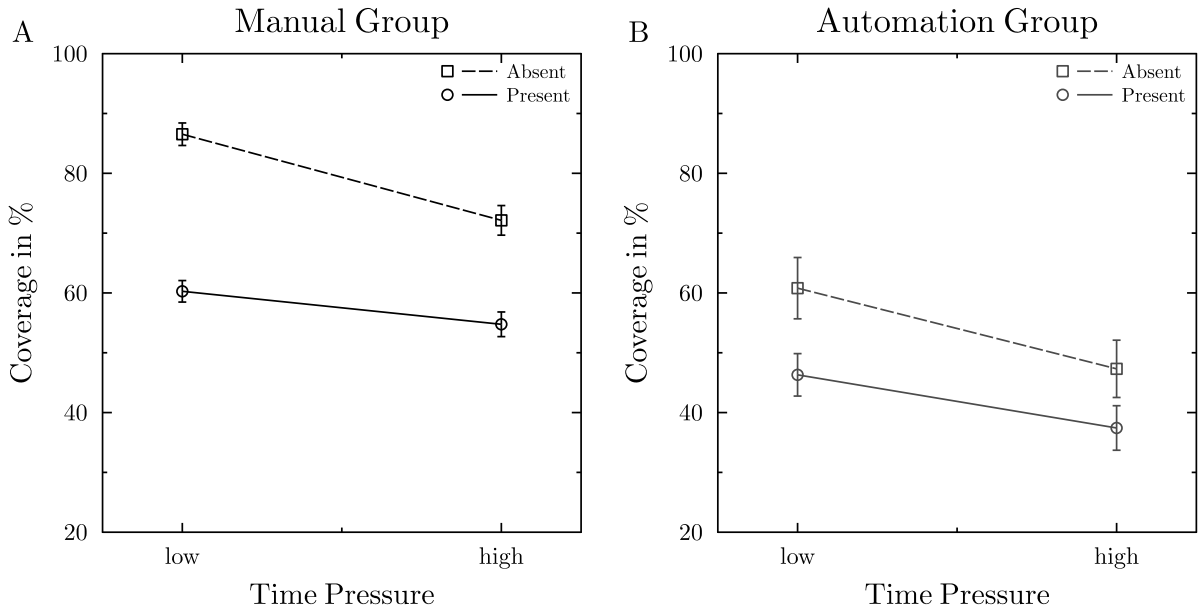

Figure 5

Mean search amount in % separately for each group (A: manual, B: automation), as a function of time pressure and target presence in Experiment 2. Error bars represent the standard error of the mean.

there were again main effects of group, $F(1, 58) = 23.172$, $p < 0.001$, $\eta_p^2 = 0.285$, time pressure, $F(1, 58) = 42.304$, $p < 0.001$, $\eta_p^2 = 0.422$, and target presence, $F(1, 58) = 153.464$, $p < 0.001$, $\eta_p^2 = 0.726$. As in Experiment 1, larger areas were searched through in the manual than in the automation group (68.4% vs. 48.0%), in the low than in the high time pressure condition (63.5% vs. 52.9%), and in absent than in present trials (66.7% vs. 49.7%). There were again several interactions. Specifically, time pressure and target presence interacted significantly, $F(1, 58) = 18.95$, $p < 0.001$, $\eta_p^2 = 0.246$—that is, as in Experiment 1, the difference between target absent and present search amount was much smaller under high time pressure (difference: 13.6%) than under low time pressure (difference: 20.4%). Moreover, as in the first experiment, the interaction of group and target presence was also significant, $F(1, 58) = 12.304$, $p < 0.001$, $\eta_p^2 = 0.175$, with a larger difference between target absent and present trials in the manual (difference: 12.2%) than

in the automation (difference: 21.8%) group. No other interaction was significant ($ps > 0.172$).

Automation Use

We again analyzed overall dependence, as well as compliance and reliance to check whether time pressure had an impact here. There was again no overall effect of time pressure (all $ps > 0.356$). When again separately checking the trials in which the automation made an error, there was no effect of time pressure on dependence in failure trials in this experiment, $F(1, 29) = 2.012$, $p = 0.167$, $\eta_p^2 = 0.065$, with only a descriptive difference in agreement with automation errors in the high time pressure (73.3%) vs. low time pressure (63.3%) conditions. As in Experiment 1, we ran an exploratory analysis separating the red/green trials of the automation and checked the response behavior to these alarms/non-alarms in comparison to how the manual group responded to the same stimuli. Similar to Experiment 1, participants in the automation group responded present on 83.7% of the trials where the DSS indicated an alarm, while only 63.0% of the participants in the manual group responded present for the same stimuli. Vice versa, for the trials in which the DSS recommended absent, participants in the automation group responded absent on 91.8% of these trials, while 89.5% of the manual group participants responded absent for the same stimuli. This seems to show that participants benefited particularly from the alarms of the DSS, but not as much from the non-alarms.

Discussion

Experiment 2 mainly replicated and extended the findings obtained in Experiment 1. Thus, the blockwise time-pressure manipulation seems to produce similar findings as the trialwise manipulation. As in Experiment 1, there were again some participants in the automation group who always followed the automation's recommendations (two cases out of 30 under high time pressure and no cases under low time pressure). Moreover, there were also again participants in the automation group who

conducted rather incomplete searches of the stimuli (i.e., mean search amount smaller than 10% in four out of 30 cases under high and two cases out of 30 under low time pressure). As was mentioned above, time pressure through opportunity costs is another version of time pressure instead of a time restriction per trial—but the results were mostly the same. We will therefore continue to discuss the two experiments jointly.

General Discussion

The goal of the present experiments was to investigate the performance consequences of automated DSS, time pressure, and their possible interactions in an applied visual search task, simulating the demands of an x-ray screening task. Moreover, to gain a better understanding of the search behavior, we used a relatively novel paradigm, which allowed us to track how much of the stimulus participants uncovered prior to their decisions. In two experiments, participants either performed the visual search task manually or with the support of a DSS. In addition, the time available for conducting the task was varied, in order to simulate conditions with varying degrees of time pressure. Specifically, where Experiment 1 used the most typical manipulation of time pressure (i.e., trial-by-trial time restrictions), Experiment 2 extended the findings to a different variation of a time-pressure manipulation (i.e., time pressure through opportunity costs).

The first key finding was that with support of a DSS, no negative effect of time pressure on sensitivity was observed. This marked an important contrast to the unsupported manual groups, where typical performance impairments of time pressure were obtained, consistently reflected in a lower sensitivity for discriminating between target-present and target-absent images (see e.g., McCarley, 2009; Rice & Keller, 2009; Rieger et al., 2021; Rieger & Manzey, 2022). This finding was highly consistent across both experiments, suggesting a general potential of automation support compensating negative effects of time pressure on decision making in such search tasks, independent of how exactly this time pressure is induced. The fact that automated DSS can reduce the

negative impact of a workload factor like time pressure is in line with some earlier research (e.g., Rice & Keller, 2009; Rice et al., 2010; Rieger & Manzey, 2022). However, in contrast to the findings by Rice and Keller (2009), who even found an increased visual search performance under high vs. low time pressure when supported by a DSS, we did not find such an effect. That is, even though the negative effect of time pressure was absent in the automation groups of both experiments, there was no positive effect of time pressure, neither statistically reliable nor descriptively. Hence, there was no evidence for a more heuristic strategy (Payne et al., 1988; Rice & Keller, 2009) of automation dependence under high time pressure in the present study. This challenges the assumption of Rice and Keller (2009) that participants just depend on the DSS without checking the image themselves when working under high time pressure and with a highly reliable DSS available. In contrast, even under time pressure, our participants seem to have used the automation in a way that they adapted their response bias in a more liberal direction, without becoming as liberal as the automation alone, though.

A second key finding of our experiments was that in the automation groups of both experiments, the joint performance (i.e., sensitivity) of the human-automation dyad was consistently below that of the automation alone. This finding is in line with earlier research which also showed a worse joint dyad performance than that of a highly reliable automation alone, implying a sub-optimal use of the automation (Bartlett & McCarley, 2017; Boskemper et al., 2021; Meyer, 2004; Meyer & Kuchar, 2021; Rieger & Manzey, 2022). In the present data, this sub-optimal use becomes evident when contrasting the dependence of our participants on the automation with the achieved overall sensitivity. That is, despite the fact that overall dependence was high (88.8% in Experiment 1, 87.5% in Experiment 2), the sensitivity was still much worse than that of the automation. This suggests that participants significantly interfered with the automation's recommendations not only when the automation made errors, but also when it was correct—which can be considered as a sort of disuse of the DSS. Moreover, as became evident when separately looking at the

dependence when the automation erred, participants also agreed with many incorrect recommendations—which can be considered as some sort of overtrust and misuse of the DSS. This latter effect even seems to become more pronounced under high time pressure.

Thus, it seems like participants in the automation group managed to show both behavior of disuse and misuse within the same experiment. This finding of misuse and disuse at the same time suggests that participants did not trust the DSS too little or too much. Instead, they just seem to avoid to always agree with the assessments of the DSS (i.e., full dependence) even though they were supported by a highly reliable aid.

Perhaps, one possible reason for this honest but failed attempt at improving the aid's recommendations is that following the aid all the time would be in contrast to their perception of the own role of being ultimately responsible for the final decision. This raises the question whether operators should be kept in the loop at all when automation is highly reliable, particularly under high time pressure where there is already AI available which outperforms pathologists in their typical time-pressured workflow (Bejnordi et al., 2017). However, in many real-world cases, it is of course not possible to remove the operator from the loop for legal reasons (Bryson et al., 2017) or to have someone there in case of emergencies. Further studies should try to follow up this assumption of whether human operators' trying to make sense of their role as final authority in a human-automation system might actually lead to suboptimal overall performance when interacting with highly reliable DSS. Moreover, in future studies (with more trials), including an SDT-driven approach of the behavioral consequences of trust by analyzing the differences in response criteria between alarms and non-alarms (e.g., Maltz & Meyer, 2001; Meyer et al., 2014) might be helpful to gain a better understanding of the use of the system.

Finally, the present paradigm also allowed us to take a closer look at the search behavior (i.e., how much of the stimulus was actually uncovered). These analyses revealed evidence for some kind of an automation bias—that is, the general search amount was lower when being supported by an automation than when working manually, showing that

less evidence was required to make a decision for participants working with a DSS. However, as was discussed above, this unfortunately did not result in some kind of a functional automation bias, where operators strictly follow the DSS' advice given its high reliability. The results obtained in search amount are generally in line with theoretical predictions from SST models of visual search (e.g., Bricolo et al., 2002; Pashler, 1987; Snodgrass, 1972; Treisman & Gelade, 1980; van Zandt & Townsend, 1993), as there were consistent interactions both between target presence and automation support and target presence and time pressure. Particularly the latter interaction seems problematic from an applied standpoint, as a less thorough search under high time pressure likely leads to overlooking crucial targets, as is also evident from the lower manual sensitivity under high time pressure. Moreover, it would be an interesting avenue for future research to explore differences between automation and manual groups in fixations of certain stimulus areas, potentially using eye-tracking.

Of course, the present study does not come without limitations. First, the experiment was conducted online and therefore comes with the typical limitations associated with an online experiment. However, we enforced rather strict exclusion criteria prior to our data analyses to ensure that only participants who reasonably worked on the experiments were included. Second, there were of course some differences between our experimental task and the real world. Specifically, the task which we used resembles the task characteristics of a medical visual search task but there is of course a difference between the real-world equivalent and a lab adaptation of the task. Further, the time pressure manipulation in Experiment 2 might only be somewhat helpful in understanding the differences between different types of time pressure, as in the real world, there is likely no scenario where one gets speed feedback every five images if it is required to adjust the work pace. Nevertheless, we used such live feedback in the present study to allow participants, who were new to the task, at least some information about their ongoing time management. Regardless, we believe that this kind of task can still help to get a basic

understanding of the phenomena studied in the present research. Third, with 50%, we had an unreasonably high base rate of targets which is required to have enough trials to analyze but future research needs to address this issue more systematically.

To conclude, the findings of the present experiments show that the negative impact of time pressure can be reduced when providing a highly reliably automated DSS. However, similar to concerns raised in earlier research (e.g., Bartlett & McCarley, 2017; Meyer & Kuchar, 2021), the joint performance of the DSS and the human was much less than an ideal symbiosis of human and machine would theoretically allow for. Moreover, the present experiments show that using a relatively novel uncovering paradigm in a simulated medical visual search task, it is possible to enhance the understanding of automated decision support.

Key Points

- Time pressure negatively impacts manual performance in a simulated medical visual search task
- Automation can reduce this negative effect but there were no positive effects of time pressure with a highly reliable automation
- The joint performance of the human-automation dyad was sub-optimal, regardless of time pressure
- Studying search behavior can help to gain a better understanding of automation use

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. https://doi.org/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. https://doi.org/10.1177/0018720817700258

Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., ... Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), 2199. https://doi.org/10.1001/jama.2017.14585

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, *50*(3), 937–962. https://doi.org/10.3758/s13428-017-0918-2

Boskemper, M. M., Bartlett, M. L., & McCarley, J. S. (2021). Measuring the efficiency of automation-aided performance in a simulated baggage screening task. *Human Factors*. https://doi.org/10.1177/0018720820983632

Bricolo, E., Gianesini, T., Fanini, A., Bundesen, C., & Chelazzi, L. (2002). Serial attention mechanisms in visual search: A direct behavioral demonstration. *Journal of Cognitive Neuroscience*, *14*(7), 980–993. https://doi.org/10.1162/089892902320474454

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9

Burgess, A. E., Jacobson, F. L., & Judy, P. F. (2001). Human observer detection experiments with mammograms and power-law noise. *Medical Physics*, *28*(4), 419–437. https://doi.org/10.1118/1.1355308

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge. https://doi.org/10.4324/9780203771587

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dorrius, M. D., der Weide, M. C. J.-v., van Ooijen, P. M. A., Pijnappel, R. M., & Oudkerk, M. (2011). Computer-aided detection in breast MRI: A systematic review and meta-analysis. *21*(8), 1600–1608. https://doi.org/10.1007/s00330-011-2091-9

Drew, T., Cunningham, C., & Wolfe, J. M. (2012). When and why might a computer-aided detection (CAD) system interfere with visual search? An eye-tracking study. *Academic Radiology*, *19*(10), 1260–1267. https://doi.org/10.1016/j.acra.2012.05.013

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

French, B., Duenser, A., & Heathcote, A. (2018). *Trust in automation* (tech. rep. CSIRO Report EP184082). CSIRO, Australia.

Gale, A. G., Mugglestone, M. D., Purdy, K. J., & McClumpha, A. (2000). Is airport baggage inspection just another medical image? In E. A. Krupinski (Ed.), *Medical imaging 2000: Image perception and performance*. SPIE. https://doi.org/10.1117/12.383105

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

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

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

Hebert, C. R., Sha, L. Z., Remington, R. W., & Jiang, Y. V. (2020). Redundancy gain in visual search of simulated x-ray images. *Attention, Perception, & Psychophysics*, *82*(4), 1669–1681. https://doi.org/10.3758/s13414-019-01934-x

Huegli, D., Merks, S., & Schwaninger, A. (2020). Automation reliability, human–machine system performance, and operator compliance: A study with airport security screeners supported by automated explosives detection systems for cabin baggage screening. *Applied Ergonomics*, *86*, 103094. https://doi.org/10.1016/j.apergo.2020.103094

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

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Luz, M., Strauss, G., & Manzey, D. (2016). Impact of image-guided surgery on surgeons' performance: A literature review. *International Journal of Human Factors and Ergonomics*, *4*(3/4), 229. https://doi.org/10.1504/ijhfe.2016.083516

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory – a user's guide*. Psychology Press. https://doi.org/10.4324/9781410611147

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, *49*(5), 773–785. https://doi.org/10.1518/001872007x230154

Mair, P., & Wilcox, R. (2020). Robust statistical methods in r using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors*, *43*(2), 217–226. https://doi.org/10.1518/001872001775900931

Matzen, L. E., Stites, M. C., & Gastelum, Z. N. (2021). Studying visual search without an eye tracker: An assessment of artificial foveation. *Cognitive Research: Principles and Implications*, *6*(45), 1–22. https://doi.org/10.1186/s41235-021-00304-2

McCarley, J. S. (2009). Effects of speed–accuracy instructions on oculomotor scanning and target recognition in a simulated baggage x-ray screening task. *Ergonomics*, *52*(3), 325–333. https://doi.org/10.1080/00140130802376059

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

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

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196–204. https://doi.org/10.1518/hfes.46.2.196.37335

Meyer, J., & Kuchar, J. K. (2021). Maximal benefits and possible detrimental effects of binary decision aids. *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. https://doi.org/10.1109/ichms53169.2021.9582632

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

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.

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

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

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception & Psychophysics*, *41*(3), 191–201. https://doi.org/10.3758/bf03208218

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. https://doi.org/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. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1326–1329. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/00221309.2012.705187

Rieger, T., Heilmann, L., & Manzey, D. (2021). Visual search behavior and performance in luggage screening: Effects of time pressure, automation aid, and target expectancy. *Cognitive Research: Principles and Implications*, *6*(12), 1–12. https://doi.org/10.1186/s41235-021-00280-7

Rieger, T., & Manzey, D. (2022). Human performance consequences of automated decision aids: The impact of time pressure. *Human Factors*, *64*(4), 617–634. https://doi.org/10.1177/0018720820965019

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258–276. https://doi.org/10.1016/j.actpsy.2007.05.004

Sha, L. Z., Remington, R. W., & Jiang, Y. V. (2018). Statistical learning of anomalous regions in complex faux x-ray images does not transfer between detection and discrimination. *Cognitive Research: Principles and Implications*, *3*(1), 1–16. https://doi.org/10.1186/s41235-018-0144-1

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1). https://doi.org/10.1518/155723405783703082

Snodgrass, J. G. (1972). Reaction times for comparisons of successively presented visual patterns: Evidence for serial self-terminating search. *Perception & Psychophysics*, *12*(4), 364–372. https://doi.org/10.3758/bf03207223

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

Sterchi, Y., Hättenschwiler, N., & Schwaninger, A. (2019). Detection measures for visual inspection of x-ray images of passenger baggage. *Attention, Perception, & Psychophysics*, *81*(5), 1297–1311. https://doi.org/10.3758/s13414-018-01654-8

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

van Zandt, T., & Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics*, *53*(5), 563–580. https://doi.org/10.3758/bf03205204

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Short Biographies

Tobias Rieger is a researcher and lecturer at the Department of Psychology and Ergonomics, Technische Universität 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 Universität 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.