An Analysis of Displays for Probabilistic Robotic Mission Verification Results

Matthew O'Brien[1], Ronald Arkin[1],

[1] School of Interactive Computing, Georgia Tech, Atlanta, GA 30332
{mjobrien,arkin}@gatech.edu

Abstract. An approach for the verification of autonomous behavior-based robotic missions has been developed in a collaborative effort between Fordham University and Georgia Tech. This paper addresses the step after verification, how to present this information to users. The verification of robotic missions is inherently probabilistic, opening the possibility of misinterpretation by operators. A human study was performed to test three different displays (numeric, graphic, and symbolic) for summarizing the verification results. The displays varied by format and specificity. Participants made decisions about high-risk robotic missions using a prototype interface. Consistent with previous work, the type of display had no effect. The displays did not reduce the time participants took compared to a control group with no summary, but did improve the accuracy of their decisions. Participants showed a strong preference for more specific data, heavily using the full verification results. Based on these results, a different display paradigm is suggested.

Keywords: display of uncertainty, decision support system, formal verification, behavior-based robotics

1 Introduction

Robotics has the potential to be a key technology for combating weapons of mass destruction [1]. This domain presents new challenges for autonomous robotic systems. In these types of missions, failure is not an option. Human operators must be confident in the success of a robotic system before the technologies can be applied. To address these problems our research, conducted for the Defense Threat Reduction Agency (DTRA), has successfully developed the methods and software to perform robotic mission verification [2].

While robotic mission verification is similar to traditional software verification, there are several additional complications. The real world is continuous, and both robotic sensors and actuators are noisy. The robotic controller is only one piece, and the result of a mission is also determined by the physical robot and its interaction with the environment, and modeling of both will always be imperfect. This means any verification is fundamentally probabilistic.

This presents a new challenge. People do not use all the available data or systematic methods when assessing probabilistic data. Instead, heuristics are applied to simpli-

fy the analysis, which can lead to systematic errors and bias [3]. The methods of displaying the information must ensure an operator can easily and accurately interpret the data. This paper explores methods to achieve this goal.

2 Related Work

Research on the presentation of probabilistic data and uncertainty has shown that participant's decisions in various tasks are not significantly affected by the format the data is presented in (graphical, numerical, or verbal) [4]. Though numeric statements offer more precision and consistency than linguistic phrases, it's hypothesized that people treat all probabilities in a vague manner, utilizing membership functions [5]. These results were extended in [6] where both display format and specificity level were varied. Display formats included linguistic, numeric, and multiple graphical icons. Specificity level was the size of the range of probabilities represented by a single icon or expression. Results agreed with previous research, showing that display format had no significant effects. However, specificity did have significant effects on performance in a simulated stock purchasing task.

This work expands on these results in two ways related to the application of robotic mission verification. First, participants in this study have access to more information than a single measure of probability. Success in a robotic mission is tied to multiple criteria, such as time to completion or allowable distance from a goal location, whose values may have some variability. The full verification results have probabilities of achieving each criterion independently over a range of values. This information is important to operators, and will affect their decisions, so it must be included. Secondly, the context of the tasks is significantly different. Participants were asked to make decisions on high-risk missions, where lives are (hypothetically) at risk. These types of risks/costs are difficult to quantify and participants may resort to different methods of reaching a decision.

3 VIPARS – The Verification Tool

VIPARS, or *Verification in Process Algebra for Robot Schemas*, is a robot mission verification tool [7] designed for use with *MissionLab*, a graphical programming environment for behavior-based robots [8]. Informally, VIPARS determines how likely a robot mission is to succeed. In formal terms it takes as input a behavior-based robotic controller (software), models of the robot hardware and environment, and performance criteria. With this information, VIPARS can calculate and return a probability of success. All of these components are descriptions of the physical system except for performance criteria, which define what a successful mission is. The two most fundamental criteria, and those used for this study, are time (how long a robot may take to achieve its goal) and space (how far from a goal a robot may be).

VIPARS achieves verification by defining the state of the system as a set of random variables. Flow functions, created from the robot's behaviors and the environmental models, describe how these random variables map from one time step to the next. This allows VIPARS to avoid the state-space explosion, caused by the continu-

ous dynamics and noisy sensing/actuation of the real world, that plague traditional verification techniques such as model checking [9]. This paper will not go deeper into the description of VIPARS. For a thorough description of the verification process please see [2].

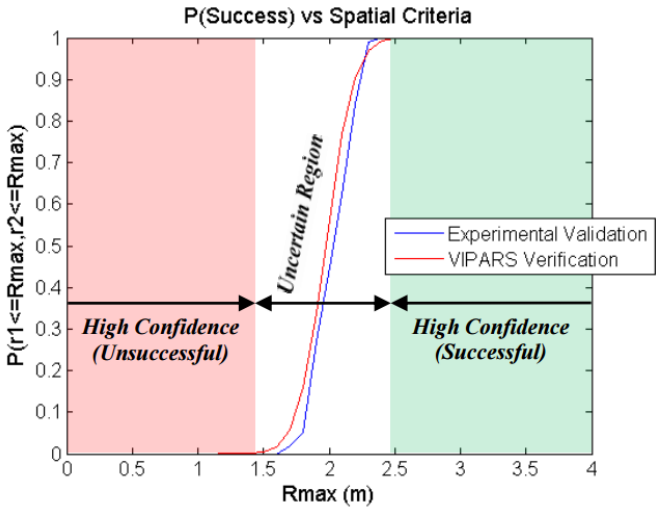

Fig. 1. Example verification results compared to empirical validation from real robots. Rmax is the spatial criterion, or the max distance from the goal location that still counts as a successful mission. The y-axis is the probability that both robots, r1 and r2, meet their spatial criterion.

Though VIPARS can produce a single probability for specific performance criteria, a more complete understanding of a mission can be gathered from observing how the probability changes over a range of criteria. See Figure 1 above for the results of a multi-robot mission verification from [7]. The red curve is the VIPARS verification result, while the blue curve is experimental data gathered from real executions of the robot mission for validation purposes. Results can be broken down into three regions. In the Unsuccessful region, the performance criterion is so strict that success is impossible, i.e., the precision or speed demands exceed the capabilities of the system. In the Successful region, the criteria are easy enough to guarantee success (ignoring unmodeled possibilities). Both of these regions are high-confidence, where it is certain the actual mission probability will match the verification results. In between lies the uncertain region, where uncertainty is introduced in two ways. First, the results are between 0% and 100%, so mission success is uncertain even with a perfect verification. Second, in this region small errors or simplifications in modeling can create moderate differences between the predicted and actual probability of success. Thus, the results of the verification itself are low-confidence. With a basic understanding of VIPARS and the data it produces, the experiment described in this paper and the displays used can be discussed.

4 Experimental Design

4.1 Task

Participants were asked to make decisions on whether to execute high risk autonomous robotic missions based on situational information and the verification results. Participants were presented scenarios appropriate for a mobile robot mission. They were given access to the VIPARS graphical interface via a laptop under the assumption the robot (hardware) and controller (software) had been decided and are fixed. The participants reviewed information on a scenario (robot's task, risk factors, time or spatial constraints) and then executed the VIPARS verification. Using the information VIPARS provided, the participant made a decision on whether to execute the robot mission or defer to a human team, and rated their confidence in both the mission and their decision. Scenarios included some limited information about the performance and risk for human teams.

Each participant was presented five total scenarios. The scenarios were divided into two categories. Certain scenarios were made to have a clear correct decision, with probabilities of success either being 0% or 100%, and high confidence in the verification. There were three certain scenarios, two successful and one unsuccessful. The uncertain scenarios had probabilities of success at 30% and 70%, as well as low confidence in the verification.

4.2 Independent Variable – The Displays

Participants were divided into four conditions. Every condition had the low-level display available, which showed the full verification results. For three conditions (A-C), subjects were presented a variant of a high-level display and could switch to the low-level display at will, while in the control condition (D) subjects could only view the low-level display.

The Low-Level Display

The low-level display provides the full probabilistic information given by VIPARS. This display is based on the validation graphs discussed in section 3. The graph is a cumulative distribution function (CDF) for the probability of achieving a performance criterion over a range of values. Figure 2 shows an example of the graph. It is augmented in several ways to aid a user. Areas with 0% or 100% success probabilities make up the "high-confidence success" or "high-confidence failure" regions. These are colored for rapid identification. The threshold a user has selected for the specific criteria is marked with a dashed line. In addition, the scales of the presented graph are selected relative to this threshold; from zero to twice the value. This presentation is limited to one criterion at a time, so a user must manually switch which criteria they are viewing.

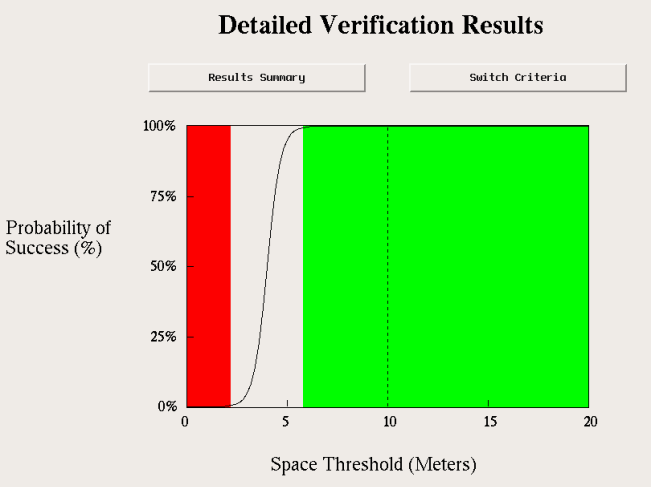

Fig. 2. Example low-level display for a spatial criterion set at 10 meters.

The High-level Displays

The high-level display summarizes the verification results in two ways. First, only the probability at the selected criteria value is used. This means information on the effect of changes in mission criteria is lost. Second, the results of all criteria are combined to give a total probability of success. This removes mental calculations from the user, but hides potential causes of failure. Three display types were chosen that vary with respect to type and specificity.

At the high end of the specificity scale is a simple numeric display of the final mission probability, which can be considered the most basic approach. A less precise means of displaying a percentage is graphically, using a bar. A bar was selected because reading position along a common scale has been shown to be the most accurate task for extracting quantitative information from a graphical representation [10] and it is commonly used in decision support systems (e.g. [11][12]). At the lowest level of specificity, a symbolic system only presenting three options (success, failure, and uncertain) could be used for the high-level display. This scheme takes advantage of the current predictions of VIPARS which typically have low confidence in probabilities between 0% and 100%. In this symbolic system a green thumbs up represents success, a red thumbs down represents failure, and a question mark represents uncertain results. Figure 3 presents the options along a scale of specificity.

4.3 Dependent Variables

For each scenario five dependent variables were recorded, shown in Table 1 below. The first three variables were automatically recorded by the software, while the last two are selected by the participant. Mission and decision confidence were presented as Likert scales with values ranging from 1 to 9. For participants in the control group,

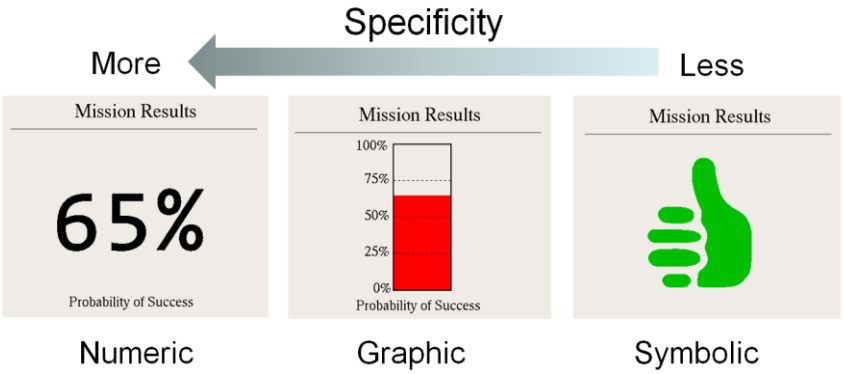

Fig. 3. The high-level displays. The numeric display was used for condition A, the graphic for B, and symbolic for C. Condition D was the control group.

time-to-decision is equal to *time-on-raw-data*, as they can only view the low-level display.

Table 1. The five dependent variables recorded in the study

User-decision	Binary choice on whether to execute the robotic mission
Time-to-decision	The time between VIPARS execution and final decision
Time-on-raw-data	Time spent viewing the low-level display
Mission-confidence	Confidence the robotic mission would be successful if ran
Decision-confidence	Confidence the user's decision (to execute or not) is correct

4.4 Hypotheses

Based on the related work discussed previously, three hypotheses were formed. This section covers the hypotheses and their predictions on the dependent variables.

> **Hypothesis 1: Displays summarizing VIPARS results can improve the comprehension accuracy and speed of users over the direct display of VIPARS output.**

Hypothesis 1 predicts that *time-to-decision* and *time-on-raw-data* will be reduced when using high level displays versus the control case, and that the accuracy of user-decision will increase for certain scenarios. If participants in the control cases achieve perfect accuracy (i.e. always select correct decision) for certain scenarios, then it will be assumed that perfect accuracy on the high-level displays validates this hypothesis.

> **Hypothesis 2: Various representations of the VIPARS output will provide similar understanding of the mission probability.**

Hypothesis 2 predicts that between the high level displays, user-decision will not vary significantly for uncertain scenarios.

> **Hypothesis 3: More precise representations of probability will bias operators towards interpreting higher certainty in the result.**

Hypothesis 3 predicts that *decision-confidence* will increase as the specificity of the high-level display increases.

Finally, additional analysis is performed to look for effects that do not have explicit hypotheses. For example, if one particular display has a higher *time-on-raw-data* on the average, it may indicate that users find the representation inadequate for decision making.

4.5 Execution Details

A total of 45 participants were tested. Participants were screened for color blindness with a shortened version of the Ishihara colorblind test, two failed and were excluded. In addition, two participants performed the tasks incorrectly[1], their data was also excluded. The results include 41 participants, 23 male and 18 female, with an average age of 24.3 (range from 18 to 54).

Each participant first went through a tutorial session that introduced the VIPARS system and allowed them to try an example scenario. Afterwards, they were given information on one scenario at a time by the proctor. The proctor was nearby and available for questions, but not able to view the computer or participant's choices. Sessions were video recorded, and the time taken for questions and answers during the test was removed from the measurements of t*ime-to-decision* and t*ime-on-raw-data*.

5 Results

5.1 Hypothesis 1

The first hypothesis made two predictions. First, that users would make faster decisions when provided with the high-level displays, lowering *time-to-decision* and *time-on-raw-data*. The second was that the accuracy of user's decisions on certain scenarios would be improved when using high-level summaries. First we examine the data on *time-to-decision* and *time-on-raw-data*.

Both *time-to-decision* and *time-on-raw-data* were analyzed using one-way ANOVA over the four conditions. *For time-to-decision,* or the total time a user took, there was no significant difference between display types when all scenarios were averaged together ($P = 0.688$). Scenarios were also tested independently, and showed no significant differences. Figure 4 shows the *time-to-decision* for each display type. In contrast, there was a statistically significant difference between *time-on-raw-data*

[1] Participants used prior situational information for new scenarios

(P = 0.012). Post hoc tests using Games-Homell showed only display A had a significant reduction (alpha = 0.05) compared to the control (means = 32.22, 56.84, SD = 24.47, 41.27).

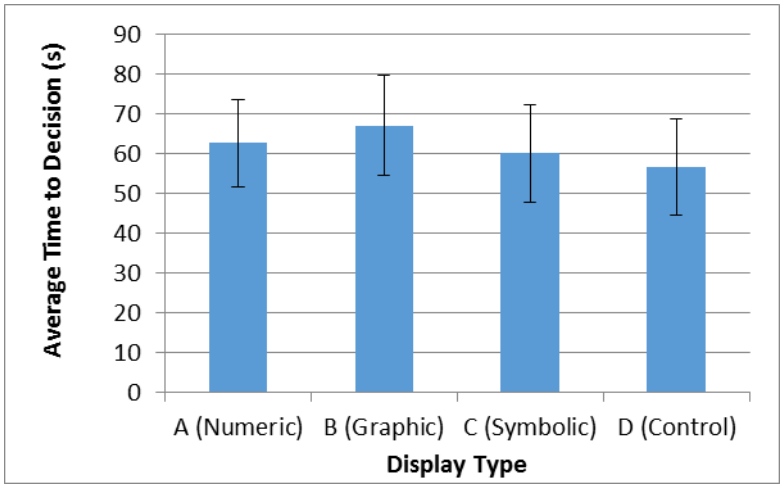

Fig. 4. Average *time-to-decision* per display type for all scenarios plotted with 95% confidence intervals.

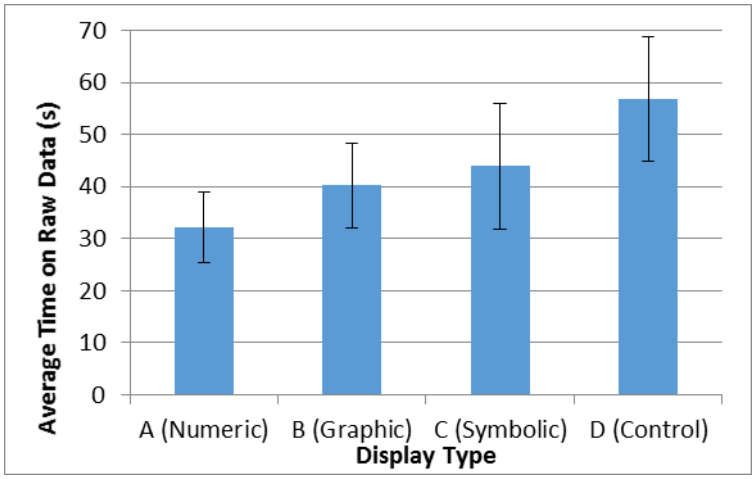

Fig. 5. Average *time-on-raw-data* per display type for all scenarios plotted with 95% confidence intervals.

For decision accuracy, uncertain scenarios were excluded as no correct decision could be assumed. This left the three certain scenarios. For each user and display, a correct decision was an “execute” for missions with 100% probability of success, and a “do not execute” for missions with a 0% probability of success. The table of decisions for each display is shown below. The reader can see that the control case D has a larger number of incorrect decisions. As the table is sparsely populated, Fisher’s exact test was used to test for statistical significance. A significant difference between display

types was found (P = 0.026). Thus hypothesis one is partially confirmed; the accuracy of users improved with the high-level displays, but their times to decisions were not reduced.

Table 2. All decisions for the certain-scenarios, sorted by condition.

		Display Type			
		A	**B**	**C**	**D**
Decision	**Correct**	28	33	32	22
	Incorrect	2	0	1	5

5.2 Hypothesis 2

The second hypothesis predicted that between the high level displays, the understanding of mission probability and thus *user-decision,* would not vary significantly for uncertain scenarios. As these scenarios had different probabilities of success (70% and 30%) they will be analyzed separately. The decisions for each scenario are broken down in Tables 3 and 4, and a Fisher's exact test reported the difference between display types was not statistically significant. (P = 0.906, 0.526 for scenarios one and two, respectively). Thus hypothesis two is confirmed.

Table 3. *User-decisions* for the first uncertain scenario, total probability of success = 70%

		Display Type			
		A	**B**	**C**	**D**
Decision	**Execute**	6	8	8	7
	Don't	4	3	3	2

Table 4. *User-decisions* for the second uncertain scenario, total probability of success = 30%

		Display Type			
		A	**B**	**C**	**D**
Decision	**Execute**	5	5	6	7
	Don't	5	6	5	2

5.3 Hypothesis 3

The final hypothesis predicted that more precise representations of probability will bias operators towards interpreting higher certainty in the results, thus *decision-confidence* will increase as the precision of the high-level display increases. This hypothesis needs to be tested per scenario, as different risks and probabilities with each scenario should affect the confidence of the user. Performing an ANOVA for *decision-confidence* versus display type showed no significant differences between displays. Table 5 and Figure 5 below display the P values and average values for each scenario. Thus hypothesis three is rejected.

Table 5. ANOVA results for *decision-confidence* versus display type for each scenario.

Scenario	1	2	3	4	5
P Value	0.56	0.48	0.09	0.49	0.52

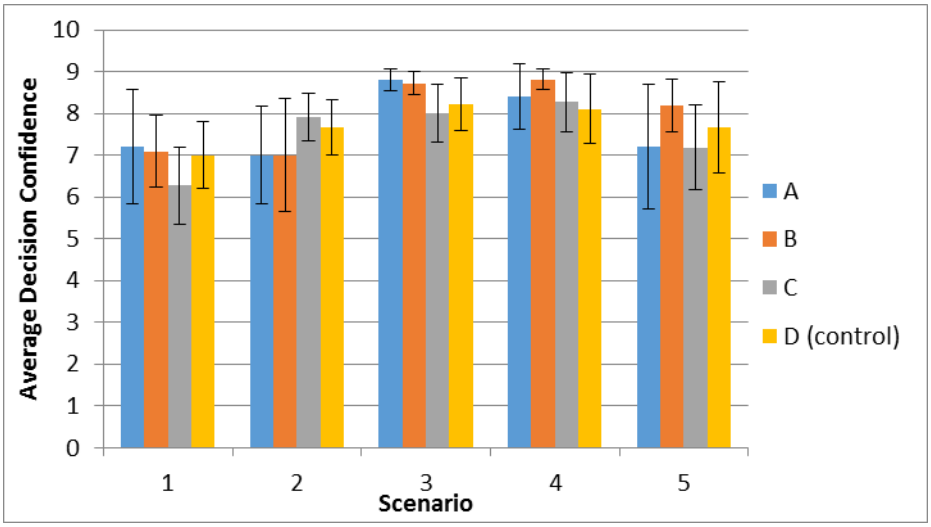

Fig. 5. Average decision-confidence per display type, per scenario, and plotted with 95% confidence intervals.

6 Discussion

This section will cover the key results from this study, and how they have impacted the design of the VIPARS interface.

1. Users wanted the most detail possible

Almost all users in high-level display conditions heavily utilized the low-level display as well. The author predicted the high-level displays would decrease the time a user needs, but the opposite was true. Users in the control condition had the lowest *time-to-decision*, though it was not statistically significant. The reason is obvious from test data, users spent time reviewing both levels of displays when they were available. As seen in Section 5.1, only condition group A (having the most specific high-level display) had a statistically significant reduction in *time-on-raw-data* compared the control group. This is consistent with previous work which found preferences for higher specificity [6].

2. The type of high-level display had almost no effect

There was no significant difference between the numeric, symbolic, or graphical displays except for *time-on-raw-data*. This is consistent with the previous work [6],[4]

that showed display format has little impact, but in this experiment specificity was also varied. Does this disagree with previous results that showed specificity had a significant effect? The authors do not believe so. In this experiment, users had access to a more specific information source in the low-level display. As most participants heavily utilized the low-level display, it seems likely that the variance in specificity at the high-level was overshadowed by the information from the low-level display.

3. The high-level displays helped reduce errors

Due to either misinterpreting the low-level graphs, or improperly combining the results of multiple criteria, more mistakes were made in the control group. While in a more realistic setting users would have additional training (reducing the likelihood of errors), the actual situations may be more complex and include several extra criteria (increasing the likelihood of errors).

Initial designs for the VIPARS user interface, and the prototype display for this experiment, utilized a layered system, where a user is presented with a high-level summary first, and would only view low-level detailed information if necessary. These results indicate that while a summary of results is useful, it likely should not be the primary focus. Instead, the complete verification results should be the primary output, with automatic summaries displayed alongside as a mental check for users. See Figure 6 below for an example design. The choice of display format for the summary is not critical, as no option showed superior performance, however results suggest users may prefer the numerical display due to its greater precision.

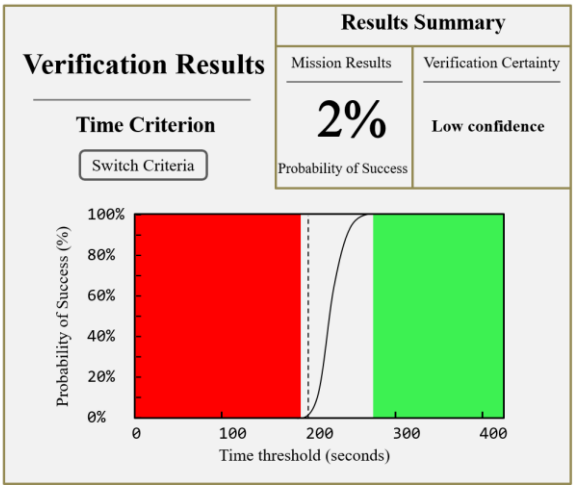

Fig. 6. New example display design that combines the high-level summary with the complete low-level results.

7 Conclusion

This paper has presented research on the display of uncertainty towards robotic mission verification. A human study on the display of probabilistic data for robot mis-

sions was performed. Three high-level summaries were chosen to present the results of a mission verification software toolkit. Surprisingly, the high-level summaries did not affect the time a user took, or their confidence with their decision. Instead, participants preferred to utilize the low-level detailed results. The control group, without access to the summarized data, made more mistakes. This implies some value in the high-level displays for the purpose of ensuring a user has accurately interpreted the verification results. The outcomes of this study have improved the design paradigm of the VIPARS interface; helping to ensures users will be able to quickly and accurately interpret the probabilistic information.

Acknowledgments. This research is supported by the United States Defense Threat Reduction Agency, Basic Research Award #HDTRA1-11-1-0038.

References

1. Doesburg, J.C., Steiger, G.E.: The Evolution of Chemical, Biological, Radiological, and Nuclear Defense and the Contributions of Army Research and Development. NBC Report, the United States Army Nuclear and Chemical Agency (2004)
2. Lyons, D. M., Arkin, R. C., Jiang, S., Liu, T. M., & Nirmal, P.: Performance Verification for Behavior-based Robot Missions IEEE Trans. on Rob. 31(3) (2015)
3. Tversky, A., & Kahneman, D.: Judgment under uncertainty: Heuristics and biaes. Science, 185(4157), 1124-1131 (1974)
4. Budescu, D. V., Weinberg, S., & Wallsten, T. S.: Decisions based on numerically and verbally expressed uncertainties. Journal of Experimental Psychology: Human Perception and Performance, 14(2), 281 (1988)
5. Wallsten, T. S., & Budescu, D. V.: A review of human linguistic probability processing: General principles and empirical evidence. The Knowledge Engineering Review, 10(01), 43-62.4 (1995)
6. Bisantz, A. M., Marsiglio, S. S., & Munch, J.: Displaying uncertainty: Investigating the effects of display format and specificity. Human Factors: The Journal of the Human Factors and Ergonomics Society, 47(4), 777-796 (2005)
7. Lyons, D. M., Arkin, R. C., Jiang, S., Harrington, D., & Liu, T. M.: Verifying and validating multirobot missions. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2014)
8. MacKenzie, D. C., Arkin, R. C., & Cameron, J. M.: Multiagent mission specification and execution. In Robot colonies (pp. 29-52). Springer US (1997)
9. Jhala R., Majumdar R.: Software Model Checking. ACM Computing Surveys 41(4) 21:53 (2009)
10. Cleveland, W. S., & McGill, R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387), 531-554 (1984)
11. Daradkeh, M., Churcher, C., & McKinnon, A.: Supporting informed decision-making under uncertainty and risk through interactive visualisation. In Proceedings of the Fourteenth Australasian User Interface Conference, Volume 139 (pp. 23-32). Australian Computer Society, Inc. (2013)
12. Masalonis, A., Mulgund, S., Song, L., Wanke, C., & Zobell, S.: Using probabilistic demand predictions for traffic flow management decision support. In Proceedings of the 2004 AIAA Guidance, Navigation, and Control Conference. American Institute of Aeronautics and Astronautics (2004)