Asymmetric effects of false positive and false negative indications on the verification of alerts in different risk conditions

Rebecca Wiczorek, Technische Universität Berlin, Germany
Joachim Meyer, Tel Aviv University, Israel

Indications from alerts or alarm systems can be the trigger for decisions, or they can elicit further information search. We report an experiment on the tendency to collect additional information after receiving system indications. We varied the proclivity of the alarm system towards false positive or false negative indications and the perceived risk of the situation. Results showed that false alarm-prone systems led to more frequent re-checking following both alarms and non-alarms in the high risk condition, whereas miss-prone systems led to high re-checking rates only for non-alarms, representing an asymmetry effect. Increasing the risk led to more re-checks with all alarm systems, but it had a stronger impact in the false alarm-prone condition. Results regarding the relation of risk and the asymmetry effect of false negative and false positive indications are discussed.

INTRODUCTION

Alarm systems and decision aids support operators' decision making in complex situations in numerous environments. Even though many systems are very good, none is ever 100% correct. They produce two types of errors – false negatives (missed detections) and false positives (false alarms). Threshold settings determine whether the system is prone to one or the other type of error.

Error frequencies are often specified through the predictive values of alarm systems. The positive predictive value (PPV) is the ratio of true to all alarms, whereas the negative predictive value (NPV) represents the ratio of correct rejections to all non-alarms (Getty, Swets, Pickett & Gonthier, 1995; Meyer & Bitan, 2002; Parasuraman, Hancock & Olofinboba, 1997).

Designers usually follow the fail-safe approach and try to minimize the number of misses (i.e., they raise the NPV). Thus, most decision aids are false alarm-prone. However, it has been argued that false alarms are as bad as misses or even more problematic with regard to safety (Dixon, Wickens & McCarley, 2007). A number of studies showed not only a reduction of trust in the alarms, but also the behavioral consequence of ignoring them, due to operators' experience of frequent false alarms (e.g. Bliss, Gilson & Deaton, 1995; Madhaven, Wiegmann & Lacson, 2006). This behavior can be described as lack of compliance, where compliance is operators' tendency to react to an alarm as if an underlying problem exists. The opposite behavior – no reaction in the absence of an alarm, is referred to as reliance (Meyer, 2004).

More recent studies showed that false alarms not only reduce compliance, but they also lower reliance, while missed detections only lower reliance (Dixon et al., 2007; Meyer, Wiczorek, & Günzler, 2014; Rice & McCarley, 2011). This asymmetric effect has been demonstrated on a number of measures. It has been referred to as the asymmetry bias, because the reduction of the other, non-related behavioral component only occurs for low PPVs, but not for low NPVs.

A suitable countermeasure to avoid reduction of compliance is to offer operators the possibility to cross-check or re-check the alarms with additional information such as raw data. Instead of ignoring alarms from a system with low PPV, participants cross-check most of them (e.g., Bliss, Jean & Prioux, 1996; Manzey, Gérard &Wiczorek, 2014).

Prior studies investigating the asymmetry bias did not use the re-checking option. Thus, the first aim of this study was to investigate, whether this type of an asymmetry effect will be found with regard to participants' frequencies of re-checking the different types of cues when additional information is available, beyond the alert.

The second aim was to explore the underlying reasons for the asymmetry effect. Some authors focus on the special role of false alarms, as they might be more salient or have a stronger impact on trust (Rice & McCarley, 2011; Dixon et al., 2007). Others suppose that the reason is more related to the reliance component (Meyer et al., 2014). Reliance might be more vulnerable to changes in reliability than compliance, no matter whether the PPV or the NPV is diminished. One possible reason could be operators' unwillingness to commit omission errors, if they are not certain about their response, because missing a critical event is usually more dangerous.

The supposed danger of an action or of refraining from it depends on the perceived risk of the situation and the expected consequences. The impact of risk on decision making with alarms is still not well understood. Parasuraman and Riley (1997) describe risk as one important factor for reliance. Wiczorek and Onnasch (2012) argue that it could moderate the relation between system properties and decision making with alarms.

Based on these theoretical assumptions, the vulnerability of reliance could be the reason for the asymmetry bias, and the variation of risk might impact on this bias as it determines the degree of vulnerability.

THE CURRENT STUDY

The aim of the current study was to investigate the potential impact of risk on the asymmetry bias and its manifestation on the level of re-checking. Therefore, we compared three cuing systems with different thresholds under two conditions of either high or low risk in a paradigm that offered a re-check option.

Participants saw images briefly (for 1 second), and the cuing system provided indications. Participants could either make an immediate decision or choose to re-check the image. One system produced the same number of false alarms and misses. The second system was prone to false alarms, and the third was a miss-prone system.

Based on prior research, the frequency of re-checking was expected to depend on the predictive values of the two types of cues, with lower values leading to more re-checks (Manzey et al., 2014). Due to earlier findings of asymmetry, we predicted that a low NPV would increase re-checks of 'no target' cues, while a low PPV should lead to both more re-checks of 'target' cues and 'no target' cues (e.g. Meyer et al., 2013).

Variation of risk level was based on the theory of Renn (2012) who stated that perception of risk is influenced by the *source of the risk* and the *risk situation*. Accordingly, we framed the high risk condition as cancer screening (source of risk) with an alarm system (situation). In the low risk condition, the same task was indicated as tissue categorization of different cells with the help of a decision aid.

In line with theoretical assumptions, we expected re-check frequencies to vary as a function of risk (Parasuraman & Riley, 1997). A higher risk level should lead to more re-checks. Additionally, we predicted that the asymmetry bias would be moderated by the level of risk (Wiczorek & Onnasch, 2012). Effects of PPV on the re-checking of both 'target' cues and 'no target' cues should only manifest themselves in the high risk condition. For the low risk condition, PPV should only influence the re-checking of 'target' cues.

Thus, participants with all three systems should re-check more cues in the high risk condition than in the low risk condition. Those working with the neutral system were expected to re-check both types of cues with the same frequency because of the equality of PPV and NPV. With the miss-prone system 'no target' cues should be re-checked more often than 'target' cues because of the lower NPV. We expected participants in the false alarm-prone condition to differentiate their behavior according to their risk condition. Under high risk, the asymmetry bias should arise, so that participants will re-check both the 'target' cues and the 'no target' cues. In contrast, the bias was expected to disappear in the low risk condition, and 'target' cues should be re-checked more often than 'no target' cues as a reaction to low PPV.

METHOD

Participants

Seventy-two students at Ben-Gurion University of the Negev and at Tel Aviv University participated in the study. Experiments were conducted in groups of up to 8 people.

Simulation environment

Participants conducted a visual search task on 19" screens. They saw a blurry picture containing nine digits in three rows and three columns (see Figure 1). Their task was to decide whether the target number 3 was present and to press the corresponding button. This task was supported by a cuing system. The system indicated its diagnoses with different colored horizontal bars under the picture and with written messages. Below the picture were two buttons for the two decision options - 'target' and 'no target'. After one second, the image disappeared and a third button was presented, labeled 're-check'. If participants were uncertain about their response, they could click on the 're-check' button and revisit the picture for another 3 seconds before making their final decision. After clicking one of the two decision buttons, a new image appeared. Participants received 10 points for every correct decision, and they lost 10 points for every wrong decision. Re-checking was penalized by the reduction of 2 points. At the end of the experiments, participants could receive a bonus payment, based on the number of points collected during the experimental blocks.

Design

The study consisted of a 2 (Risk Framing) x 3 (Cuing System) x 2 (Type of Cue) x 2 (Block) design with repeated measures on the third and fourth factor. In the high risk condition, the cuing system was framed as an alarm system with red and green cues and 'target' and 'no target' indications. The task was introduced as cancer screening. In the low risk condition, the systems was described as a decision aid, and it provided dark blue and light blue cues with the written messages 'tissue A' and 'tissue B', respectively. The task was described as tissue categorization of two different types of cells.
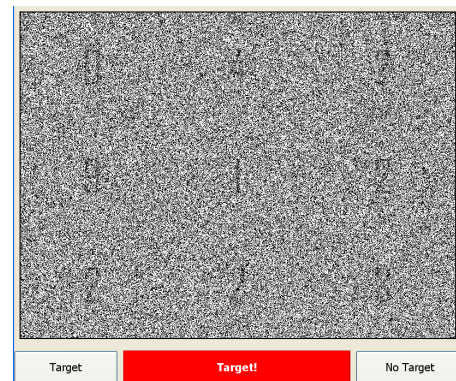

Figure 1. Experimental screen with stimulus image for a participant in the high risk condition.

The type of cue and the corresponding button was either 'target'/'tissue A' to indicated that the image contained the number 3, or 'no target'/'tissue B', signaling the absence of number 3.

The cuing systems differed with regard to the errors they committed. Participants either saw a false alarm-prone, a miss-prone or a neutral system. The neutral system produced the same number of false alarms and misses. Its PPV and NPV were 0.8. The false alarm-prone system had a PPV of 0.6 and a NPV of 0.8. The miss-prone system had a PPV of 0.8 and a NPV of 0.6. The error base rate was 0.5 in all conditions. Participants' behavior was recorded in two consecutive blocks to control for potential learning effects.

Procedure

After arriving at the lab, participants first signed a consent form. They then received standardized instructions on the screen. Training and data collection intermitted over the

course of four blocks. Participants trained the task without the cuing system. The two manual training blocks consisted of 40 trials each. Both of the two experimental blocks included 80 trials. In those blocks, participants were supported by one of the three systems, either framed as a decision aid or an alarm system. Participants' behavior was recorded. At the end, participants were thanked and paid.

RESULTS

We used a multifactorial ANOVA with repeated measures to analyze the results. The factors Cuing System and Block were not significant. Participants' re-checking frequencies did not change over time. The analysis revealed main effects for Risk Framing, $F(1,66)=4.23$, $p<.05$, $\eta^2_p=.06$, and for the Type of Cue, $F(1,66)=16.75$, $p<.001$, $\eta^2_p=.2$. Re-checking appeared more often in the high risk conditions, compared to the low risk conditions. In addition, participants re-checked significantly more often the 'no target'/'tissue B' cues than the 'target'/'tissue A' cues. This indicates greater insecurity with 'no target' cues, compared to 'target' cues.

These main effects were further qualified by two two-way interaction effects. The significant interaction Risk Framing x Cuing System, $F(2,66)=3.14$, $p<.05$, $\eta^2_p=.09$, was caused by the percentage of re-checking being highest with the false alarm-prone system in the high risk condition and lowest with the false alarm-prone system in the low risk condition, while the other two systems lead to medium re-checking rates under both risk conditions. Percentage of re-checking varied as a function of risk only in the false alarm-prone condition. The second two-way interaction effect was Risk Framing x Type of Cue, $F(2,66)=5.26$, $p<.05$, $\eta^2_p=.07$. Over all systems, re-checking of 'no target'/'tissue B' cues increased more under high risk, compared to low risk, than did the re-checking frequency of 'target'/'tissue A' cues. Thus risk had a larger effect on the re-checking of 'no target'/'tissue B' than on the re-checking of 'target'/'tissue A' cues. None of the other interaction effects was significant. These statistical effects of the current study are in line with the predictions made before. However, as can be seen in Figure 2 and Figure 3, the actual re-checking behavior did not completely correspond with our predictions.
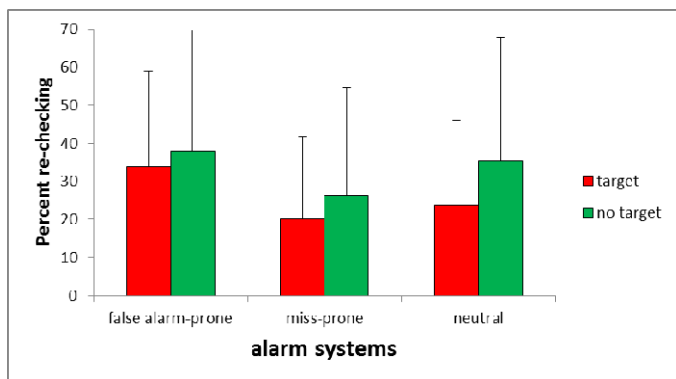

Figure 2. Means of re-checking frequencies (averaged across two experimental blocks) for the three alarm systems in the high risk condition.
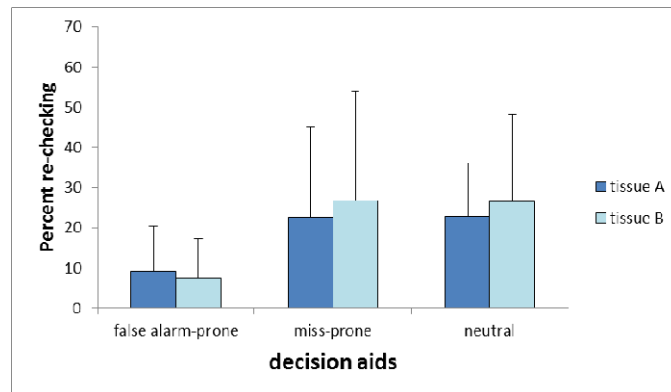

Figure 3. Means of re-checking frequencies (averaged across two experimental blocks) for the three decision aids in the low risk condition.

Participants in the false alarm-prone high risk condition frequently re-checked both, 'target' and 'no target' cues, showing an asymmetry effect, where PPV influenced re-checking of 'target' and 'no target' cues. As predicted, this effect disappeared in the low risk condition. Re-checking was done much less. Unlike predicted, no real difference between re-checks of 'target' cues and 'no target' cues was found.

Participants' interaction with the miss-prone system under the condition of high risk also corresponded with our expectations. As a consequence of lower NPV, participants re-checked the 'no target' cues more often than the 'target' cues. As expected, the pattern in the low risk condition was the same. The predicted reduction of re-check frequencies due to the lower risk was not found.

Similar, re-checking with the neutral system did not change as a function of risk. More surprisingly however, participants under both risk conditions re-checked the 'no-target'/'tissue B' cues more often than the 'target'/'tissue A' cues, even though PPV and NPV were the same.

DISCUSSION

The results of the current study should be interpreted with caution, and attention should be given to the different effects that brought about the overall behavioral pattern.

Participants' frequent re-checking of 'target' cues and 'no target' cues, likewise, in the false alarm-prone condition under high risk represents a type of asymmetry effect on the level of re-check behavior and thus, corresponds to prior findings of asymmetry (e.g. Dixon et al., 2007; Meyer, Wiczorek & Günzler, 2013; Rice & McCarley, 2011). The disappearance of the effect, as well as the strong reduction of re-checking in the low risk condition, provide evidence for the proposed moderation of the asymmetry effect through the level of perceived risk.

In the neutral group under high risk, the differences between re-checking frequencies for 'target' cues and 'no target' cues are not in line with predictions. Because the system had the same PPV and NPV, re-checking rates were predicted to be similar for both types of cues. The disproportion, in favor of the more frequent re-checking of 'no target' cues, might be the result of the vulnerability of the reliance discussed earlier. While a PPV of 0.8 is interpreted as

sufficient, a NPV of 0.8 might already be seen as too low, because missing a critical event is perceived as more problematic than making a commission error and erroneously reacting to an alarm. Similar effects have been found before in a way that few reduction of high NPV led to an over proportionally strong reduction in reliance (Manzey et al., 2014).

The unexpected difference between re-check frequencies for 'tissue A' and 'tissue B' cues in the neutral low risk condition may be due to the design we used. The stimulus material used in the current study may have interfered with the risk manipulation in the low risk condition. Whereas the cues (blue and light blue), as well as the labels ('tissue A' and 'tissue' B'), were chosen to make both options identical, the images differed with regard to one, maybe important detail. One type of image contained the number three, the other did not. While we framed the task as a categorization task, the stimulus material we used still corresponded to a signal detection task. Thus, the more frequent re-checks of the target absent images may represent the same aversion to commit omission errors described above.

However, it is not clear, why variation of risk did not reduce overall re-checking, either in the miss-prone or in the neutral condition. One possible explanation is that risk not only moderated the asymmetry effect, but that the effect of risk is also moderated by system characteristics. Only participants working with false alarm-prone systems were sensitive to a variation of the risk.

Further research is needed to answer the remaining open questions. The variation of risk conditions should be repeated with stimulus material consistent with the task. In addition, future studies have to design the systems with regard to a potential omission avoidance bias. A neutral system should have predictive values of at least 0.95 (see Meyer et al., 2013), to prevent confounding the two possible explanations for the asymmetry effect (i.e. impact of false alarms vs. vulnerability of reliance).

Results of the current study show the existence of an asymmetry bias also on the level of re-check frequencies. However, a more complex analysis of response frequencies, including also analyses of direct compliance and direct reliance, in addition to re-checks is needed to complete the picture.

This study does not provide a definite answer to the question whether the asymmetry effect is due to a disproportionally strong weighting of false alarms compared to misses or whether it is the result of a greater vulnerability of reliance, due to the tendency to avoid misses of critical events. However, it provides evidence in favor of the latter explanation, because 'no target' cues were re-checked more often.

REFERENCES

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, 38(11), 2300–2312.

Bliss, J. P., Jeans, S. M., & Prioux, H. J. (1996). Dual-Task Performance as a Function of Individual Alarm Validity and Alarm System Reliability Information. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 40, 1237-1241.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*(4), 564-572.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied*, 1(1), 19-33.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, 48(2), 241–256.

Manzey, D., Gérard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, 57(12), 1833-1855.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 46(2), 196-204.

Meyer, J., & Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors*, 44(3), 343–353.

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors*, 56(5), 840-849.

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

Parasuraman, R., Hancock, P.A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision warning systems. *Ergonomics,* 40(3), 390-399.

Wiczorek, R., & Onnasch, L. (2012). Development of a model predicting the use of automated decision aids. In D. Waard, N. Merat, A. Jamson, Y. Barnard, & O. Carsten (eds.), *Human Factors of Systems and Technology* (pp. 51–61). Maastricht, NL: Shaker Publishing.

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