On the Relation Between Reliance and Compliance in an Aided Visual Scanning Task

Rebecca Wiczorek[1], Joachim Meyer[2], & Torsten Guenzler[1]
[1]Berlin Institute of Technology, Berlin, Germany
[2]Ben-Gurion University of the Negev, Beer Sheva, Israel

Alarms, alerts, and other binary cues affect user behavior in complex ways. One relevant distinction is the suggestion that there are two different responses to alerts – compliance (the tendency to perform an action cued by the alert) and reliance (the tendency to refrain from actions as long as no alert is issued). An experiment tested the dependence of the two behaviors on the Positive and Negative Predictive Values of the alerts (PPV and NPV) to determine whether these are indeed two different behaviors. Results suggest that the compliance is relatively stable and unaffected by irrelevant information (the NPV), while reliance is also affected by the PPV. The results are discussed in terms of multiple-process theories of trust in information sources.

INTRODUCTION

Responses to binary alerts, alarms and dynamic warnings have attracted considerable interest in recent years. One issue that arises in this context was the question whether the trust in such systems is a single entity or whether there are actually two different forms of trust in binary cues (Meyer, 2004). One of them is *compliance*, which is the degree to which the binary cue, when it is present, causes the operator to act in accordance to the cue. Another type of trust, referred to as *reliance*, is the degree to which operators dare to avoid taking precautions when the binary indicator does not point to a signal.

The question whether reliance and compliance are indeed two separate responses or whether they are two expressions of a single trust entity has been the subject of some recent research (Dixon & Wickens, 2006, Dixon, Wickens & McCarley, 2007; Rice, 2009; Bahner, Elepfandt & Manzey, 2008; Rice & McCarley, 2011). These studies looked at differential effects of various variables on the two responses. Namely, compliance should be affected by the likelihood that a cue indeed indicates a signal (i.e., a malfunction or other problem). This variable is the Positive Predictive Value (PPV) of the cue, which is the probability that there is a signal, given that there was an alert (p[Signal|Alarm]). It depends on the probability that a cue appears, given that there is a signal (pHit in terms usually used in Signal Detection Theory [SDT]), the probability of a cue, given that there is no signal (pFalseAlarm in SDT), as well as on the prior probability of a signal (pSignal in SDT). From these probabilities one can use Bayes Theorem to compute PPV. Similarly, one can compute the Negative Predictive Value (NPV), i.e. how likely is it that there is no signal when no cue was given (p[noSignal|noAlarm]).

Whereas the PPV decreases with an increase of false alarms, the NPV depends on the number of misses generated by the cuing system.

An analysis of the normative responses to cues (Meyer, 2004) shows that compliance, i.e., the tendency to act as if there is a signal when a cue was issued, should depend on PPV and should be independent of NPV. Reliance, i.e. the tendency to act as if no signal exists when no cue was issued should depend on NPV and should be independent of PPV. Hence, the trust in the information from the cue when it indicates a signal should be independent of the trust in the cue when it indicates that no signal exists.

In previous studies (e.g. Dixon, Wickens, & McCarley, 2007; Rice & McCarley, 2011) the authors showed that automation false alarms lowered compliance and reliance, while automation misses only affected reliance. It was shown that the effect of the two failures was not symmetrical, as PPV influenced both types of trust. In these studies the PPV and NPV values were extreme (PPV = 1 and NPV = 1 respectively). Hence, the asymmetric patterns found in former studies might have been due to the partly perfect automation.

The current study aims to investigate whether the same asymmetric effects exist when both types of system failures can occur and influence trust simultaneously. We therefore varied PPV and NPV in a balanced manner without using extreme (PPV or NPV = 1) values. Conditions differed in one characteristic (the PPV or the NPV), while the other was kept constant. This allows us to assess the independence of the two responses in a controlled setting – in this case a simulated inspection task, resembling the visual inspection of images for signals, as in airport luggage scanning.

One explanation that was proposed for the asymmetry in the effects was that perhaps false alarms are more easily detected than misses. Therefore operators will base their response to the automation more on the variable that depends on false alarms (PPV) than on the variable that depends on misses (NPV), and therefore PPV will have a stronger effect than NPV (Rice & McCarley, 2011). By trying to maintain similar levels of salience for both misses and false alarms, it may be possible to assess this hypothesis. We account for the issue of perceptual asymmetry by keeping misses and false alarms equally salient and providing feedback after each trial.

Three possible patterns of results can appear in our experiment:

1. The two responses may be independent (i.e., the response to a given PPV will be unaffected by NPV, and the response to a given NPV will be unaffected by PPV).
2. Both responses are related (i.e., the response to a given PPV will also depend on NPV and the response to a given NPV will also depend on PPV).

DOI 10.1177/1071181312561060

3. There is an asymmetry of the effects of NPV on PPV and PPV on NPV, namely PPV (due to false alarms) will affect responses to a given NPV level, while NPV (due to misses) will have no effect on responses to a given PPV level.

An additional difference between our study and most previous studies on reliance and compliance was in the variable we used to measure the types of trust.

The computational framework used for computing reliance and compliance here is Signal Detection Theory (see Swets, Dawes & Monahan, 2000, for a description). The tendency to respond to a cue and to identify an event as a signal or not was evaluated through the response bias or threshold settings operators use with and without cues. A lower threshold setting means that people had an increased tendency to declare that a signal exists. In our study we also included blocks in which no cues were issued and computed the threshold setting (c) there, too, where c is defined as

$$C = -0.5(Z_{pHit} + Z_{pFalse\ Alarm}) \qquad (1)$$

As pointed out in a study on physicians' responses to clinical reminders (Vashitz et al., 2009), compliance (and reliance) should ideally be computed relative to responses when no cues are available. Thereby, allowing not only comparisons of two different systems but also to achieve information about the absolute amount of compliance. A compliance value of 0 means that there is no difference between users' response bias for alerts and without alerting system indicating a total lack of compliance with the cues.

We computed the measure of compliance by subtracting the threshold with a cue indicating a signal (“Alert”) from the threshold when no cue was available (baseline):

$$\text{Compliance} = C_{Baseline} – C_{Alert} \qquad (2)$$

And similarly, reliance was computed subtracting the threshold when no cue was available from the threshold with a cue indicating the absence of a signal (“No-Alert”):

$$\text{Reliance} = C_{No\text{-}Alert} – C_{Baseline} \qquad (3)$$

This computation created two measures that should be positive if operators expressed either reliance or compliance.

METHOD

Participants

Sixty undergraduate students from Ben-Gurion University of the Negev participated in this study. Participants had normal or corrected-to-normal vision and reported no color vision deficiency. Participants received 30 ILS (Israeli Shekels, about $8.4) for their participation and also took part in a lottery of 4 times 100 ILS (about $28). Each performance score represented a lottery ticket for this lottery so that participants had an incentive to collect as many points as possible.

Task Environment

Participants performed a visual scanning task. Participants viewed monochrome images showing a 3x3 matrix of single digit numbers, displayed on 19" screens. It was difficult to identify the digits because 39% of the pixels were inverted, resulting in a blurred image as shown in Figure 1. Pictures had to be classified according to the presence or absence of the target digit 3 by clicking on either of two buttons labeled “threat” or “no threat”. Images were presented for 2 seconds and were not repeated. In the blocks in which cues were provided participants saw either a red or a green cue 2 seconds before the image appeared. These cues indicated a “threat” or “no threat” diagnosis, respectively. The reliability of the cues differed between the experimental conditions. After participants responded, they were asked how confident they felt about their decision. After each trial visual feedback was given, informing the participant about the correctness of their decision.

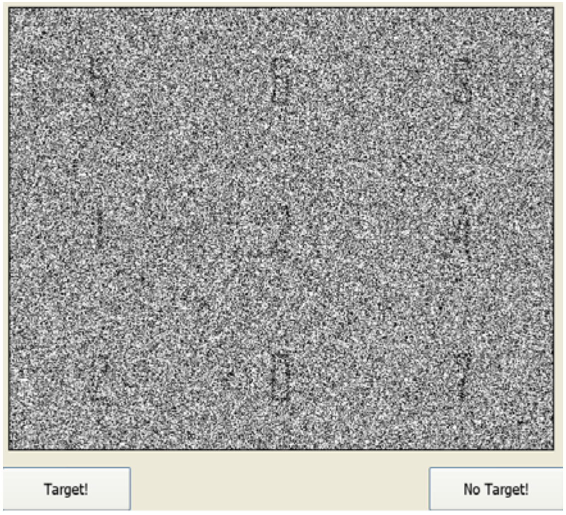

Figure 1: Example of a stimulus.

Procedure

Participants were randomly assigned to one of the four experimental conditions and one of six computer work stations. After reading the instructions, participants performed the experimental task, consisting of 280 trials. The task was subdivided into 4 blocks. The first and the third block served for baseline measurements and consisted of 40 trials each. During these trials, participants were not supported by the cues. The second and the fourth block consisted of 100 trials each, and in them performance was aided by binary cues, indicating the presence or absence of a target. In all blocks, 50% of the trials were "signals" (i.e., in 50% of the trials the target digit 3 appeared on the screen). Participants received 10 points for each correct response and lost 10 points for each incorrect response. At the end of the experimental session, participants were debriefed and were paid.

Experimental Design

The study consists of two complementary 2x2 designs. In one part of the study, NPV was manipulated while PPV was kept constant, whereas in the other part PPV varied between groups with the same NPV for both conditions. Block served as a within-subject factor in all conditions. NPV and PPV of the diagnostic support system were varied symmetrically. In the “NPV.75” condition 75% of the green cues where correct while 95% were correct in the “NPV.95” condition. In both conditions, PPV (percentage correct during red trials) was held constant at .90. The levels of the PPV factor were manipulated similarly. The system’s NPV was held constant at .90, while PPV was .75 vs. .95 in the two PPV conditions. Characteristics of the cueing systems used in the different conditions are shown in Table 1.

Table 1: SDT parameters of the four different cueing systems

	PPV	NPV	c	d’
NPV.75	.9	.75	0.44	1.93
NPV.95	.9	.95	-0.23	3.03
PPV.75	.75	.9	-0.44	1.93
PPV.95	.95	.9	0.23	3.03

Measures

Sensitivity, as a measure of the task performance, was quantified through the d' value (cf. SDT) of overall performance, which is a measure of the combined sensitivity of the user from observing the image and using the information from the cue when it was given.

Participants’ compliance and reliance measures were computed as described above. A value of 0 means, that there is no compliance or reliance, while larger values indicate more compliance or reliance.

RESULTS

Sensitivity

We analyzed the sensitivity scores separately for the two groups with .75 and .9 for PPV and NPV, respectively, and for the two groups with .95 and .9. This separation was necessary, because the maximal possible d' is much larger when the NPV and PPV values are both equal or above .9 than when one of the values is .75. As shown in Table 1, the predicted d' of decisions that are only based on the cue are 1.93 for the PPV = .9, NPV = .75 and PPV = .75, NPV = .9 conditions, respectively, and the predicted d' values for the PPV = .9, NPV = .95 and PPV = .95, NPV = .9 conditions are 3.03.

In the two-way ANOVA of d' for the conditions with .75, with the condition and the block as independent variables, neither the effect of the condition, *F*(1, 28) = .08, *p* = .78, nor the interaction Condition x Block were significant, *F*(3, 84) = 1.23, *p* = .30. There was, however, a significant main effect of the block, *F*(3, 84) = 7.03, *p* = .0003, with d' values of 0.98, 1.64, 1.41 and 1.46 for the four blocks. Since blocks 1 and 3 were not aided by any cues, we should expect lower d' values for them. In fact, the d' in block 1 was indeed lower than in the other blocks, but there was no significant difference between blocks 3 and 4. Thus, when the cueing system has only limited validity in one of its indications, performance of users with some experience with the task did not significantly benefit from receiving these cues.

The two-way ANOVA for the groups with the .95 PPV or NPV also showed only a significant main effect of the block, *F*(3, 84) = 62.63, *p* < .0001. Here the sensitivity in the two blocks in which cues were available was clearly higher than the sensitivity in the blocks in which no cues were available (with d' values of 0.88, 2.05, 1.18 and 2.39 for the four blocks). Thus performance in blocks 2 and 4 (when alerts were given) was clearly superior to performance in blocks 1 and 3.

It should be noted that in all four conditions, the d' values for users who could rely on the combination of the image and cue for making the decision did not reach the sensitivity level that could have been attained if participants would have responded only to the cues (compare Table 1.). Thus one can argue that participants showed insufficient trust in the cueing system, in the sense that more trust in this system could have helped them attain better discrimination performance.

Compliance and Reliance

We analyzed compliance and reliance with two-way ANOVAs with either NPV or PPV as a between-subject factor and Block as a within-subject factor.

Results for the analysis of Compliance are shown in Figure 2. All values are significantly larger than 0 (as can be seen by comparing the standard error whiskers to the 0 value). None of the effects in the analysis were significant.

The results indicate that compliance is fairly stable and does not depend on the (irrelevant) NPV values.

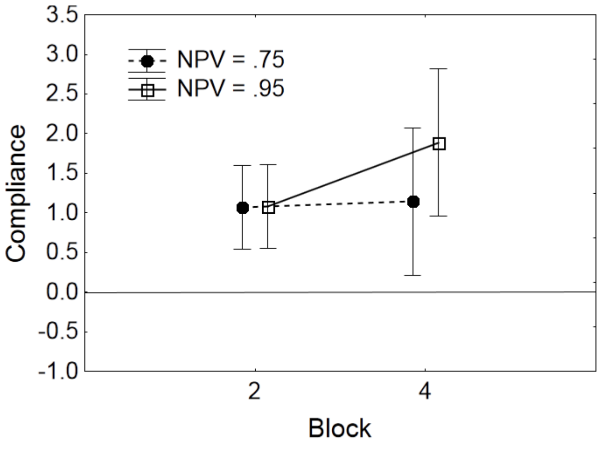

Figure 2: Compliance value for Block 2 and 4 for conditions with NPV = .75 and NPV = .95.

The parallel analysis of reliance showed somewhat different results, as depicted in Figure 3. Here we found a significant main effect of the PPV value, *F*(1, 28) = 9.90, *p* = .004. When

PPV = .75, there was no reliance. It seems that in this case, the system seemed to be so unreliable that participants did not use it to determine whether a situation was safe when no cue was issued. When PPV = .95 there was evidence for reliance. Since neither the main effect of the block, $F(1, 28) = .15$, $p = .70$, nor the interaction, $F(1, 28) = 1.40$, $p = .25$, were significant, it seems that the reliance must have developed very quickly.

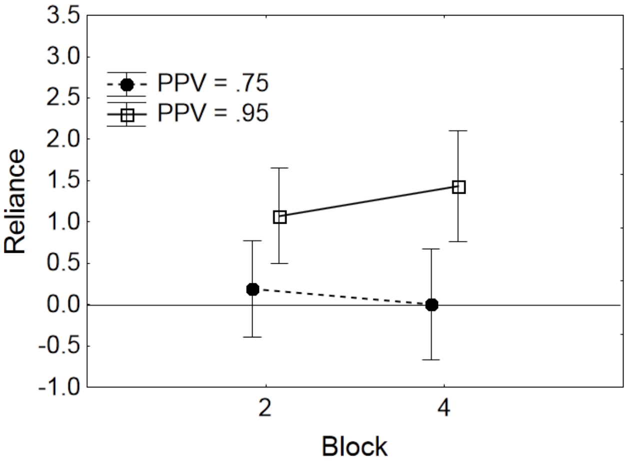

Figure 3: Reliance value for Block 2 and 4 for conditions with PPV = .75 and PPV = .95.

DISCUSSION

We present the results of an experiment that assessed user responses to different cues in a visual scanning task. Our results show that discrimination performance in this task does not reach the level of performance that could have been attained if the task would have been done automatically (i.e., if the alerting system would have made the categorization decision). Involving the human operator in the task did not contribute towards better performance of the task. This was mainly due to the fact that the operator gave insufficient weight to the information from the cue. This was true when the cue was only partly diagnostic (when NPV or PPV were .75), but it was also true when NPV or PPV were .95.

Our study was mainly intended to assess the relation between the two aspects of trust in a system, identified by Meyer (2004) - the users' compliance with the system and the reliance on it. The two aspects could theoretically be two expressions of the same underlying trust. They could be two entirely independent types of trust. Finally, they can be somewhat related, with either affecting the other or with only one affecting the other.

Our results support the last possibility. Differences in NPV had no effects on compliance. Apparently the only factor that affected the operators' response was the question how well the system could be trusted when it issued an alert. Thus, it seems that the response to the alert, as expressed in compliance, is the primary response.

In contrast, reliance was affected by PPV. In fact, when PPV=.75, there was no evidence for reliance. Participants in this condition apparently did not use the cues indicating the absence of a signal at all for deciding if a situation was intact.

Reliance became evident when the PPV was high. Thus, for a user to trust a system when it indicates that everything is fine, the user seems to consider the likelihood that the system detects a failure if one exists.

It seems therefore, that the distinction between reliance and compliance has some value. These are not two expressions of the same underlying trust in the system. However, the two types of trust are not independent. It seems that a relatively high level of PPV is a precondition for reliance. Compliance, however, does not require a relatively high level of NPV.

It is not quite clear what causes this asymmetry in the responses to information. One proposed explanation (Rice & McCarley, 2011) suggested that the asymmetry may be due to the greater salience of false alarm events. Thus people are more aware of the failures of the system when it issued an alert when there was no signal and may not notice system failures when the system failed to detect a signal. This explanation is particularly plausible when only some of the events are accompanied by a cue (the alert is presented only when a signal is detected, otherwise it is silent).

In our study the different cues were visible in all trials in blocks in which alerts were available, either as a red or a green indicator. Additionally, visual feedback was provided after each trial, informing participants about system failures. Thus, at least from the perspective of information presentation, the conditions with and without cues were not different. Hence, the salience hypothesis, at least in its simpler form, does not account for our findings.

It might, however, be possible to maintain a salience hypothesis that is based on the apparent salience, rather than on the actual one. It is possible that participants for some reason attend more closely to events for which an alert is issued compared to events for which no alert is issued. Consequently, failures in the former type of events may be more vivid and affect responses stronger than failures in the events when no alert is given.

An alternative explanation may be that people are more afraid of causing misses than false alarms. As a result the reliance component may be more vulnerable to any type of system failures and peoples' reliance decreases whenever their trust in the system diminishes, irrespective if it is due to an insufficient NPV or PPV. Gérard and Manzey (2010) found that reliance decreased disproportionally strong when the system provided only a few misses, whereas a larger decrease in PPV was necessary to lower compliance.

Future research should address these issues. Still, our findings provide new insight into the scope of the phenomenon. The asymmetry between reliance and compliance is not limited to partly perfect systems (where either the PPV or the NPV is perfect). This pattern rather seems to develop in both systems which commit only one type of error as well as in systems which commit both types. Also, reliance seems to be a less robust phenomenon than compliance, and for it to appear, the system needs to have a fairly high level of validity.

REFERENCES

Bahner, J.E., Elepfandt, M. & Manzey, D (2008). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. *Proceedings of the 52nd Meeting of the Human Factors and Ergonomics Society, New York, September 2008.* Santa Monica: HFES.

Dixon, S.R. & Wickens, C.D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48(3)*, 474–486.

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.

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. Weikert (eds.), *Human Factors. A system view of human, technology and organisation* (pp. 55-70). Maastricht: Shaker Publishing.

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

Rice, S. (2009). Examining single- and multiple process theories of trust in automation. *The Journal of General Psychology, 136*(3), 303-319.

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*(4), 320-331.

Swets, J.A., Dawes, R.M. & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26.

Vashitz, G., Meyer, J., Parmet, Y., Peleg, R., Goldfarb, D., Porath, A. & Gilutz, H. (2009). Defining and measuring physician's responses to clinical reminders. *Journal of Biomedical Informatics*, 42, 317-326.