Taking Advice from Intelligent Systems: The Double-Edged Sword of Explanations

Kate Ehrlich, Susanna Kirk, John Patterson, Jamie Rasmussen, Steven Ross, Daniel Gruen
IBM Research
1 Rogers St, Cambridge, MA 02142
{katee, sekirk, john_patterson, jrasmus, steven_ross, daniel_gruen}@us.ibm.com

ABSTRACT
Research on intelligent systems has emphasized the benefits of providing explanations to accompany recommendations. But can explanations lead users to make incorrect decisions? We explored this question in a controlled experimental study with 18 professional network security analysts doing an incident classification task using a prototype cybersecurity system. The system provided three recommended choices for each trial. The choices were displayed with explanation (called "justifications") or without. On half the trials there was a correct choice amongst the three and the other half there was no correct choice. Users were more accurate when there was a correct choice. Although there was no overall benefit of explanation, we found that a segment of the analysts were more accurate with explanations when a correct choice was available but were less accurate with explanations in the absence of a correct choice. Based on our analysis of these results we conclude that explanations may be perceived as compelling for reasons other than just supporting interpretation. We discuss implications of these results for the design of intelligent systems.

Author Keywords
Intelligent systems, recommendations, explanations, adaptive agents, evaluation, individual differences,

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

CHI 2009, April 4–9, 2009, Boston, MA, USA.

INTRODUCTION
Intelligent systems help users make decisions in a variety of domains. Consider, for example, a network security analyst who is faced with diagnosing and recommending remedial action for a network intrusion within a short time. An intelligent system can help the network security analyst by speeding up the analysis, or allowing the system to automatically handle cases where it has sufficient confidence in the automated diagnosis. Or consider financial professionals who might use an intelligent advisor that analyzes large complex financial models to make investment recommendations [6]. In these mission critical settings, the user needs to trust the system but also to have access to the reasoning behind suggestions to determine whether or not to accept the provided suggestion.

Numerous studies have demonstrated the benefits of explanations in intelligent systems [18], including building user trust [8, 14], supporting the evaluation of system conclusions [9, 13, 21], increasing transparency of system reasoning [6] and making suggestions more compelling [10]. However, many of these studies have been done in the context of consumer behavior where the consequences of a user following an incorrect recommendation are not serious. On the other hand, studies of intelligent systems in mission critical settings that have also examined the impact of providing erroneous information [11, 22] have been limited to suggestions without also examining explanations.

This paper addresses correct and incorrect recommendations and explanations, in a mission-critical setting. The key issue is whether explanations of system reasoning make it easier to detect erroneous suggestions by, for instance, letting users discover flaws in the system's reasoning, or whether explanations compel users to select a suggestion provided, even if none are correct. To address these issues we conducted a controlled empirical study in which we systematically varied the accuracy of suggestions and whether suggestions were also accompanied by explanations. The study was carried out with network security analysts who need to respond quickly and accurately to alerts generated by intrusion detection systems. The present study, which addresses issues of explanation and accuracy is an extension of a previously

published paper that focused principally on display characteristics with the same group of analysts [15].

The paper is organized as follows. After presenting related research we describe the setting in which our study took place and the prototype we developed for testing purposes. We then describe the design of our study and its results and conclude with a discussion of the broader implications.

PREVIOUS RESEARCH

Benefit of explanations

It is well established that user interaction with intelligent systems can be facilitated by explaining system reasoning in a way that users find intelligible [4, 9, 13]. Studies have explored a range of benefits related to providing explanations; many focusing in particular on the advantage of earning users' trust and confidence in the systems [8, 14, 17, 21].

Wang and Benbasat [21] conducted research on the benefit of different types of explanations (how, why and trade-off) for enhancing different types of beliefs. They found that explaining the "how" of the system supported competence and benevolence beliefs, "why" explanations increased benevolent beliefs, and "trade-off" descriptions increased integrity beliefs. Their "how" and "why" distinction maps to the difference other researchers have identified between "explanations" and "justifications" [20].

Several researchers have touted the importance and benefits of transparency. For instance Fleischmann and Wallace argue that it is imperative, for ethical, political and legal reasons, as well as improved judgment, for intelligent decision support systems to provide users with access to the underlying models [6, 7].

Response to erroneous recommendations

There can, however, be negative consequences of being too trusting in a system, especially when the system provides erroneous suggestions. Most of the research with respect to erroneous recommendations has focused on how explanations can support users in establishing appropriate expectations such that they derive value from good recommendations and tolerate or provide effective feedback concerning poor recommendations. If intelligent agents are to be effectively deployed in arenas where decisions are more consequential, it will be necessary to explore whether and how explanation facilities may support users not only in tolerating poor recommendations, but in accurately discriminating between correct and incorrect suggestions.

How might explanations affect acceptance of inaccurate recommendations? In one study with medical practitioners on recommendations from a case-based decision support system [3], the researchers included an inaccurate recommendation with a poor explanation in each of the test conditions in order to test the degree to which users were providing considered answers. The results indicated that assessment of the inaccurate predictions was significantly lower than for the accurate ones. However, the intention of the study was to examine whether users were "paying attention" during the test rather than to systematically study the effect of inaccuracy in the user interface.

For systems designed to support decisions such as media consumption or consumer purchases, the consequences of a poor decision is not serious. However, systems which support mission-critical tasks have more responsibility to their users. It is often assumed that users of these systems will tend to be less trustful and more reluctant to accept intelligent system support [10]. In either case, the key objectives of earning trust in good recommendations and overcoming potentially negative reactions to bad recommendations has led to a focus on the role that explanations may play in persuading users of the soundness of reasoning or helping users develop appropriate understanding of perceived reasoning limitations [10, 18].

So, while the support of effective decision making is an acknowledged objective of explanation [19], only a handful of studies have tested how accuracy of recommendations is affected by explanations [1, 2]. Some useful insight into this dilemma can be gained from research in the field of human interaction with automated decision support systems, where complex cost-benefit evaluations have been conducted with systematically varied levels of accuracy [11, 16, 22]. Additionally, research on explanation facilities for expert systems such as those designed to support medical diagnosis [5] have examined performance outcomes. These studies, however, have focused primarily on automating decisions in settings that impose a high cognitive load on users which is a different set of requirements than in many of the other systems we have described. In those systems, there is less stringent time requirement and there is less focus on decision accuracy Moreover, while these studies have examined the effect of accuracy on performance they have not, to our knowledge, addressed the combination of accuracy and explanation.

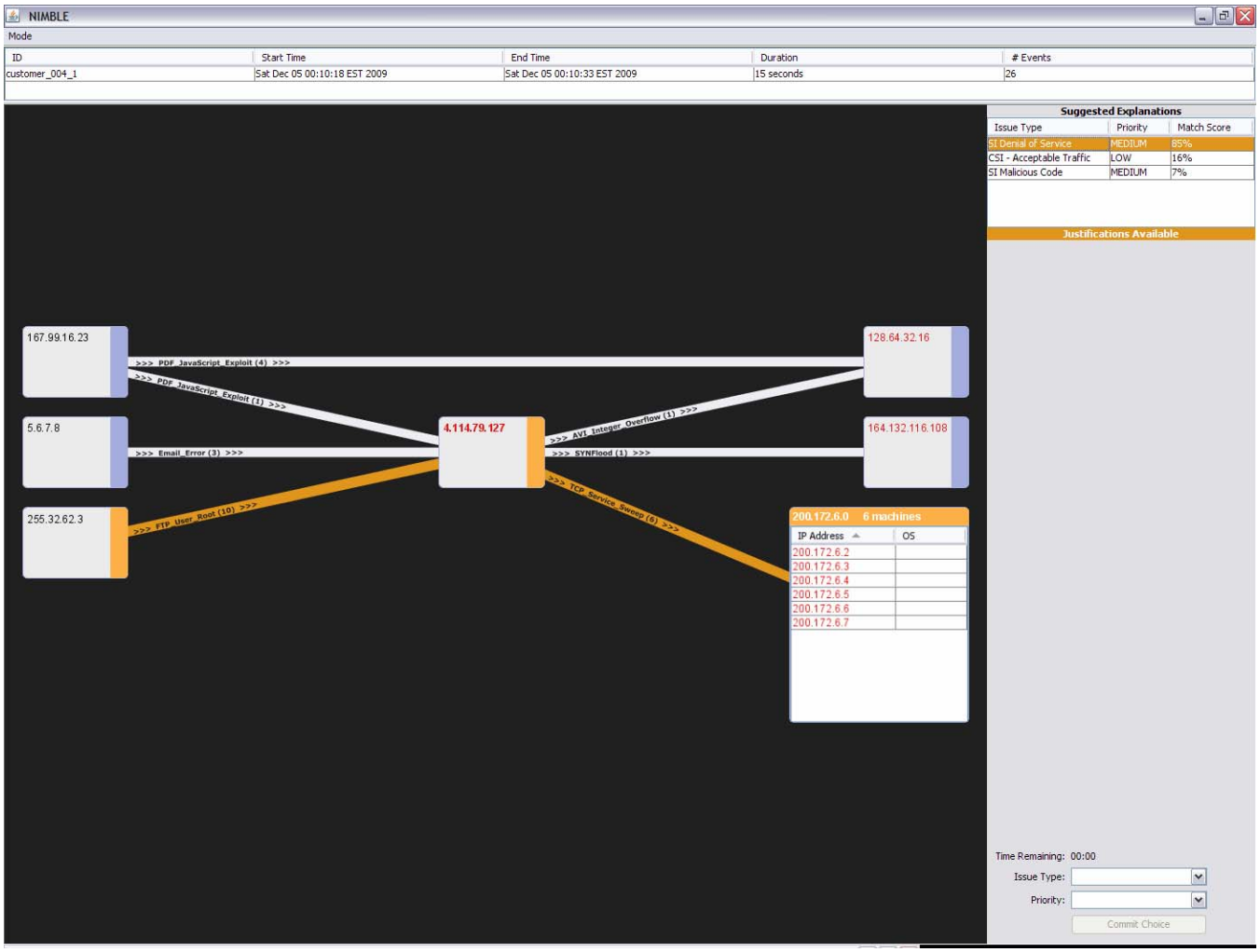

Figure 1: Screenshot showing alert with suggestions and justification with highlighting (data are simulated for confidentiality).

STUDY CONTEXT

The study was conducted with analysts who are engaged in real-time monitoring of cybersecurity incidents based on alerts that are generated from sets of events identified by intrusion detection systems (IDS), systematically categorizing and prioritizing threats.

Displaying Explanations

For the purposes of our study, we developed a prototype cybersecurity environment, NIMBLE (Network Intrusion Management Benefiting from Learned Expertise). The NIMBLE software reads correlated event data from an input file, creates a semantic model for each alert, matches each alert model against historical models in order to create recommendations, and displays alerts to the analyst. A sample screen with explanations is shown in Figure 1. Additional details and evaluation of the NIMBLE interface features, including a comparison between visual and tabular displays can be found in [15] from which much of the following description is derived.

Figure 1 shows a graph in which each node or "card" represents one or more machines and the edges connecting nodes represent sets of IDS signatures involving the connected machines. Cards are connected with labeled edges that indicate IDS event signatures and counts. The width of the edge is an indication of the total number of events it represents. The width is scaled by the natural logarithm of the event count; however a minimum size is enforced to ensure legibility of the label, which is drawn inside the edge.

NIMBLE can display incident classification suggestions with explanations, suggestions without explanations, or no suggestions at all. In Figure 1, the panel to the right of the alert display lists three suggestions under the headings of Issue Type, Priority and Match Score. The Match Score was only shown when the suggestions were accompanied by explanations. In this figure, the best suggestion is selected, and the display has updated to show the region of the concern graph that corresponds to the rule model underlying the suggestion.

Explanations take the form of selective highlighting in the main display. When the user selects one of the suggestions, the display highlights the portions of the currently viewed alert that match that suggestion. The degree of match is indicated using three shades of orange. The colors mean slightly different things for cards and edges, but in both cases the darker the orange the closer the match.

For the machine cards:

Dark Orange: Exactly the same machines.

Medium Orange: Not the same machines, but the same clustering, i.e. a single machine mapped to a single machine, or many machines mapped to many machines.

Light Orange: Single machines mapping to multiple machines or vice-versa.

For the edges:

Dark Orange: Exactly the same set of signatures (but counts may vary).

Medium Orange: Some overlap in the set of signatures.

Light Orange: No overlap in the actual signatures, but the system interprets something about the event activity as corresponding to the template model. (E.g. could have been the same TCP port in both cases.)

Hovering over an orange card or edge would show a tooltip detailing the differences between the currently viewed alert and the historical alert that was the basis for the recommendation.

Computing Suggestions

To make incident classification recommendations, NIMBLE calculates the similarity between the model for a given alert and historical alert models. The scoring algorithm is based on a general purpose semantic matching algorithm, which attempts to find the least-cost correspondence between two semantic models. This is a classic inexact graph matching problem. While there are many sophisticated approaches to doing this kind of matching (see for example [12]), for the NIMBLE prototype we used a simple best-first search of the space of possible correspondences between the claims. Our matching procedure is asymmetric. We wish to treat one model as a template graph, for which we seek correspondences in the other model's matching graph. Thus a smaller template model may find a good match embedded in the context of a larger matching model, and our system will have detected a target attack embedded within the context of a larger alert. As our cybersecurity ontology does not use relationship hierarchies, correspondences only need to be considered between claims involving identical properties. The cost function for matching corresponding claims depends on the sum of the ontological distance between unequal corresponding source and destination entities, which itself is determined by the percentage of classes in the ancestry of the template entity that are not found in the ancestry of the matching entity. The search finds the set of correspondences which result in the lowest cost, thus achieving the highest degree of match. The reported match score ranges from 0 to 1.0, representing the degree of match found between the two semantic graphs. Although users were explicitly told that the match score was not a confidence measure, it was often interpreted as if it was.

One of the benefits of this approach is that it is possible for the system to provide a justification for the suggestions that it makes. As part of the similarity calculation, we identify how the features of the current alert model correspond to the features of a similar historical model, and we can visualize this alignment to analysts curious about the reasoning behind the suggestion.

Although we used this simple case-based reasoning approach to generate incident classification suggestions for use in our study, our ontology-based models can also be aggregated and generalized to form a more succinct set of abstract rules. Justifications remain possible with generalized rules.

EXPERIMENTAL DESIGN

Participants

Nineteen analysts participated in the study. All had a minimum of three years experience in the job and most had worked as an analyst for over five years. Data from one of the analysts was removed from our dataset due to the analyst's lack of experience with the particular event signatures that were key to accomplishing the task.

Procedure

Each analyst was tested individually in a two-hour session. Sessions began with an introduction to the study and a detailed training on the NIMBLE test console, lasting about 30 minutes. During the training, participants had an opportunity to ask questions as they viewed an example of each of the display and suggestion conditions and completed two hands-on examples. They were also explicitly informed that the system was a prototype and could be delivering incorrect information.

Following the training, analysts completed 24 timed analysis trials, with a break at the midway point. They were instructed to complete each trial within two minutes and to give their best guess if they ran out of time. A chime sounded 15 sec before the end and again at the two minute mark. The alert, however, remained displayed until the analyst completed the task, even if it took longer then two minutes. The purpose of imposing a two minute limit was to mimic the limited time constraints under which analysts often operate. Pre-testing with analysts, who did not participate in the main study, confirmed that two minutes was realistic for completing the tasks.

The task had three parts. First the analyst determined the category of alert and its priority by selecting the alert

category from a list of 11 items and the priority from a list of 2 items (Low, Medium). We did not provide "High" as a priority choice, as we had no examples of high-priority alerts in our dataset, so no suggested explanation could be high-priority. The analyst indicated their completion of this task by clicking on a button. A mockup of the task is shown in Figure 2.

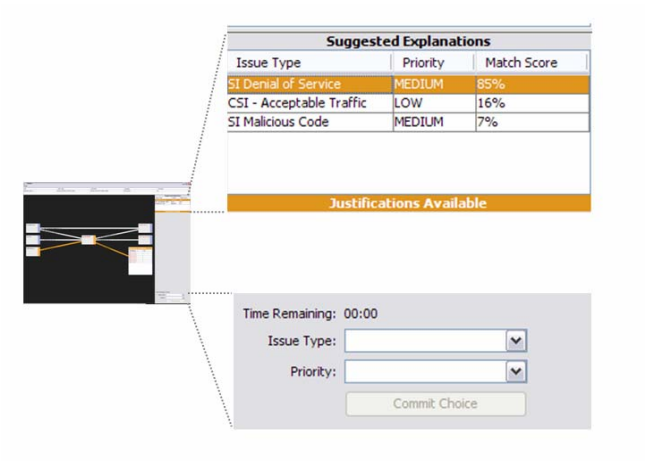

Figure 2: Mockup of a user task with justification.

Analysts were asked to talk aloud during the trials about what they noticed in the displays and how they were solving the task. They were given an opportunity between trials to make additional comments and observations on the tasks and the user interface. We recorded audio from the entire session, with their permission. Individual sessions concluded with a survey in which analysts rated the value of suggestions and justifications, and provided general feedback and reflections on their experience. After all the analysts had completed their individual sessions, they attended a two-hour focus group to discuss their impressions of the study. The analysts were given no feedback either during the trials or after, about the accuracy of their responses.

Research Variables

We tested the research goals with a balanced parametric design in which we independently varied two variables:

- **Recommendation**. The alert was either presented on its own (baseline condition), with suggestions or with justifications. In the suggestion and justification conditions, we always provided 3 choices. Each choice included the Issue Type and the Priority as illustrated in Figures 1 and 2. In the baseline condition, the system generated suggestions but these were not presented to the user.
- **Correctness.** There was either a correct choice available or all the choices were incorrect. When all the choices were incorrect, none of the issue types alone were correct.

There were 4 instances of each condition for a total of 24 trials. There were 24 different alerts were balanced across the different conditions such that each user saw all 24 just once. The order in which the 24 trials were presented was randomized for each participant.

It should be noted that the current study was part of a larger study [15] that included additional variables that are not reported here since they were not relevant to the issues we are examining.

Dependent Measures

As participants completed each trial, the NIMBLE software logged their response. These log data were converted into our primary dependent measure of accuracy[1]. The data were analyzed using ANOVA repeated measures design.

After the timed trials were completed, each subject was asked to rate the helpfulness of suggestions ("*In general how helpful were the suggestions in this task*") on a scale from 1 (very unhelpful) to 5 (very helpful). We collected additional demographic information, from management, for each analyst: Tenure (years with the organization), Level of Expertise (ranging from 1 for a junior person to 5 for the most senior).

Quantitative measures from the trials and surveys were augmented by qualitative data from audio recordings of each session and from a group debrief session which took place after all the trials had been completed.

RESULTS

Figure 3 shows accuracy across all conditions.

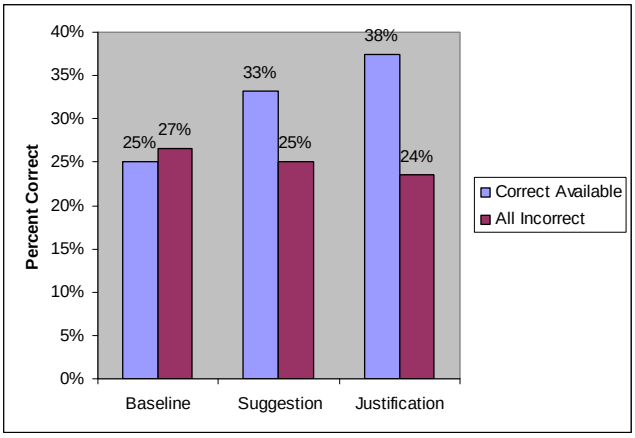

Figure 3: Mean response accuracy for recommendations and correctness.

[1] The term "accuracy" is used as shorthand to refer to agreement between the combination of issue type and priority by the analyst in the study and the designation given to the same alert in the historical record.

Users performed better when there was a correct choice compared to no correct choice ($F_{1, 17} = 4.21$, $p = 0.06$). There was no overall statistical effect of recommendation (comparing baseline, suggestion and justification irrespective of whether correct choice was available) ($F_{2, 34} = 1.03$, ns), nor was there an interaction between suggestion and suggestion accuracy ($F_{2, 34} = 1.31$, ns).

Considering only those cases where there was a correct choice available, there was a marginally significant improvement for suggestions and justifications over the baseline condition ($F_{2, 34} = 3.05$, $p = 0.06$). In pairwise comparisons when the correct choice was available, there was a significant difference between the baseline and justifications ($p < 0.05$), but not between suggestions and justifications. Nor was there any significant difference between the baseline and suggestions. Under the conditions when no correct choice was available, there was no difference between the baseline condition and either suggestions or justifications ($F_{2, 34} = 0.09$, ns).

These results point to the benefit of justifications over a baseline, when a correct response is available. When no correct response is available, neither suggestions nor justifications lead to any improvement over a baseline condition. Neither is there a decline in performance relative to the baseline, suggesting that the absence of a correct choice was not harmful. The lack of a substantial improvement for justifications over suggestions implies that our system's justifications are not providing sufficient additional information to promote either trust or transparency.

Individual Differences

In the course of conducting the study we observed that some users seemed to respond more strongly to the recommendations than others. Users strongly value self-sufficiency, independent analysis, and individual judgment. Most participants expressed disinclination to follow system-generated suggestions without their own confirmation:

"I was using the justifications just to nudge my own analysis a little bit, but I was pretty much doing my own analysis."

"I used my discretion and outweighed the automatic choices."

However, other users seemed to use the suggestions to guide their thinking rather than as confirmation.

"having suggestions was helpful just to clear out some of the other stuff"

To examine possible individual differences in the processing of justifications, we used the ratings from the post-trial survey, which measured perception of the helpfulness of suggestions, to segment our population. Of the 18 analysts in our study, six gave a rating of 4 (on a scale of 1-5 where 5 is high), seven gave a rating of 3 and five gave a rating of 2. We designated these analysts as high, medium and low respectively, to indicate differences in their perception of suggestions. There was no difference between these three groups in years of experience or level of expertise.

To test for differences between the groups, we re-ran the analyses for recommendation and correctness, to include Group as a between subjects variable. The results are shown in Figures 4 and 5. The low and medium groups did not differ significantly in their responses so their data are combined in the Figures; the analyses included all 3 groups.

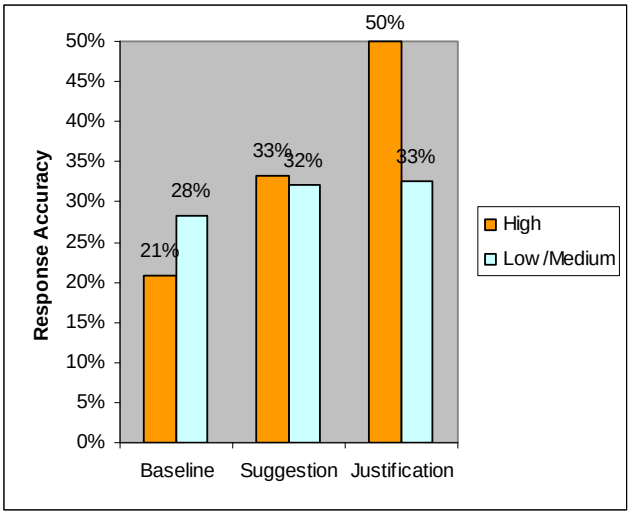

Figure 4: Mean response accuracy across user groups when there is a correct choice available.

Analysis revealed a significant 3-way interaction between Recommendation, Correctness and Group ($F_{4, 30} = 3.80$, $p < 0.05$). The main effect of Correctness reported in the previous section remained significant ($F_{1, 15} = 4.8$, $p < 0.05$) but the interaction between Recommendation and Correctness did not ($F_{2, 30} = 1.3$, $p > 0.10$). There was no significant difference between the groups overall ($F_{2, 15} = 0.61$, ns) nor did the groups differ in the baseline condition.

Figures 4 and 5 indicate that the High group is benefiting much more than the other groups, from justifications, but only when a correct choice is available. When there is a correct choice, they are more accurate with suggestions and justifications compared to the baseline ($F_{2, 10} = 10.23$, $p < 0.01$). In pairwise comparisons, suggestions were marginally better than the baseline ($p < 0.10$); justifications were significantly better than the baseline ($p < 0.001$). Additionally, they performed better with justifications than with suggestions alone ($p < 0.05$). This is an important result because it suggests an additional benefit for justifications over suggestions alone, for this group.

Conversely, the High group performed worse when no correct choice was available ($F_{1, 5} = 14.76$, $p < 0.05$). There was no overall difference in performance between baseline, suggestion and justification when no correct choices were

available. However, in pairwise comparisons, there was a marginally significant difference between suggestions and justifications ($p < 0.10$). Again, this is an important result because it indicates that the justifications led them to select rather than reject the bad choices.

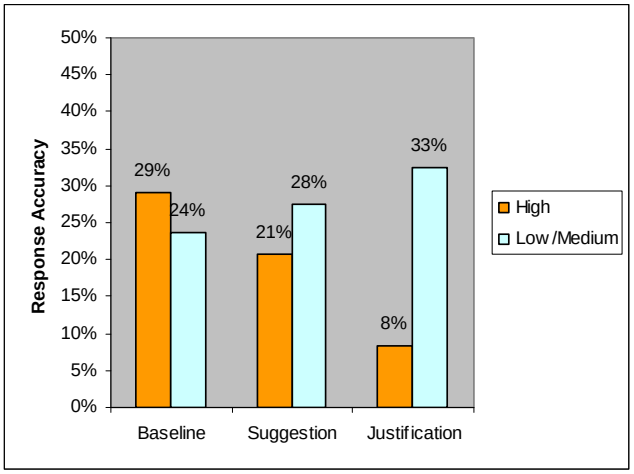

Figure 5: Mean response accuracy across user groups when there are NO correct choices.

Are users following the recommendations?

Before we draw conclusions about whether users are influenced by recommendations, we need to look more closely at their responses. The additional examination is necessary in order to draw conclusions about whether the presence of suggestions or justifications is influencing users' choices over and above the accuracy of their choice.

In this study, users were always presented with 3 choices in the Suggestion and Justification conditions. We are interested in whether users are influenced by the presence of a choice. In half the cases, one of these choices was correct and two were incorrect. In those cases there were 2 out of 3 opportunities for users to provide a wrong but consistent response. In the other half of the cases, none of the three choices were correct. In those cases, the only way for a user to respond accurately was to ignore the choices and respond with an answer that was not provided. Similarly, if the user gave a wrong response it could have been consistent with one of the choices presented or not. Note that 3 choices were computed for all alerts used in this study but they were not visible in the baseline condition.

We first examine whether there is any evidence that users are following the suggestions, irrespective of whether their response is correct. That is, we are only looking at whether the response is consistent with one of the three choices. Figure 6 shows the overall frequency of giving a response that is consistent with one of the choices provided. There was a significant difference between baseline, suggestions and justifications ($F_{2, 35} = 8.4$, $p < 0.01$) and between one correct or no correct choices ($F_{1, 7} = 43.2$, $p < 0.001$). In pairwise comparisons there was a significant difference between the baseline and suggestions, and, between the baseline and justifications ($p < 0.05$). There was no difference between suggestions and justifications alone. These results suggest that users are indeed being influenced by the choices since they are more likely to give a response consistent with one of the recommendations when it is visible than when it is not. However, they may also be applying judgment to their decision because the data show that they are less likely to select one of the choices when none of them are correct. There was no interaction between Recommendation and Correctness.

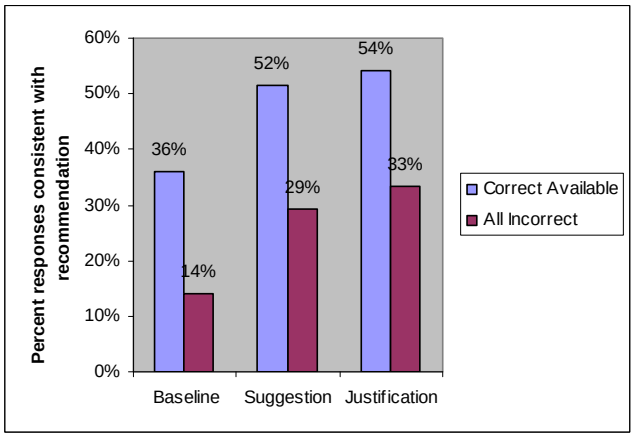

Figure 6: Mean consistency of responses for Recommendation and Correctness.

In our previous analysis, we reported that the High group was less accurate with a justification than they were with suggestions alone or in the baseline condition, if there was no correct response provided. But they could have given a wrong answer that was not part of the recommendation. In order to examine whether the justifications were influencing users to give one of the responses on the list, we analyzed the "follow" responses segmented by user group as well as by our two main variables. The data are shown in Figures 7 and 8.

Analysis of the data revealed a significant interaction between Recommendation and Group ($F_{4, 30} = 3.2$, $p < 0.05$) and a significant 3-way interaction ($F_{4, 30} = 2.7$, $p < 0.05$). When no correct choice was present, the High group continued to respond with one of the choices, especially in the presence of justifications, but the other groups did not.

This result provides additional evidence of the influence of justifications for the High group. For other users, the justification had no additional benefit over suggestions. These results extend the findings, reported in the previous section, that people in the High group benefit from justifications when there is a correct choice available but suffer when there is no correct choice.

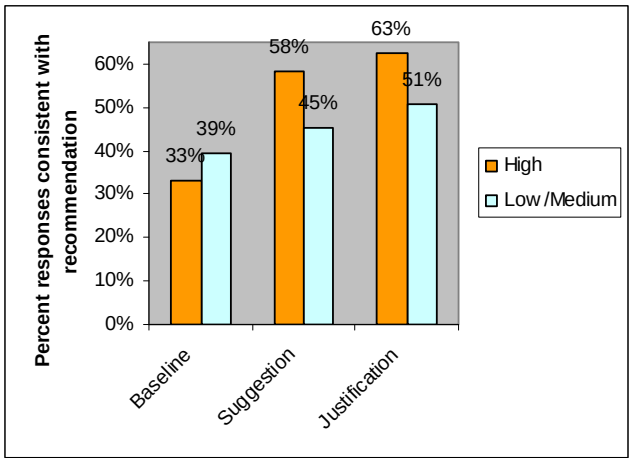

Figure 7: Percent responses consistent with any recommendation when one correct choice was provided.

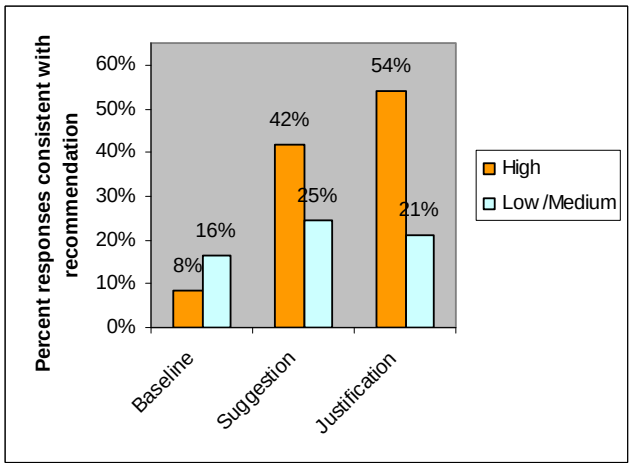

Figure 8: Percent responses consistent with any recommendation when NO correct choice was provided.

DISCUSSION

Intelligent systems can be an important aid to user's decision-making. While previous research has established that users can benefit when these systems enhance their recommendations with system reasoning [4, 9, 13, 14], there has been much less focus on whether there are detrimental effects when the system offers incorrect advice. Studies on automation in decision-support have explored the consequences of providing imperfect information [11, 22]. However, these studies have mostly focused on lower level processing and workload and have not systematically examined incorrect information for explanations.

The study reported in this paper sought to examine the benefit of a form of explanation, justifications, and possible negative consequences of justifications in the absence of correct recommendation. There were two main manipulations in the study: the kind of recommendation and the accuracy of the recommendation. These two manipulations allowed us to evaluate both the potential benefit and risk associated with justifications. By comparing suggestions with a baseline, we can evaluate whether providing suggestions at all is helpful. By comparing justifications with suggestions, we can see whether the additional rationale aids users in making decisions. By comparing cases where there is one correct recommendation with cases where there is no correct recommendation, we can examine whether justifications help users make better decisions or lead them to make a wrong decision.

Additionally, we explored whether users might systematically vary in their responses to justifications and to the absence of correct information. The users in our study differed in their perception of the value of suggestions which we used as the basis for segmentation. Our rationale in using this method of segmentation was that users who perceive value in the suggestions are more likely to make use of them.

Considering the main effects of recommendation and correctness, users benefited from the recommendations over the baseline condition; there was no additional benefit for justifications over suggestions. On the other hand, there was also no reduction in performance in the absence of correct choices. There were, however, differences between the three groups of users. Those who valued the suggestions the most were also most affected by them. They benefited from the additional rationale provided by the justifications to perform better than with suggestions alone. But they also performed worse with justifications in the absence of any correct choices. In other words, we obtained a pattern of results, for one segment of users, in which the benefit of justifications carries with it a potential liability of following wrong advice in the absence of correct choices.

Previous research has suggested that explanations are beneficial because they increase the transparency of the system and increase trust in the system. The transparency argument suggests that users should be better equipped to decide which, if any, recommendation to follow because they have more insight into the system's reasoning process. Or, in cases where there is only a single recommendation, users should be able to evaluate whether to follow it at all. Transparency predicts that all users should perform better with explanations than either suggestions or baseline, even if none of the recommendations are correct. The trust argument suggests that providing explanations gives users greater confidence in the system which translates into a greater propensity to accept the recommendations. This argument predicts that users will be more likely to select one of the recommendations even when none are correct.

The results of the current study are consistent with a hybrid model, at least in the case of users who perceived the most value from suggestions. These users followed the recommendations even when none were correct, implying

that they trusted the system. Had the benefit of justification been to increase transparency then they should have been able to reject the recommendations when none were correct. However, their performance was markedly poorer implying that they trusted the system to provide correct recommendations and selected one of them leading to a wrong response.

There are other interpretations for the results we observed. One is that users formed own decisions from the alert and used the recommendations as confirmation if it agreed with their decision, or rejected the recommendation if it did not agree with their decision. Based on comments during the trials, we believe this is how some participants operated. Another interpretation is that the presence of recommendations "primed" or biased the users towards selecting one of the choices. We see evidence of this argument in the data that shows that users tended to follow the suggestions offered, although more often in the case where one of the recommendations was correct, thus demonstrating judgment in addition to merely "following" behavior.

Limitations

As with any lab based study, the requirements imposed by the need for control in the study design may limit the generality of some of the findings.

Justifications. The justifications were derived from a restricted set of raw data which limited the breadth of information we could provide. If the reasoning behind the justifications seems solid and reasonable, we expect the users to have more trust in the suggestion, whereas if the explanation seems suspect, they should distrust it more. The justifications in our study may not have been sufficiently compelling to give all users the level of trust over suggestions alone. Moreover, the effectiveness of justifications may depend on how understandable they are to the user. Some of the justifications in the present study could highlight spurious correspondences, for example. So the user, looking at the justification wouldn't necessarily have much insight into why the system made its recommendation. If the justifications were strongly highlighting a particular path through the network, and saying "*See these signatures, that's why I made this particular recommendation*" the user might have been able to agree or disagree with it better.

Individual differences. We used a rather blunt instrument for segmenting users which needs further research to establish a more robust measure. The instrument, did, however, highlight the importance of considering individual differences for consideration in design of intelligent systems which has been largely neglected in the literature. We were not able in this study to pinpoint the source of the difference, but we did hear from the analysts about acknowledged stylistic differences in approaches to decision-making. These differences did not appear to relate to performance directly but is an area that warrants additional research.

The present study differs from many in the literature in several ways. Firstly, we are examining recommendations in mission critical settings where there is a single correct answer rather than a range of acceptable alternatives. This setting imposes a different set of requirements, expectations and especially consequences than settings in which the system is providing recommendations for the purchase or use of consumer products. Secondly, we are offering users an explicit and fixed number of choices as compared with systems that might offer only a single recommendation, from which users are to only make one selection. The added complexity of multiple recommendations is consistent with a model in which the user first decides whether to use any of the recommendations, and then decides which recommendation to use. Thirdly, our system provided no feedback to users during the trials that they might have used to adjust their behavior.

Implications

The results of this study have implications for the design of intelligent systems that strive to provide recommendations to aid users in their decisions without penalizing them. Most systems unintentionally provide incorrect advice some or most of the time. Since the system cannot detect when its advice is wrong, users' apply their own reasoning to decide when to accept the system's advice. The present study suggests that some users are more inclined to accept the system's recommendation, even when it is wrong. If designers of these intelligent systems want to limit the effects of poor recommendations for these and other users, one approach is to only present recommendations that are at or above a threshold of confidence which has been advocated by other researchers [22]. Although an intelligent system can't detect when a recommendation is wrong, most systems can generate confidence parameters for their suggestions.

Errors induced by compelling explanations can also spell trouble for case-based reasoning systems that update their rules and heuristics with data from current users. If these users are themselves making errors which the system can't detect, the core data used by the system will degrade over time rendering the system less reliable.

Summary

Research on intelligent decision support systems has advocated that systems provide explanations of their reasoning. This study found that providing these explanations can be a double-edged sword. They are helpful when accompanied by correct recommendation but can be detrimental in the absence of any correct recommendation, for some segments of the user population. These results point to the importance of considering weaknesses as well as strengths, of explanations.

REFERENCES

1. Bilgic, M. and Mooney, R.J. Explaining Recommendations: Satisfaction vs. Promotion. *Proc. Beyond Personalization Workshop, IUI* (2005).

2. Celma, Ò. and Herrera, P. A new approach to evaluating novel recommendations. *Proc. RecSys '08* (2008) 179-186.

3. Cunningham, P., Doyle, D. and Loughrey, J. An evaluation of the usefulness of case-based explanation. *Proc. Case-based reasoning: Research and Development '03* (2003) 122-130.

4. Dhaliwal, J.S. and Benbasat, I. The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. *Information Systems Research* 7 (1996).

5. Doyle, D., Cunningham, P. and Walsh, P. An evaluation of the usefulness of explanation in a case-based reasoning system for decision support in Bronchiolitis treatment. *Computational Intelligence* 22 (2006) 269–281.

6. Fleischmann, K.R. and Wallace, W.A. A covenant with transparency: opening the black box of models. *Communication ACM* 48 (2005) 93-97.

7. Fleischmann, K.R. and Wallace, W.A. Ensuring transparency in computational modeling. *Communication ACM* 52 (2009) 131-134.

8. Glass, A., McGuinness, D.L. and Wolverton, M. Toward establishing trust in adaptive agents. *Proc. IUI '08* (2008) 227–236.

9. Gregor, S. and Benbasat, I. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly* 23 (1999) 497–530.

10. Herlocker, J.L., Konstan, J.A. and Riedl, J. Explaining collaborative filtering recommendations. *Proc. CSCW '00* (2000).

11. Lee, J.D. and See, K.A.T. Trust in automation: Designing for appropriate reliance. *Human factors* 46 (2004) 50.

12. Luo, B. and Hancock, E.R. Structural Graph Matching Using the EM Algorithm and Singular Value Decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (2001) 1120-1136.

13. Mao, J.Y. and Benbasat, I. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems* 17 (2000) 153–179.

14. Pu, P. and Chen, L. Trust building with explanation interfaces. *Proc. IUI '06* (2006).

15. Rasmussen, J., Ehrlich, K., Ross, S., Kirk, S., Gruen, D. and Patterson, J. Nimble cybersecurity incident management through visualization and defensible recommendations. *Proc. VizSec '10* (2010).

16. Rovira, E., McGarry, K. and Parasuraman, R. Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors* 49 (2007).

17. Sinha, R. and Swearingen, K. The role of transparency in recommender systems. *Proc. CHI '02 extended abstracts* (2002).

18. Tintarev, N. and Masthoff, J. Effective explanations of recommendations: User-centered design. *Proc. RecSys '07* (2007).

19. Tintarev, N. and Masthoff, J. A survey of explanations in recommender systems. *Proc. ICDE'07 Workshop on Recommender Systems and Intelligent User Interfaces* (2007).

20. Vig, J., Sen, S. and Riedl, J. Tagsplanations: explaining recommendations using tags. *Proc. IUI '09* (2009) 47–56.

21. Wang, W. and Benbasat, I. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23 (2007) 217–246.

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