Asymmetries in Human Tolerance of Uncertainty in Interaction with Alarm Systems: Effects of Risk Perception or Evidence for a General Commission Bias?

Torsten Guenzler, & Dietrich Manzey
Technische Universitaet Berlin, Berlin, Germany

Providing access towards raw data is often considered to be a good solution for improving human decision making in interaction with imperfect automated decision support such as alarm systems. However, there is some evidence that such cross-checking measures are used in an asymmetric manner with respect to the amount of uncertainty involved in the decision. Namely, people seem to accept low amounts of uncertainty when complying with an alarm cue, but not when contradicting it. The current study investigates the question whether this phenomenon is limited to alarm systems and a high risk environment. Within a multi-task PC simulation participants performed a low risk monitoring task which was supported by a system neutrally framed as "assistant system". In one group the cues emitted by the system were 90% correct, in the other 10% were correct, thus causing a 10% uncertainty about the real state in both conditions. Results show a strong asymmetry as participants in the latter condition spent a high amount of effort in reducing their uncertainty, while participants in the former condition did not. Furthermore participants' behavior almost exactly replicates the asymmetric cross-checking pattern found in a former study which employed a comparatively high risk monitoring task supported by an "alarm system". This supports the hypothesis that the observed commission bias represents a general phenomenon in the context of automated decision support, irrespective of the risk attributed to the environment and irrespective of whether the system represents an alarm system or not.

INTRODUCTION

Although automations such as decision support systems usually work according to clearly defined decision-criteria (e.g. specific thresholds in case of alarm systems), humans do not, at least not necessarily. In fact decision research has proven that humans show a variety of decision biases. That is the case particularly when uncertainty is involved in the decision situation, a setting which is also found in the interaction with decision support systems. Among the probably most prominent examples are alarm systems. On the one hand they provide indispensable decision-support in typical supervisory control tasks. On the other hand, however, they never provide perfectly reliable information but emit a certain number of false alarms and/or misses, depending on the specific design of their signal-detection characteristics. This usually leaves the user with some degrees of uncertainty about the validity of a given alarm cue. In case that operators have no possibilities to double-check a given alarm's validity towards other available data they usually rely on some sorts of response selection heuristic which have been referred to as extreme responding or probability matching (e.g. Bliss, 2003; Bliss, Gilson, & Deaton, 1995). However even more interesting is the case when operators have the possibility to verify a given alarm towards other data. This way they may effectively reduce the level of uncertainty before making their decision whether or not to follow an alarm. Nevertheless, it is yet not well understood how people deal with such opportunities to reduce uncertainty.

Gérard and Manzey (2010) followed the straight-forward hypothesis that double-checking behavior would directly depend on the level of uncertainty. They assumed that people would not check raw data if uncertainty is low, but make use of double-checking when uncertainty is high. In their study the level of uncertainty was manipulated by varying the PPV (positive predictive value) of an alarm within a range of .1 (10% of alarms correct) to .9 (90% of alarms correct) between groups. Note that the extreme PPVs – in this case .1 and .9 – cause the lowest levels of uncertainty while the medium PPV of .5 causes maximum uncertainty. Specifically, when a PPV=.1 alarm occurs it is 90% certain, that there actually is no critical event. Hence there is a 10% error risk if not double-checking such an alarm but ignoring it altogether. When a PPV=.9 alarm occurs, it is 90% certain that there really is a critical event. Therefore there is a similar 10% error risk if not double-checking such an alarm but directly following it. In case of a PPV=.5 alarm uncertainty is at a maximum of 50% and so would be the error risk without double-checking. Accordingly, Gérard and Manzey assumed that the checking rate would relate to PPV in an inverted u-shape. Interestingly, participants' decision behavior at the same time supports and contradicts this uncertainty hypothesis. In accordance to

the hypothesis participants double-checked most of the PPV=.5 alarms and rarely checked PPV=.9 alarms. In contrast to the assumptions however checking rates for the PPV=.1 alarm were as high as for the PPV=.5 alarm. So while participants most of the time accepted the 10% uncertainty in the PPV=.9 case without looking at the raw data they did not accept the 10% uncertainty in the PPV=.1 case. In the latter case they spend a high amount of effort in checking the raw data, thereby compromising their overall monetary outcome as supervising the alarm system was only part of their multi-task mission. It has to be noted that objective costs of *false alarms* and *misses* do not account for this asymmetry as the monetary payoff structure was symmetrical and both errors had the same costs.

While the asymmetry in participants' uncertainty tolerance is pronounced, the understanding of its causes thus far is not. Considering the experimental setting of the study there might be two possible explanations: the palpable risk of the cover story as well as the alarm context. In their instructions Gérard and Manzey presented the test environment (a multi task PC-Simulation) as part of a control room of a chemical plant, where participants had to monitor reaction chambers. Consequently one might argue that the asymmetric uncertainty reduction is a somewhat reasonable strategy to compensate for severe risks attributed to *misses* committed in chemical plants (e.g. contamination, explosion), because no such risks would apply to *false alarms*. This raises the question whether a similar decision pattern would evolve in a low risk environment.

Apart from the cover story the alarm wording could also play a crucial role. Participants had to monitor a detection aid which was presented as "alarm system". Alarms usually indicate potentially dangerous events. It might therefore be reasonable for users of such systems to assume that it is more dangerous to miss a critical event behind an alarm than to erroneously act on a false alarm. Hence, the monetary payoff which was equal for both, misses and false alarms, might have been blanketed by the combination of high risk environment and alarm context.

However, the asymmetry in uncertainty tolerance could also be a more basic issue of decision automation in general. In contrast to the famous omission bias in economics, where people refrain from taking an action, the asymmetry might point to a general commission bias in the interaction with decision aids. The mechanisms of which could be quite similar to what has been referred to as *action bias* in other decision making contexts (e.g. Bar-Eli, Azar, Ritov, Keidar-Levin, & Schein, 2007). That is, people might feel that they are expected to actively comply with recommendations by an automation, and that even making a false decision by complying with a system would be more tolerable than committing an error by ignoring the automated device. After all, users might feel more accountable for errors committed against a system's advice, because responsibility for errors committed in accordance with the system advice could be considered a shared responsibility and thus feel reduced. As a consequence, uncertainty tolerance should be higher for compliance with cues than for objection to cues. In this case the results found in the alarm research referred to above would reflect a sort of general commission bias in interaction with decision support systems instead of a specific effect in responses to alarms in high risk contexts.

The current study was conducted in order to explore whether the asymmetry reported by Gérard and Manzey (2010) also emerges in a context where 1) the perceived risk is much lower and 2) a neutral framing is used for the automated decision support system. For this purpose the test environment of Gérard and Manzey (2010) was reused, but with a different framing. The "monitoring of reactions in a chemical plant" became the "monitoring of the labeling process in a brewery" and the "alarm system" was now described as an "assistant system".

METHOD

Participants

A sample of 42 students participated in the experiment. The data of four participants had to be excluded from the analysis, twice due to technical problems during the experiments and twice due to problems with the comprehension of the experimental task. Hence, the data of 38 participants (20 male, 18 female, age: 19-35 years) was included in the analysis. Participation was compensated by 7 Euro (about 9.50 US$) plus a performance-based bonus of up to 15 Euro (about 20 US$).

Apparatus and Tasks

The same laboratory multi-task environment M-TOPS (Multi-Task Operator Performance Simulation) as in the study of Gérard and Manzey (2010) was used for the investigation. In this task, participants have to work on either two or three tasks concurrently, one of which is representing a monitoring task that is supported by an alarm system. The tasks were developed to simulate basic work demands of control room operators. However, for the present experiment the cover story around this task was modified to provide participants a different framing of the context and the automated

system they work with. Instead of the usual chemical plant cover story participants were now instructed to operate in the control room of a brewery. Furthermore the alarm system was now introduced as an assistant system helping in classifying the brew.

The user-interface is shown in figure 1. It was exactly the same for both studies apart from some minor changes in terms of technical wording and the corporate logo. Furthermore also the logic and procedures of the tasks were similar to the tasks presented in Gérard's and Manzey's study.

Resource Ordering Task (ROT). This task was situated in the upper left quadrant of the interface. Basically it represented a mental arithmetic task. Participants were instructed that they always have to assure the availability of a specific amount of required ingredients to keep the brewing process running. For this purpose the actual and the overall required amount of an ascertained ingredient were given. Participants then had to calculate the difference, type the result in the designated ordering field, and send the order by clicking the respective button below. This would irreversibly finish the ordering process. Participants had 15 seconds to respond to a given request. After 15 seconds a new request appeared, irrespective of the status of the preceding one. However, participants could speed up the sequence by actively initiating a new trial via mouse-click on the "arrow" button (upper right). Each correct order was worth a bonus of 0.06 Euro. There was no penalty for incorrect or missed orders.

Coolant Exchange Task. This task, presented in the top-right quarter, was not part of the current study and could be ignored by the participants. This is analogous to the configuration of Gérard and Manzey (2010).

Monitoring Task (MT). This task was displayed on the lower-right quadrant. Here participants had to monitor the value of the wort concentration as this determines the type of beer brewed – e.g. in terms of taxation – and therefore the labeling needed. Above a certain value labeling needed to be changed from the default *Vollbier* to *Starkbier* (unfortunately these rather technical terms for two different categories of beer have no English translation). Brewing tanks were presented in a serial manner and automatically analyzed by an automation which in this case was called an *assistant system*. Participants were instructed that the system would inform about whether the wort concentration detected from the system was below (no labeling changes needed) or above a certain threshold (labeling changes necessary). The assignment of a tank to the Starkbier labeling line needed to be done manually by clicking on the respective button on the lower left. Participants had the possibility to inspect the raw data (i.e. the actual wort concentration) directly, thereby reducing uncertainty upon a given system cue to zero. However, the double-checking procedure was somewhat tedious and time consuming. First the denomination of the corresponding brewing tank had to be selected in a drop down list of quite similar tank denominations and afterwards wort concentration had to be determined by counting red areas in a pop up picture. Participants had a time window of 10 seconds for each presented tank in which they had to decide whether to directly comply with the automation, ignore it or first double-check the raw data. The system would proceed to the next brewing tank after this time window had elapsed or after the participant had entered a decision. For each correctly labeled tank 0.08 Euros were added to the bonus, for each incorrectly labeled tank 0.08 Euros were subtracted. That means a bonus was added or subtracted for every trial, even if there was no input from the participant.

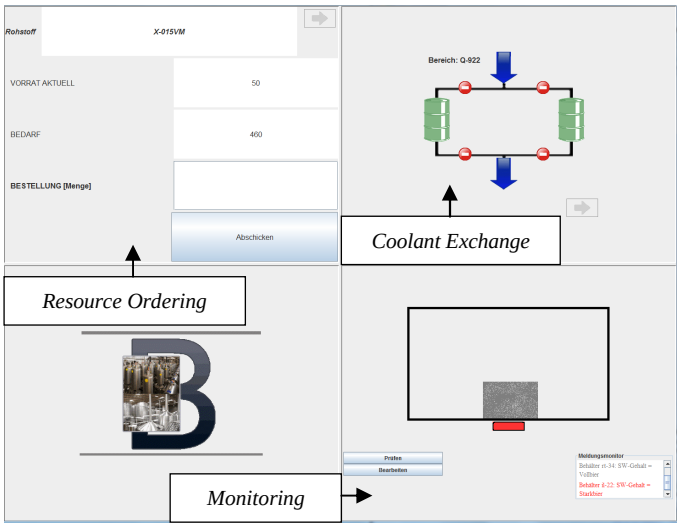

Figure 1. User interface of M-TOPS.

Design

The study comprises a one factorial design. The factor represents two different levels of PPV manipulated between subjects, with one group being supported by a PPV=.1 system and the other by a PPV=.9 system.

Dependent variables

Behavioral measures for both tasks were derived from log-files which contain the complete input of each participant.

As this study investigates the reaction towards alarm type of cues it has to be noted, that the *Starkbier cue* corresponds to the alarm cue in signal detection terms. Analysis will thus focus on the reaction to Starkbier cues. It was assessed to what percentage such cues were

ignored (*ignore ratio*), double-checked (*double-check ratio*) or directly followed (*comply ratio*). Any activity directed towards the analysis of relevant system data was tagged as double-checking behavior. Note that therefore "ignore", "double-check" and "comply" mark the complete scope of possible reactions towards a cue.

Performance in the *ROT* was assessed by the number of correct orders. Furthermore subjective measures were collected for manipulation check purposes.

Procedure

Participants were welcomed and randomly assigned to one of four computer work stations. They were informed about the experiment system, the tasks and the payoff structure by instructions presented on the screen. This way they were all confronted with the same brewery cover story and neutral assistant system framing in a standardized manner. They performed a 120s training block with the *ROT*, a short 180s training block with the *MT* for comprehension of the general functionality and again a 100 tank *MT* training block for getting to know the reliability of the system cues. Then they answered a questionnaire where they were first asked and afterwards informed about the performance contribution of their assistant system. Subsequently participants performed the 800s experimental block working on both tasks at the same time. After answering a final questionnaire participants where paid and thanked for their participation. An experimental session took between 1.5 and 2 hours.

RESULTS

Manipulation Check

In accordance with the manipulation of the PPV between groups it was ascertained, whether the different PPVs were reflected in participants' perception of the assistant system. Perceived PPVs were calculated with the help of participants' estimated quantity of Hits and False Alarms. As expected, participants in the PPV=.1 condition perceived their system's PPV to be significantly lower (M = .22, SD = .17) than participants in the PPV=.9 condition (M = .75, SD = .15), $t(36) = -10.04$, $p < .001$. Furthermore it was ascertained, that the equal amount of uncertainty generated by the two antipodal PPVs was reflected in participants' perception of the respective uncertainty. Uncertainty was calculated with the help of perceived PPVs, and was overestimated in the PPV=.1 (M = .19, SD = .10) as well as in the PPV=.9 condition (M = .23, SD = .12). Consistent with the manipulation groups did not differ in the subjective amount of uncertainty caused by the two differently reliable Starkbier cues, $t(36) = -.96$, $p > .20$.

The brewery cover story was intended to be comparatively neutral in terms of risk perception. Comparing a brewery (M = 2.76, SD = 1.08) to a chemical plant (M = 5.32, SD = .66) on a six point scale participants indeed reported to attribute significantly greater risks to the latter, $t(37) = -13.90$, $p < .001$.

The "assistant system" framing was intended to eradicate the alarm context from the M-TOPS paradigm. Indeed only two out of 38 participants reported their system to be an alarm system.

Behavioral Measures

Monitoring task. The different responses to the Starkbier cues are shown on the left of figure 2. As expected, the most frequent response in the PPV=.9 condition was the direct compliance with the cue (M = 62.7%, SD = 30.8%). In this condition only every third cue was double-checked by participants (M = 32.2%, SD = 31.2%). In accordance with the high PPV almost no cues were ignored (M = 5.1%, SD = 5.5%). As predicted a different pattern emerged in the PPV=.1 condition. About one fifth of the Starkbier cues were ignored (M = 18.9%, SD = 29.1%) and virtually none directly followed (M = 0.2%, SD = 0.7%). However, in this case double-checking was by far the most frequent behavior (M = 80.1%, SD = 29.0%). Interestingly this pattern almost replicates the results of Gérard and Manzey (2010) which are depicted on the right of figure 2. In order to analyze the statistical significance of the asymmetry in double-checking behavior both groups were compared with a *t*-test. In accordance with the hypothesis participants who were

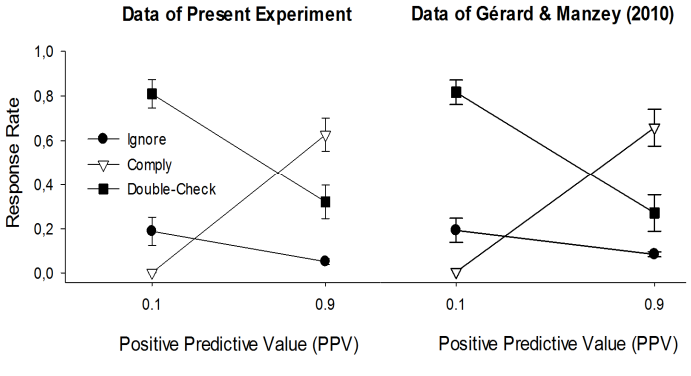

Figure 2. Means and standard errors for the response rates to system cues indicating a Signal, depending on PPV. Left: Low risk cover story and neutral system framing in the current experiment. Right: high risk cover story and alarm system framing in the former study.

supported by the PPV=.1 system double-checked significantly more Starkbier cues than participants working with PPV=.9 system, $t(36) = 4.99$, $p < .001$.

Ordering task. Participants in the PPV=.1 condition in average sent 69.0 correct orders ($SD = 16.3$) compared to 60.5 correct orders ($SD = 18.7$) sent by participants in the PPV=.9 condition, $t(36) = 1.48$, $p = .15$.

DISCUSSION

This study investigated the phenomenon of asymmetric human tolerance of uncertainty in the interaction with alarm systems. Specifically, users do rarely accept a small amount of uncertainty when it is very likely that the alarm is false (low PPV), but they do accept a similar amount of uncertainty when it is very likely that the alarm is correct (high PPV). That means in the former case users do double-check the raw data behind an alarm, in the latter case they mostly abstain from doing so, as reported by Gérard and Manzey (2010). A potential explanation was that the high risk environment – Gérard and Manzey presented a chemical plant cover story – renders subjective costs of *misses* higher than costs of *false alarms*, thereby causing the asymmetry. Another explanation was that the alarm context already causes asymmetry. By attribution the word alarm might point to a critical state which is comparatively dangerous or costly to be missed.

It was however assumed that the asymmetry phenomenon would rather represent a basic phenomenon in the context of automated decision support, irrespective of risk attributed to cover story or alarm wording. To test this hypothesis, participants of the current study were presented a neutral cover story and neutral system framing for the same task as in Gérard's and Manzey's study.

As expected, participants in the PPV=.1 group exhibited a significantly lower uncertainty tolerance than participants of the PPV=.9 group. Even more interestingly there were almost no differences to the behavioral patterns reported by Gérard and Manzey (2010). This strongly indicates that the asymmetry in uncertainty tolerance is a general problem in the interaction with decision support systems. Although a lot remains to be learned about the causal mechanisms of the phenomenon, the results might point to a general commission bias in the decision support domain. In accordance with findings about diffusion of responsibility (Milgram, 1963; Latané, & Darley, 1968) there might be a tendency to be less diligent when responsibility is shared, in this case shared with the system. Hence there might be a stronger reluctance to accept uncertainty when the responsibility for the decision cannot be shared. While in both cases, the PPV=.1 case as well as the PPV=.9 case, uncertainty is similar, subjective responsibility in case of an error might not be so. In both cases one would commit an error in 10% of the decisions by accepting the uncertainty and making a decision without checking the raw data. However, in the PPV=.9 case the decision would be in accordance with the system, i.e. responsibility for errors is shared. In contrast, in the PPV=.1 case the decision would be against the system, i.e. responsibility for errors is not shared.

Further research should therefore investigate the role of subjective responsibility for errors in the decision support context. A better understanding of the psychology behind biases such as the asymmetry in uncertainty tolerance will be crucial for a more comprehensive design not only of alarm systems but of decision support systems in general.

REFERENCES

Bar-Eli, M., Azar, O. H., Ritov, I., Keidar-Levin, Y., & Schein, G. (2007). Action bias among elite soccer goalkeepers: The case of penalty kicks. *Journal of Economic Psychology, 28*(5), 606-621.

Bliss, J. P. (2003) An investigation of extreme alarm response patterns in laboratory experiments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2003, 47*, 1683-1687.

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

Gérard, N., & Manzey, D. (2010). Are false alarms not as bad as supposed after all? A study investigating operators' responses to imperfect alarms. In D. de Waard, A. Axelsson, M. Berglund, B. Peters & C. Weickert (Eds.), *Human Factors: A system view of human, technology and organisation (pp. 55 - 69)*. Maastricht, The Netherlands: Shaker Publishing.

Latané, B., & Darley, J. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377-383.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology 67*(4), 371–378.