

User-Centered Evaluation of Technosocial Predictive Analysis

Jean Scholtz and Mark Whiting

Pacific Northwest National Laboratory
PO Box 999, Richland, Washington 99352
{jean.scholtz| mark.a.whiting}@pnl.gov

Abstract

In today's technology filled world, it is absolutely essential to show the utility of new software, especially software that brings entirely new capabilities to potential users. In the case of technosocial predictive analytics, researchers are developing software capabilities to augment human reasoning and cognition. Getting acceptance and buy-in from analysts and decision makers will not be an easy task. In this position paper, we discuss an approach we are taking for user-centered evaluation that we believe will result in facilitating the adoption of technosocial predictive software by the intelligence community.

The Use of Models and Simulations

The use of technical models to understand diverse domains such as weather, ocean behavior, military technologies, power and energy consumption, etc. has been employed for many years by researchers and those involved in military operations. In recent years, the intelligence community has also begun to use models and simulations, but sporadically at best. There are several issues with the use of models and simulations. First, models and simulations are difficult to understand and usually require interpretation by an expert in both the domain and the use of models. Secondly, to vary parameters and rerun models, even with the high-speed computational resources today, takes a good deal of time. Given these issues, why should the intelligence community be interested in using models and simulations in their analytic work? In the 2007 winning Galileo paper, Hanig and Henshaw (2008), propose that simulations would improve integration and collaboration within the intelligence community, would be a way for outside experts to share their knowledge with analysts, and would help the intelligence community better anticipate complex developments in the geopolitical arena. Additionally, simulations would be useful for helping to train new analysts.

Adding in the Social Aspects

While the use of technical models could certainly enhance the rigor of analysis, the geopolitical world consists not only of technology but also of people. Blending technical and social models is one way to provide more expertise to the intelligence community. The use of social models has increased since the events of 9/11. For example, the Terrorist Risk Model produced by Rand researchers (Willis et al. 2007) developed a probabilistic risk assessment model to help the Department of Homeland Security better utilize their resources. Another social model, described in *Visualizing the Political Landscape* (Dixon and Reynolds 2005), helps analysts understand political factions. Technosocial models combine both the technical models and the social models to help analysts understand possible outcomes and the degree to which varying different parameters in the models impacts these outcomes.

Introducing these models into the intelligence community will be challenging. This constitutes a fundamentally different way of doing analysis. As mentioned earlier, the use of models and simulations in the intelligence community is currently not a common analytic technique. Adding the social models to this will be a big deviation from the more typical analytic procedures. Additionally, in the case of our particular project, we will be adding a gaming environment for users to interact with the models, a knowledge repository, containing the information used in the models, and a trace of user interactions with the various simulations. In order to facilitate the transition of these new analytic environments into the community, we propose using user-centered evaluation.

User-Centered Evaluation

User testing normally focuses on the usability of a software system. That is, usability evaluations use representative users and give them tasks that involve using the software to complete a software related task. Effectiveness, efficiency, and user satisfaction are measured. For example, a typical word processor task would be to "edit line 30 in document A, changing the font from Times Roman to Courier." The

measures collected would be if the user could complete the task (effectiveness), how long it took the user to complete the task (efficiency), and how satisfied the user was with her completion of the task. While these evaluations can assess whether the users can do these specific tasks, we have little information about the importance of these tasks to the user. Would users actually do these tasks during their normal work? How would such tasks fit into their current work process? For the case of systems such as word processors, we are reasonably confident about the usage of lower level tasks. To adequately evaluate software used for solving complex tasks, we need to devise higher level, realistic tasks that users would actually do. It is difficult to conduct such evaluations due to the time needed to complete the tasks and the difficulty evaluators have in understanding the domain and constructing the appropriate evaluation tasks (Redish 2007). While usability evaluations are necessary for software use, they are not sufficient. We argue that it is necessary to conduct what we call user-centered evaluations, especially for research software that will significantly change the current processes of the anticipated end users.

User-centered evaluation goes beyond typical usability testing. Not only must software be usable so that the functionality is usable, the functionality must provide significant utility to the end user population (Scholtz 2006). For software that is to be used in critical situations, it is also necessary that the end users trust the software output. This does not necessarily mean that there is no uncertainty that results from the use of the software. But it does mean that there has to be sufficient transparency so that the end user understands the inputs and algorithms used in the software to correctly interpret the output.

Utility can be measured in a number of ways, including cost of use, the quality of the product produced using the software, and the confidence in any recommendations contained in the product. Cost can be the amount of time and expertise needed to produce an analytic product. The quality of the product could include more situations or hypotheses investigated or new insights gained. Confidence in recommendations is a more subjective measure. An analyst might have increased confidence due to her ability to analyze more situations, investigate more hypotheses, or marshal more evidence.

Analysts are actually intermediaries in the intelligence process. They need to produce analytic reports for customers. To this end, they need to provide the justification for interpretations, judgments and recommendations. Assume that we use as a base case an analyst reading a number of documents and producing an analysis of a situation. The analyst cites certain facts or evidence from these documents, adds her assessment and explains the rationale based on her knowledge of the area or people involved. If we expect analysts to use technical and social models as sources for exploring situations, then the analysts will need to produce explanations. Analysts' understanding of the system rationale is essential for them to produce quality explanations.

Trust is the most difficult and most subjective measure. We trust the judgment and advice of people based on past performance. One means for analysts to trust software systems is for them to observe the software's performance on data where an outcome has already occurred or been agreed upon by experts. Another possibility is for analyst to observe others they trust using the software. (Moore 2002). Another possibility (and the one we plan to use in our work) is that analysts' will trust the technosocial models based on reviews from domain experts.

Possible Approaches to User-Centered Evaluation of Technosocial Analytics

One possible way to measure utility is to do a comparative study. That is, analysts could use a current method and then use the technosocial models embedded in a software environment. We could measure the time to produce a product; the number of analysts/experts needed and the quality of the product using expert reviews. The problem is that we need to have two comparable analytic questions OR we need to have two sets of comparable analysts solve one question, with one group using the current method and the other group using the new software. Training and expertise inequalities quickly become confounding factors.

If we agree on certain measures of utility as being more important, then the evaluation task becomes somewhat easier. For example, if we agree that cost is important, then we might show that the same quality report can be produced with fewer analysts or with more junior analysts. If we focus on more insights as a desirable measure of utility, then we might have analysts look at the problem using current methodologies, then look again using the technosocial methodology and see if they find more relations, connections, or possible outcomes. This type of study could be enhanced if we used data in which we embedded relationships and connections. We could then have a quantitative measure of "insights."

The big drawback with these types of evaluation is that they depend on the system being fully developed. Our goal is to do user-centered evaluations as soon as possible to provide the necessary feedback for researchers to make revisions before investing too much development effort. In that respect, we will be developing hierarchical metrics. That is, we will develop metrics for components in the technosocial modeling system based on their contribution to our high-level user-centered metrics. It will be essential that we evaluate the components early to assess these metrics. For the system we are developing, components include technical models, social models, a knowledge encapsulation component, and a gaming environment that is the user interface for the system.

It should be noted that the system we are evaluating consists of a technosocial model and a knowledge encapsulation framework that will automatically locate relevant material that can be incorporated into the model. Users will access these components using a gaming interface.

The following sections outline our evaluation plans for the various components of our system and for the overall systems. We also outline our initial metrics but these will evolve as we work with end users and researchers.

Model Evaluation

Understanding of the technosocial model inputs and outcomes is less problematic to evaluate and can be done much earlier in the development cycle. We can obtain information on the model inputs and outputs from experts and then do some experiments to determine if users understand this. This would also include evaluating users' understanding of what they can adjust in the models to investigate other outcomes. This type of evaluation can be done at a formative stage in the research. Paper prototypes will suffice or even extremely rough electronic mock-ups.

Knowledge Encapsulation Evaluation

The utility of the knowledge encapsulation framework (KEF) must be evaluated from the point of view of two end users. First, we look at the utility to the modelers of the information located by KEF. We can instrument this component to count the documents that are read by the modelers and those that contain information that is incorporated into the models. Once we have this list of documents, we can have the analysts rate these documents or, more specifically, the information gleaned from the document, for its' contribution to the insights gleaned from the model.

Gaming Environment Evaluation

As the gaming environment is the user interface, usability issues must be addressed in order to properly assess utility. We will develop a heuristic review process to assess the usability of the gaming environment. This will be a standard usability review to ensure interactions are consistent and that other guidelines for serious games have been followed. This will include evaluation of the gaming environment for involvement, presence and flow (Takatalo et al. 2007).

There are also a number of utility measures to assess. First, are the roles provided to the analysts by the gaming environment appropriate and understandable? Are the variables they are allowed to manipulate during the game understandable and do these variables allow them to explore a good range of hypotheses? It is most likely that we will have to situate these evaluations in a number of different scenarios to ensure that the gaming environment will work in diverse analytic situations. This will involve classifying analytic tasks (Greitzer and Allwein 2005).

Summative System Evaluation

One possibility for evaluation concerns a comparison with current analytic gaming exercises. To gain a better understanding of a situation and possible reactions to different strategies, analysts devise exercises. Objectives

are determined and a number of experts are asked to participate. Depending on the question to be investigated and the expertise required, exercises can last anywhere from 4 hours to 4 days. Thus, the exercises are quite expensive to run. The participants are usually quite receptive to these exercises and feel that it is well worth the time. That is, their understanding of the analytic question is significantly increased using these exercises. Decision makers are not always as accepting. A particular complaint is the lack of sources that are cited. (Personal Communication 2008). By comparison, one might hypothesize that technosocial analytic systems would:

- necessitate fewer experts
- consume less time
- produce the same or better level of understanding
- produce the same or better insights
- help analysts produce a debrief of the same or better quality
- provide sources and simulation data for supporting recommendations in the debrief.

These are metrics that could be collected in a summative comparative evaluation to determine if any of the above hypotheses are supported.

Our Current Implementation of Evaluation Procedures

We are currently working with stakeholders to determine metrics that will resonate with them and help them in deciding if technosocial systems will provide utility in their analytic work. In working with the modeling researchers, we are developing materials for doing user-centered evaluations of the technical models alone, the social models alone, and the combined technosocial models. These evaluations will be expert reviews and will focus on whether the modelers have captured the essentials and whether the variables that the analysts can manipulate are appropriate. Based on a number of runs of the simulations, we will also get feedback as to whether the outcomes are appropriate given the inputs. When we evaluate the combined models, we will be working with two experts, one with expertise in the technical model and another with expertise in the social model.

One of the distinctive advantages of the technosocial approach to model integration and software development is that systems will take advantage of the best features drawn from each of these domains, integrated in sensible ways. In an evaluation effort, we must assess the impact of this blending for users (the analysts). This is a different problem from validating the integration, although the development team will need to do this validation to create a successful product. We will examine approaches to assess the differences between a purely technical approach versus a blended one. Other considerations to add to our user-centered evaluation include:

- What impact did the specific social features as blended with the technical features have on usability/utility?

- Did the social and technical model feature sets blend well? Were there unexpected direct or side effects that impacted software performance for the user?
- Were any technical model features removed or downgraded due to the social model integration? Conversely, were social model features impacted when integrated? How did these impact usability/utility?

We will evolve this list as we become more familiar with specific technosocial software development efforts.

When we have completed evaluating the models and the combined models, we will start evaluating the gaming environment that analysts will use to interact with the models and outcomes. We will conduct both usability and utility evaluations. If the usability heuristic review locates a number of usability problems, those must be fixed before we bring in end users for simple utility evaluations. This will include ensuring that the analyst can interact with the model and understand the inputs and the outcomes of the models. For all of these evaluations, it will be necessary to develop an overall scenario and a task for the analyst.

The KEF evaluations will be done in the same time frame. We can do the end user evaluations at the same time as we do the gaming evaluations. The evaluations based on the input from the modelers will be done as soon as the modeling efforts and the KEF efforts have been integrated.

The final step would be summative evaluation conducted near the end of the project. As discussed earlier, we plan to do a comparison to a current approach. The best measure would be to have a scenario analyzed by one group of analysts using a traditional approach and a second group using the technosocial predictive analysis approach. This would allow us to test our hypotheses that this system could produce the same or better quality debrief, the same or better level of understanding, and be done in less time by fewer experts. As a follow on to this evaluation, we also plan to distribute the products produced from both approaches to decision makers to review, both from the quality point of view and from their confidence in the recommendations made.

Conclusions

While the process outlined in this paper is extremely complex and will be expensive to carry out, readers should note that this user-centered evaluation work is in the research phase. We do not expect that future technosocial efforts will need to conduct the full set of evaluations. In our particular project, we will be applying evaluation efforts to two different models. In addition, we hope to be able to track the adoption and use of this technology. We plan to do some follow-up interviews after the fact to ascertain how helpful our evaluations and metrics were to researchers, end users, and decision makers. Information from these two efforts should allow us to identify the criteria that should be used to select the main areas and metrics that future evaluations should address.

References

- Dixon, D. and Reynolds, W. 2005. Visualizing the Political Landscape. The International Conference for Intelligence Analysis. Analysis, McLean, Va. <https://analysis.mitre.org/proceedings/index.html>. Accessed 12/2008.
- Greitzer, F. and Allwein, K. 2005. Metrics and Measures for Intelligence Analysis Task Difficulty. The International Conference for Intelligence Analysis. Analysis, McLean, Va. <https://analysis.mitre.org/proceedings/index.html>.
- Hanig, R. and Henshaw, M. 2008. A National Security Simulations Center. Studies in Intelligence, Journal of the American Intelligence Professional, Vol. 52 (2). <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol52no2/to-improve-analytical-insight.html>. Accessed 12/2008.
- Morse, E., Potts, M., and Scholtz, J. 2005. Methods and metrics for evaluating technologies for intelligence analysts. MacLean, VA.
- Moore, G. 2002. Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers. Harper Collins.
- Personal communication with Instructors of Analytical Gaming Classes in the Intelligence Community. September, 2008.
- Redish, J. 2007. Expanding Usability Testing to Include Complex Systems, Journal of Usability Studies, Vol. 2(3), pp 102-111. May.
- Scholtz, J. 2006. Methods for evaluating human information interaction systems. Interacting with Computers, Vol. 18 (4), July, pp 507-527.
- Takatalo, J., Häkkinen, J., Lehtonen, M., Komulainen, J., Kaistinen, J., Nyman, G. 2007. Presence, Involvement and Flow in Digital Gaming, CHI 2007 Workshop Evaluating User Experience in Games, April 28 – May 3. San Jose, CA.
- Willis, H., LaTourrette, Kelly, T., Hickey, S., Neill, S. 2007. Terrorism Risk Modeling for Intelligence Analysis and Infrastructure Protection. Rand Corporation.