In *Proceedings of the 11th Computer-Generated Forces and Behavior Representation Conference* (2002).
p. 553-562, 02-CGF-100. Orlando, FL: U. of Central Florida http://acs.ist.psu.edu/papers/avraamidesR02.pdf

Using Multidisciplinary Expert Evaluations to Test and Improve Cognitive Model Interfaces

Marios N. Avraamides
Department of Psychology
The Pennsylvania State University
University Park, PA 16802
+1 (814) 865-4455
marios@psu.edu

Frank E. Ritter
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
+1 (814) 865-4453
ritter@ist.psu.edu

Keywords:
Subject-matter experts, situation awareness, cognitive models

ABSTRACT: *Typically, the design of cognitive models has not emphasized the role of interfaces for describing the models' behavior. Models that populate synthetic environments are particularly complex and need support in this area. Using a variety of subject-matter experts we evaluated the use of the Situation Awareness Panel (SAP) as a tool for inspecting the behavior and reasoning of Soar agents in a JSAF simulation. We gathered suggestions on how to improve future implementations of the SAP from experts in a variety of disciplines, including military pilots, cognitive psychologists, an HCI specialist, a logistic specialist, and a software designer. Because of their diversity, we found that all were able to report unique problems with the interface, and thus now suggest about twice as many experts be used for evaluations than were previously suggested and that the experts should vary in their perspective. We used their behavior and reports to develop a task analysis that can be used as a general guide for future designs of user interfaces for cognitive models in general and for the design of interfaces for models in synthetic environments in particular. We suggest this approach of having multiple types of experts review an interface as a general method for improving complex interfaces such as interfaces to cognitive models.*

1. Introducing Evaluating Interfaces Using Multiple Expert-types

Newell's [9] call for a unified theory of cognition has led to a new way of research in cognitive psychology. During the last two decades many researchers have started to examine psychological phenomena through situation-specific theories that they implement as computer programs. These theories, called cognitive models, remain confined to the constraints imposed by grand theories of cognition, generally known as cognitive architectures.

Cognitive architectures and cognitive models have been used extensively as means for exploring the mechanisms involved in human cognition. However, cognitive models have been also built with the goal of being used as surrogate users. Models with human-like behavior can replace humans in many situations ranging from cognitive tutoring [1] to usability testing of interfaces [20, 23]

The use of cognitive models as surrogate users is especially appealing for situations where human expertise is costly or difficult to recruit. Military training has been one of those fields that typically require a great amount of human resources. Cognitive modeling provides an alternative avenue for supporting military training. Cognitive models as intelligent agents can populate synthetic environments representing some or all of the entities involved in real combats, thus enabling the use of realistic environments for training purposes [14, 19]. One such attempt has been the TacAir-Soar system [24] which employs cognitive models developed with the Soar cognitive architecture [7, 9] to simulate the behavior of military personnel in fixed-wing aircraft missions. The benefits of using TacAir-Soar are particularly evident in

Avraamides, M., & Ritter, F. E. (2002). Using multidisciplinary expert evaluations to test and improve cognitive model interfaces. In Proceedings of the 11th Computer Generated Forces Conference, 553-562, 502-CGF-002. U. of Central Florida: Orlando, FL.

large-scale simulation exercises in which many of the entities involved can be driven by Soar models instead of human users. For example, up to 3,700 computer-generated forces were involved as both friendly and enemy entities in the Stow '97 exercise [6].

While using cognitive models to either answer psychological questions or to replace human users provides great advantages, serious problems have been identified as well. One of the problems is the limited reuse of cognitive models. It seems fair to say that cognitive models are not typically used by researchers other than the ones who developed them. Part of the problem can be attributed to the lack of graphical-user interfaces for many of the models developed [17]. Without graphic displays, observing and understanding the behavior of the cognitive models is restricted, which can contribute to limiting their adoption by others.

The non-optimal design or the total absence of graphical displays that are needed to make the behavior of the models visible, make the validation of the models problematic as well. Subject-matter experts, who are not programmers themselves, have difficulties evaluating the behavior of the model based on traces of the running program. In order to understand the model, these users need a clearer form of output.

The need for improved user interfaces for cognitive models is particularly important for models that populate synthetic environments. These environments are typically loaded with such a great number of computer-generated forces that their behavior must be easily observable if it is to be understood. As Ritter, Jones, and Baxter [17] point out, graphical user interfaces have led in the past to new understanding about the behavior of models. When a graphical interface was added to Soar [18], it became evident that the Soar models searched through the problem spaces hierarchically rather than spending much time searching in a single one.

The present study tested one such graphical interface [5]. We recruited a number of subject-matter experts from a wide variety of related fields and asked them to observe the behavior of Soar agents that fly fixed-wing aircraft missions in a JSAF simulation. Their understanding of the model in some cases led to a type of Turing-like test, where they were attempting to judge if the model' s behavior was similar to a human' s.

We have used comments on ways to improve the specific graphical interface as well as their behavior with the interface to provide a list of suggested improvements. We have also created a task analysis that can be used to improve the design of the interface we studied and of modeling interfaces in general, based on their comments and our own experience with models.

2. Expert Evaluations of the Situation Awareness Panel (SAP)

The goal of the project was to understand and improve the Situation Awareness Panel (SAP) [5] as a tool for inspecting the behavior and reasoning of the Soar agents that populate the JSAF simulation environment. Our attempt was focused on the validity and usability of the SAP, but our results make suggestions for other interfaces and for other modeling architectures.

Validity refers to whether the type of information displayed on the SAP is truly in the awareness of actual pilots engaged in air combat. Usability refers to whether people using the SAP can understand the model based on what they can see through the SAP.

In order to examine these issues we recruited people with a variety of expertises and asked them to perform a number of basic tasks while observing the awareness of the agents during a preprogrammed scenario. In addition to expert pilots, our list of participants included experts from various other domains that we thought were related to different aspects of the JSAF simulation. Such domains included cognitive psychology, geographical information systems (GIS), human-computer interaction (HCI), software development, and the military. Our goal was to assemble a multidisciplinary pool of experts in order to get feedback about the functioning of the SAP from a variety of perspectives.

This work is similar to a variety of evaluation techniques, including heuristic evaluation, cognitive walk-throughs, and semi-structured interviews with the addition that we used a wide variety of experts. This work created in a task analysis as a result rather than being based on one.

2.1. Cognitive walkthroughs, heuristic evaluation, and semi-structured interviews

There is a variety of approaches that we could have used to examine the usability of the SAP interface. We could have done a task analysis [21] if we had a list of what tasks users were performing, but we were attempting to create such a list. We were not trying to optimize performance, per se, so timing users on tasks was not appropriate either. We were not just looking at learning of it, as our users would either just be watching it for the first time in a demo (with very little learning, we would hope), or would be working with it for a while (with quite extensive learning) and did not have a task analysis in hand, so cognitive walkthroughs [15] seemed not quite appropriate either.

In the end, we did what could be described as guided heuristic evaluation. We prepared a subset of tasks that

we knew the interface would be used for. We had potential and existing users and a variety of usability experts (broadly defined) perform these tasks with the interface. We observed these users and also had them comment on the problems they had. After performing these tasks, we debriefed them in order to find out what other tasks they would like to have been able to perform with the interface. In some cases, they could do these tasks, in others, we were able to add these tasks to our developing task analysis. In some ways, our approach was similar to semi-structured knowledge acquisition interviews (see, for example, SigArt ACM Special Interest Group on Artificial Intelligence [22]).

We believe that using a multidisciplinary participant pool for validating interfaces of military simulations is a necessity due to the variety in the nature of the information that is typically displayed. For example, a situation-awareness display, such as the one we evaluated, contains information that varies from the execution of standard combat routines to awareness about the terrestrial terrain, memory for past events, perception of various sorts of input, aircraft logistics (i.e., fuel and weapon status) and so on. Instead of relying on our own common sense to evaluate the way the various types of information are presented, we have employed subject-mater experts from fields that relate to the nature of information contained in the interface. We believe that this approach is preferred over relying on common sense and we agree with Jones et al. [6] in that "…what is common sense to an experienced pilot is quite different from the common sense of an AI researcher" (p. 8).

2.2. The Situation Awareness Panel

The Situation Awareness Panel is a graphical tool that enables the user of the JSAF simulation to observe a Soar agents' understanding of a situation, their goals, and their history [5]. The JSAF environment is depicted in a Plan View Display (figure 1).

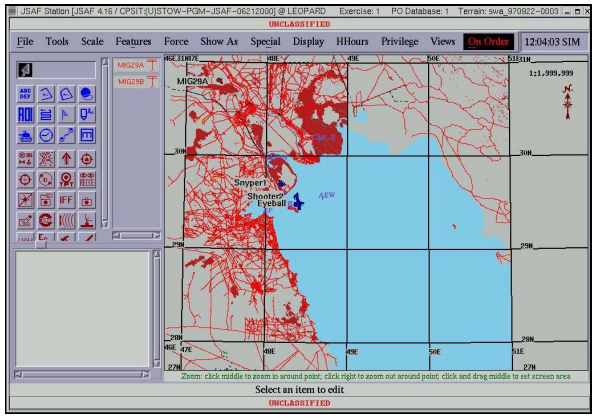

Figure 1. A screenshot of JSAF's Plan View Display.

This map-like display depicts along with features of the terrain, the agents -- both friendly and hostile -- that are involved in the simulation. Each of these agents is driven by a Soar model. The user can observe each model's external behavior by inspecting the Plan View Display but also examine its internal state by examining its SAP. The SAP is, in essence, a window into the Soar agent' s awareness of the current situation. Figure 2 shows a screenshot of the SAP of agent Shooter2.

The SAP is useful for examining and verifying the behavior of the Soar agent. With the SAP, the user is able to observe the awareness of a model and do things like examining whether the model' s behavior reflects its understanding of the situation or its intentions for action, evaluating the model' s current actions within the context of its history, and so on.

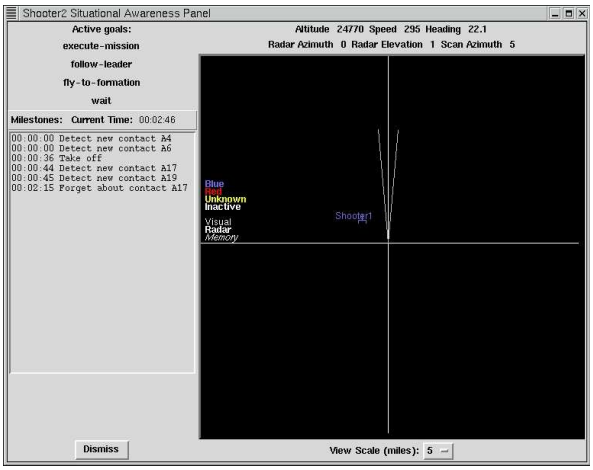

Figure 2. A screenshot of the SAP.

The depicted version of the SAP is realized in Tcl/Tk. Tcl/Tk is an extension language [13] that is jointly compiled with Soar. The SAP interacts with Soar agents running in a variety of JSAF scenarios. Any of the Soar agents can be explored with it, and there is nothing to preclude it being applicable to other Soar agents, although the plan-view display would not be useful for many of them.

Detail about the functions of the various displays of the SAP is provided by Jones [5]. In short, the SAP is divided into four displays.

(a) The Active Goal Display is located at the top left part of the SAP and it contains the model' s current stack of goals and subgoals. By enabling the user to observe what the model is trying to achieve at any moment, a comparison of the internal intentions of the model and its external behavior can be made.

(b) The Milestone Display is located underneath the

Active Goal Display. Each milestone event is recorded as a new line in the window. A time stamp for each milestone event is also recorded. This display enables the user to review quickly the model' s past activity and reasoning.

(c) The Aircraft Status Display is located at the top of control panel and stretches to the right. It is a short strip that provides some basic aircraft information that is available to the model. The Altitude, Speed, Heading, Radar Azimuth, and Radar Elevation measurements are displayed in this strip.

(d) The Agent Awareness Display occupies the rest of the SAP. This display enables the user to inspect the current state of the model' s awareness. It is basically a view of the model' s reasoning about what is going on in its world (which is not necessarily an accurate depiction of what is really going on). Entities with which the agent had contact (through vision, radar, or radio) are all represented in the display and marked with different colors to indicate whether they were friendly, hostile, unknown, or inactive. The type of contact is represented by different type styles. The user can adjust the scale of the Agent Awareness Display by choosing a different number from the "View Scale" drop-down menu.

2.3 The Study

To analyse the SAP and suggest improvements to it we had a variety of experts interact with it. The primary focus of the study was to determine the success of the SAP interface at revealing to users the reasoning of the agents and pinpoint their limitations, particularly in assessing whether the reasoning of the Soar models was realistic.

The feedback we got from our subjects enabled us to determine ways of improving the visual interface of the SAP in future designs. Questions that required the subjects to interact with the simulation in order to initiate some action provided a way for our subjects to evaluate the reasoning and the behavior of the agent, while it was engaging in action to achieve or prevent a user-initiated goal.

Participants. The participants were twelve experts coming from different disciplines. Table 1 lists the area of expertise of our participants. The first eight participants -- several of which are faculty members of the School of IST -- completed the study in our laboratory, while the last four did so under the same equipment but on the site of their employment[1]. All participants received monetary reimbursement in exchange for their participation. Participants were run individually with each experimental session lasting between an hour and two hours.

It should be noted that only subject-matter expert K was a prior user of the SAP. Including an actual user in our subject pool allowed us to examine whether the problems identified by the other experts are predictive of end-user problems. Although having just one actual user does not allow us to draw definite conclusions, it at least gives us an idea of the degree of overlap between problems identified by inspectors and those that are encountered by users

Table 1. List of expert participants.

	Area of expertise
A	Plan view/geographic information systems specialist.
B	Graduate student in AI and cognitive modeling.
C	Marine Major, specializing in logistics and infantry.
D	Former software developer in Silicon Valley with Fortune 100 companies.
E	Former merchant marine officer and expert on social and group processes.
F	Navy fixed and rotary wing pilot. RWA instructor.
G	Cognitive psychologist.
H	Cognitive psychologist with some amateur flying experience.
I	Former military aviator from BMH Associates.
J	Former military aviator from BMH Associates.
K	Former military aviator from BMH Associates.
L	Former military aviator from BMH Associates.

[1] The last four participants also observed a new and improved version of the SAP that became available to us shortly before the time they were run. Identified problems that were unique to the new SAP are presented in Ritter & Avraamides [16] and are not included in the present analyses.

Materials and equipment. The JSAF simulation environment was presented on a 19-inch monitor attached to a Dell Optiplex computer running Red Hat Linux 6.1. All experimental sessions were videotaped with the use of a SONY TRV-120 Hi-8 digital camcorder. In addition, the computer desktop activity of the first eight subjects was videotaped on VHS tape. Participants read and signed an informed consent form prior to the beginning, and they were debriefed upon completion of the study according to the IRB guidelines.

Procedure. Each experimental session started by providing the participant with a short description of the SAP taken from Avraamides' s manual [2] and a description of the scenario within the JSAF environment that would executed by the models. Soar Technology provided us with three pre-programmed scenarios, from which we selected the Defensive Counter Air (DCA) scenario for our testing purposes. Using a prescribed scenario allowed us to evaluate the SAP in a way that corresponds to how it will be used by actual users. As Nielsen and Mack [11] point out, scenarios provide a task-oriented perspective on the interface and ensure that certain interface features will be evaluated. The description of the DCA scenario given to subjects before the study read as follows:

> "The Defensive Counter Air mission involves defending an area against airborne threats. An Airborne Early Warning (AEW) aircraft is used for its long-range radar to watch for distant threats. When threats arise, the AEW dispatches an airborne 2-ship flying a Combat Air Patrol (CAP) to engage the bogeys.
>
> Of interest:
>
> - The fighters's Situation Awareness Panel (SAP) will demonstrate the agent's attention to the overcoming air threats
>
> - Coordination between AEW and fighters"

The study was divided into three parts. The goal of the first part was to familiarize our subjects with the apparatus and our data collection methodology. We therefore asked them to perform a number of basic actions (e.g., "As soon as the Plan View Display becomes visible, zoom into the map and locate the position of the AEW"). These simple tasks were introduced in order to guide our subjects to explore the various windows of the JSAF simulation and learn some important functions, such as zooming into the map with either pressing simultaneously the two mouse buttons or using the Scale menu. We believe that some familiarity with the other windows of JSAF is needed in order to make possible the efficient use of the SAP. For example, inspecting the Plan View Display a user could determine which of the agents is more likely to have a target in its awareness, and then use the SAP to examine this feature. This part lasted longer for the first eight participants because they were completely unfamiliar with JSAF.

The second part of the study involved tasks that required that subjects observe the four displays of the SAP at the time the agents were engaged in combat as defined by the on-going scenario. Subjects were asked a variety of questions that differed in terms of both what they were required to do and what aspect of the SAP was brought into focus. Some of the questions required that subjects simply observed what was going on (e.g., "What friendly agents are in Eyball's awareness and what is their status?") and others required some interaction of the subject with the interface (e.g., "Select 'MIG29 FWA sweep base' and find the agent that you think is the most likely to have an enemy plane in its awareness").

Observation questions were aimed at assessing whether the SAP was successful at conveying the information that it was supposed to convey. We were primarily interested in seeing whether our subjects could easily pick up from each panel of the SAP the information that the SAP was meant to provide. Difficulties with and sometimes misunderstandings of information were of particular interest since they provide points to consider for future implementations of the SAP.

Finally, we allowed our experts to provide further comments on the SAP out of the context of the scenario they observed. As Nielsen [10] points out, there are some advantages to giving evaluators open-ended instructions. For example, more diverse aspects of the interface can be examined in the absence of a prescribed scenario.

2.4 Analysis and Results

All videotapes were reviewed at a later time by the experimenters and a list of potential problems with the SAP interface was generated. The majority of the problems that were reported came along with suggestions for fixes. This is in line with discussions by Jeffries [4] and Desurvire [3], who point out that typically knowing about a usability problem is sufficient for finding an obvious fix for it. The problems noted along with feedback from our subjects on how to deal with them enabled us to generate a set of suggestions for the improvement of the interface. Suggestions are subjective to the experimenters but they depend wholly on feedback obtained from our subjects.

Because our study focused on providing feedback for the improvement of the SAP, the present paper does not address any of the positive feedback received from our participants, which was substantial.

The experts found between 3 and 13 problems, with an average of 6.83 problems per expert. These could be aggregated into a total of 35 unique problematic issues for the SAP display. A detail list of those problems along with suggestions for overcoming them is presented in Ritter and Avraamides [16]. This list was passed back to the developers and it has been used in generating new versions of the SAP and related displays [25].

We computed how many unique problems would be found, on average, as the number of experts increased. We did this by looking at all the possible combinations of our experts (as sets without order). Figure 3 shows the average number of unique problems that would be found as the number of experts increased. Figure 4 shows a similar calculation for the average number of unique problems found per expert as more experts look at our interface.

There is not an obvious bend in these curves, although clearly, the most problems are found on average by the first expert, and this is a monotonically decreasing function. A bend would indicate when the payoff of adding another expert became notably less helpful. In this case, however, even the last expert in a series of 12 was able to report unique problems.

As noted earlier, participant K was a user of the SAP. He identified 8 problems, a number that is slightly higher than the average of 6.83 problems that are identified on average by a single expert (figure 3). From these 8 problems, 5 were also identified by at least one of the other participants from BMH Associates (i.e., participants I, J, and L), and 1 problem was spotted by other participants as well. Participants I, J, and L identified 18 problems in total. Six out of the 18 problems were also spotted by participant K. In summary, the SAP user spotted 33% of the problems that were identified by evaluators with a similar background. Only 37% of the problems identified by the user were unique. The remaining 63% of his problems were also spotted by other subjects. Overall, these results suggest that our subject-matter experts from BMH did fairly well at identifying problems that are predictive of the problems encountered by users.

A closer examination of the overlapping problems suggests that these were problems that related to military aviation expertise (e.g., "a negative heading measurement is not meaningful"). Problems of this nature were not typically spotted by subject-matter experts without cockpit experience. In total, only 33% of the problems identified by at least one of the last four participants were also spotted by at least one of our remaining subject-matter experts. The overlapping problems were problems of a more general nature and did not rely on fighter-plane pilot experience (e.g., "labels for the status of agents cannot be distinguished easily").

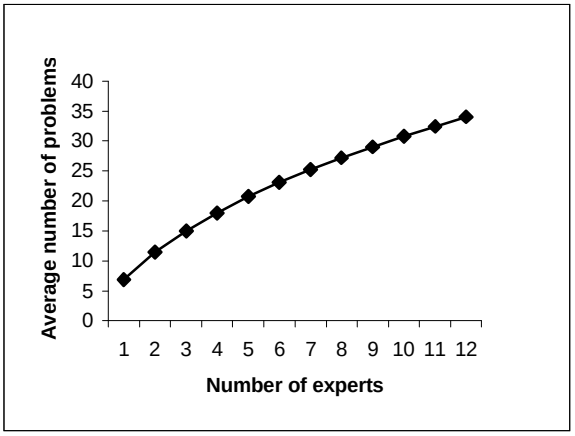

Figure 3. Average number of identified problems per number of subjects ran.

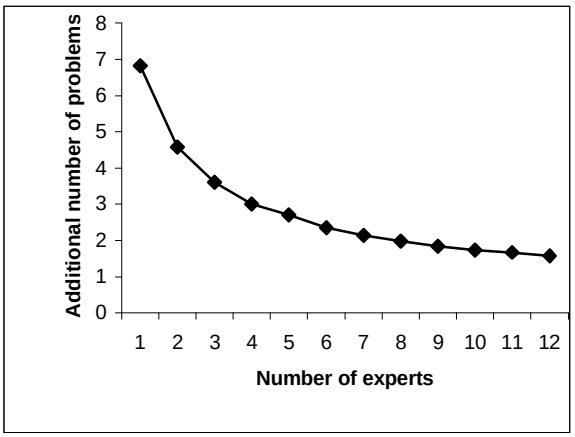

Figure 4. Increase of average number of identified problems for each additional expert.

The absence of a complete overlap between the problems identified by the SAP user and the rest of the former-pilot subject-matter experts was expected. In many cases, inspection problem reports fail to predict end-user problems, producing thus *false positives* [4]. Although in our study the proportion of false positives was rather large (67% when counting only participants I, J, and L), the fact that we had just one SAP user in our subject pool casts doubts as to whether these are indeed false positives. Assuming that among the potential users of the SAP will be people with no combat aviation experience (e.g., programmers), it should be expected that additional problems to the ones pointed out by participant K will be encountered by other users. A laboratory usability test (i.e., using just users) was beyond the scope of this

project, but it might be necessary if we wish to accurately assess the validity of our problems.

However, the fact that we have included a rather large number of subject-matter experts in our participant pool might be all that is needed to assume that many of the problems our subjects spotted are indeed problems that SAP users would encounter. Nielsen [11] suggests that with only 4 or 5 evaluators the majority of usability problems can be identified, and presumably with more experts the problems that are spotted by users can be approximated.

Nielsen [11] recommends the use of 4 or 5 evaluators and at least 3. With his *discount usability* perspective, he argues that 80% of the total usability problems can be spotted with 4 or 5 evaluators. Our figure 3 shows that our curve is somewhat shallower and we only reach 80% somewhere between 8 and 9 evaluators. As figure 4 shows, we continue to find ways to improve the interface by adding more evaluators well after 5 evaluators. In fact, gains are obtained even after adding the 12th evaluator, although, just like in Nielsen, our curve tends to asymptote as we add more experts.

The rather obvious cause for our steeper curve is the fact that our pool of experts consisted of experts from rather diverse backgrounds. As a result, many of the problems identified were unique to one expert. Indeed, 21 out of the total of 35 problems were spotted by only one expert. This explanation is also supported by the fact that in our study, a single subject identified on average about 20% of the total problems (6.83 out of 35 problems), while other studies [8, 12] report an average of up to 35% of identified problems by a single user.

If a significant proportion of the problems identified by our subject-matter experts are indeed problems that would be encountered by users, our results suggest that there is an advantage from using evaluators that come from different backgrounds. Our results suggest that using multidisciplinary experts allows the examination of an interface from various perspectives and provides a more comprehensive problem-report list.

2.5 The task analysis

Based on the feedback we got from our participants and our experience building interfaces for cognitive models [17, 18], we created the task-analysis shown in the Appendix. We believe this task analysis can be used as a guide for designing interfaces for cognitive models of military content. These tasks include what all users need to know to understand their models, so interfaces that supported these tasks would also be useful more generally.

This task analysis includes many user tasks that would be expected. Making the perceptions and actions of the model visible by analysts will not be a surprise to most modelers. Likewise, access to the mental environment of the model should not be surprising, but this is not fully supported by every modeling environment. Similarly, because models are increasingly becoming embodied and subject to their environment, the modelers need to know what aspects of the environment influence a model.

What is somewhat novel, is suggesting that the social environment of the model should be explicitly explained. These agents clearly have social aspects to their behavior and reasoning. This appears to be a different type of knowledge and processing, a type that interfaces should make available to modelers. The mental models of other agents is of increasing importance for cognitive models as they become team members, and this is particularly true for models in synthetic environments that need to understand colleagues and advisories. Finally, the model and the modeler need to keep in mind aspects of the environment related to their specific task. In the case of these models, the domain is a military one. Other models are likely to require additional information related to their domain of performance.

3. Conclusions

Using cognitive models as surrogate users in military simulations provides the capability of training military personnel even individually. In a JSAF simulation, thousands of entities can be represented as computer-generated forces providing the feel of a realistic environment without the need to recruit great numbers of humans to participate in the simulation. The success of simulators depends greatly on how realistic the behavior of the cognitive models is.

Graphical interfaces that make the behavior of these models visible to human users support the validation and the improvement of these models. So far, not much emphasis has been placed on the design of graphical interfaces for cognitive models. Even when graphical interfaces have been supplied along with models, they have been designed based on the "common sense" of their developer.

We believe that graphical interfaces are very important for cognitive models. By making their internal state more visible, graphical interfaces allow a better understanding of the reasoning and actions of the model and therefore lead to easier debugging, better validation, and more powerful demonstrations of the models.

Given the importance of graphical interfaces, we argue that they should undergo testing and revision to improve

their usability. The present study provides an example of how this can be done. It suggests that for interfaces with a wide variety of types of users, a wide variety of people can fruitfully examine it to help find problems.

Our resulting task-analysis can be used to guide the design of future modeling interfaces. Keeping this analysis in mind and extending it further will help design better interfaces so that they support the user performing their tasks and reduce the need for usability testing.

References

[1] Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R.: Cognitive tutors: Lessons learned. Journal of the Learning Sciences, Vol. 4, pp. 167-207, 1995.

[2] Avraamides, M. N.: A brief manual for running a JSAF Demo and examining the Situation Awareness Panel (Tech. Note No. 2001-1). Applied Cognitive Science Lab, School of Information Sciences and Technology, Penn State, 2001.

[3] Desurvire, H. W.: Faster, cheaper!! Are usability methods as effective as empirical testing? In J. Nielsen & R. L. Mack (Eds.), Usability Inspection Methods, pp. 173-202, New York, NY: Wiley & Sons, 1994.

[4] Jeffries, R.: Usability problem reports: Helping evaluators communicate effectively with developers. In J. Nielsen & R. L. Mack (Eds.), Usability Inspection Methods, pp. 273-294, New York, NY: Wiley & Sons, 1994.

[5] Jones, R. M.: Graphic visualization of situation awareness and mental state for intelligent computer-generated forces. In Proceedings of the Eighth Conference on Computer Generated Forces and Behavioral Representations, pp. 219-222, Orlando, FL: Division of Continuing Education, University of Central Florida, 1999.

[6] Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., and Koss, F. V.: Automated intelligent pilots for combat flight simulation. AI Magazine, Vol. 20, pp. 27-41, 1999.

[7] Laird, J. E., Newell, A., and Rosenbloom, P. S.: Soar: An architecture for general intelligence. Artificial Intelligence, Vol. 33, pp. 1-64, 1987.

[8] Molich, R., and Nielsen, J.: Improving a human-computer dialogue. Communications of the ACM, Vol. 33, pp. 338-348, 1990.

[9] Newell, A.: Unified theories of cognition, Cambridge, MA: Harvard University Press, 1990.

[10] Nielsen, J.: Heuristic Evaluation. In J. Nielsen & R. L. Mack (Eds.), Usability Inspection Methods, pp. 25-62, New York, NY: Wiley & Sons, 1994.

[11] Nielsen, J., and Mack, R. L.: Usability Inspection Methods, New York, NY: Wiley & Sons, 1994.

[12] Nielsen, J., and Molich, R.: Heuristic evaluation of user interfaces. Proceedings ACM CHI'90 Conference, pp. 249-256, Seattle, WA, 1990.

[13] Ousterhout, J. K.: Tcl and the Tk toolkit, Reading, MA: Addison-Wesley, 1994.

[14] Pew, R. W., and Mavor, A. S. (Eds.).: Modeling human and organizational behavior: Application to military simulations, Washington, DC: National Academy Press, 1998.

[15] Polson, P. G., Lewis, C., Rieman, J., and Wharton, C.: Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. International Journal of Man-Machine Studies, Vol. 36, pp. 741-773, 1992.

[16] Ritter, F. E, and Avraamides, M. N.: Improving interfaces for CGF's through multidisciplinary evaluations: A new, broad approach, (Technical Report No. 2001-1), Applied Cognitive Science Lab, School of Information Sciences and Technology, Penn State, 2001. URL: http://acs.ist.psu.edu/acs-lab/reports/ritterA01.pdf

[17] Ritter, F. E., Jones, R. M., and Baxter, G. D.: Reusable models and graphical interfaces: Realising the potential of a unified theory of cognition. In U. Schmid, J. Krems, & F. Wysotzki (Eds.), Mind modeling - A cognitive science approach to reasoning, learning and discovery, pp. 83-109, Lengerich, Germany: Pabst Scientific Publishing, 1998.

[18] Ritter, F. E., & Larkin, J. H.: Using process models to summarize sequences of human actions. Human-Computer Interaction, Vol. 9, pp. 345-383, 1994.

[19] Ritter, F. E., Shadbolt, N. R., Elliman, D., Young, R., Gobet, F., and Baxter, G. D.: Techniques for modeling human performance in synthetic environments: A supplementary review. Wright-Patterson Air Force Base, OH: Human Systems Information Analysis Center, in press

[20] Ritter, F. E., and Young, R. M.: Embodied models as simulated users: Introduction to this special issue on using cognitive models to improve interface design. International Journal of Human-Computer Studies, Vol. 55, pp. 1-14, 2001.

[21] Schraagen, J. M., Chipman, S. F., and Shalin, V. L. (Eds.).: Cognitive task analysis. Mahwah, NJ: Lawrence Erlbaum, 2000.

[22] SigArt ACM Special Interest Group on Artificial Intelligence, Knowledge Acquisition. Sigart Bulletin (Special issue), 1989.

[23] St. Amant, R.: Interface agents as surrogate users. Intelligence, Vol. 11, pp. 29-38, 2000.

[24] Tambe, M., Johnson, W. L., Jones, R. M., Koss, F., Laird, J. E., Rosenbloom, P. S., and Schwamb, K.: Intelligent agents for interactive simulation environments. AI Magazine, Vol. 16, pp. 15-40, 1995.

[25] Taylor, G., Jones, R. M., Goldstein, M., Frederiksen, R. : VISTA: A GenericToolkit forVisualizing Agent Behavior. To appear in Proceedings of the 11th Conferenceon Computer Generated Forces and Behavioral Representation. Orlando, FL. 2002

Acknowledgments

We thank all of our subject-matter experts for participating in this study. The experts at BMH Associates were particularly helpful at a distance. Glenn Taylor and others at Soar Technology provided valuable help in installing and maintaining JSAF on our equipment. Lael Schooler suggested the analysis presented in Figure 3. Robert St. Amant provided useful comments on this manuscript.

Author Biographies

MARIOS AVRAAMIDES is a doctoral candidate in Cognitive Psychology and a research assistant for the School of Information Sciences and Technology at the Pennsylvania State University. He has previously obtained an MS in Cognitive Psychology from Penn State, and a BA in Psychology from the University of Texas at Austin. His current research examines how people update spatial information provided in texts. In the School of IST, Avraamides has worked on several projects building and supporting cognitive models, including helping update the Soar FAQ.

DR FRANK RITER is one of the founding faculty of the School of Information Sciences and Technology, a new interdisciplinary academic unit at Penn State to study how people process information using technology and to train leaders for the digital economy. Ritter works on the development, application, and methodology of cognitive models, particularly as applied to interfaces and emotions. Ritter is a member of the editorial board of Human Factors, and is on the board of the UK' s Society for the Study of AI and Simulation of Behaviour (AISB). His review (with others) on applying models in synthetic environments will be published as a book this year by HSIAC as a State of the Art Report.

Appendix. Set of tasks found in this analysis method.

Perception (Inputs) - What inputs does the model get?

Inputs does the model get from instruments
- Radar and IFF values (if from display), and visual input
- Voice input/communication from other agents
- Other perceptual events

Constants in perception, e.g., due north
Self-perception, physical status of pilot: healthy, tired, bored
Where is our agent's attention (for analyst) - perhaps with a spotlight metaphor (this was used by Chong in the AMBR project to good reviews)

Actions (Outputs) - What actions has the model done?

What plane/pilot/RIO has said and done
- details of those actions if complex

What milestones are there, and what's the range of types of milestones, i.e., what could have been there but are not, and thus why are they not?

Physical environment features that affect the agent's body

Weather
Terrain, including base location, feet wet/dry, ground threats, places to land for RWA
Unknown but suspected ground threats will be an interesting thing to display

Mental environment -- Current Goals and Active Plans

Active goals, and their current status
Inactive goals, and why inactive (complete list of all possible goals and plans, and their status)
Details of the goals
Remaining steps in a goal/plan with associated physical location
Distance to target or other key events that agent would keep in attention and update, such as time left on CAP (thus timing)
Long term memory contents and active elements
Structure of memory and other mental objects
Contents of short term memory
Contents of perceptual (iconic) memory
Capacity remaining in each capacity, e.g., working memory, idle (slack) time in central processor.
Pop-up display of changes/targets of goals for turning, climbing, accelerating, i.e., when the plane starts to do any of these, a pop-up window appears over or in the SAP indicating what is being attempted
An articulate model that comment on its behavior
- what other operators were available
- why operators were or were not chosen (cf. Lewis Johnson's work)

Social environment

Cultural/political/historical facts that influence behavior (declarative facts)
Rules of engagement (perhaps available but not displayed if they don't change often)
Other social context of team, broadly defined

Mental models of other agents

(actual vs. mental may indeed be different)
x, y, z, heading, roll, yaw, pitch, speed, weapons
dx, dy, dz, d(heading),d(roll, yaw, pitch, speed)
Model of what the other planes are and what they are doing and what they are going to do (this list repeated one level down based on what they think you are going to do!)
What other planes have said
What other planes have been told, perhaps from a specific range of time
What other planes can see (their radar might not be as good, and you might know it)
Physical status of other plane, damaged or not, fuel status, munitions, etc.
Physical status of other plane pilot, healthy, tired, bored

Military environment (task and hardware of own agent)

Written instructions
x, y, z, heading, roll, yaw, pitch, speed, weapons
dx, dy, dz, d(heading),d(roll, yaw, pitch, speed)
What other planes have said to agent
Physical status of own plane, damaged or not, fuel status, munitions, etc.
Physical status of pilot, healthy, tired, bored
Munitions capabilities if novel (otherwise, assumed or reconstructed), and range