Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

VALIDATING AGENT BASED SOCIAL SYSTEMS MODELS

Gnana K. Bharathy
Barry Silverman

Ackoff Collaboratory for Advancement of Systems Approaches (ACASA)
University of Pennsylvania, 120 Hayden Hall, 3340 South 33rd Street
Philadelphia, PA 19104-6316, USA

ABSTRACT

Validating social systems is not a trivial task. The paper outlines some of our past efforts in validating models of social systems with cognitively detailed agents. It also presents some of the challenges faced by us. A social system built primarily of cognitively detailed agents can provide multiple levels of correspondence, both at observable and abstract aggregated levels. Such a system can also pose several challenges including large feature spaces, issues in information elicitation with database, experts and news feeds, counterfactuals, fragmented theoretical base, and limited funding for validation. Our own approach to validity assessment is to consider the entire life cycle and assess the validity under four broad dimensions of methodological validity, internal validity, external validity and qualitative, causal and narrative validity. In the past, we have employed a triangulation of multiple validation techniques, including face validation as well as formal validation tests including correspondence testing.

1 INTRODUCTION AND PREVIOUS WORK

Models are frequently evaluated by their ability to estimate an observe phenomenon over a specified range. In terms of traditional modeling and simulation parlance, the processes involved are called verification and validation. Obtaining valid inputs and validating outputs are critical steps in any modeling and simulation endeavor (Axelrod 1996, Moss and Davidsson 2001).

As crucial as these steps are, there is neither single established definition nor a process for model validation. Non- statistical models, especially agent based and systems dynamic models have often been criticized for relying extensively on informal, subjective and qualitative validation procedures or no validation at all (Edmonds and Chattoe 2005, Hartley 1997). Many a researchers have made clarion calls for systematizing and improving validation (Leombruni et al. 2006, Fagiolo et al. 2006, Midgley et al. 2007).

In this paper, as practitioners of social systems modeling with cognitively detailed agents, we will provide a summary of verification and validation from our perspective.

Several macro or abstract level validation approaches have been proposed for social system models, including (Carley and Gasser 1999): (1) theoretical verification or internal validation by subject matter expert determining conceptual validity (Gluck et al. 2008); (2) external validation against real world comparing the results from the model to observations in the real world (Gluck et al. 2008); and (3) cross-model validation that compares different models (e.g., Axtell et al. 1996, Hales et al. 2003). Useful as these approaches are, these approaches have been employed at macro-level. While relevance and significance of macro-level phenomena is only of one of the reasons for staying at an abstract level.

Using the typology of Schreiber (2002) and Rom Harré (1970), most agent-based models are significantly abstracted, and do not produce the exact same outputs as their target. As Schreiber suggests that most agent-based models could be classified as paramorphic analogues, showing similarity in working, "but producing output that is similar or analogous, but they not exactly the same". Validation of such

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

models are carried out at an abstract level. Frequently, in these cases, validation is carried out by interpreting and telling a story from the patterns that are observed. In some cases, a macro-level correspondence test may also be carried out. Typically, if the story matches that of the real world phenomenon, then it is deemed validated. At such high levels of abstraction, it is really difficult to impose any more stringent conditions of validation than analogy. Gilbert (2004) on the other hand claimed that "to validate a model completely, it is necessary to confirm that both the macro-level relationships are as expected and the micro level behaviors are adequate representations of the actors' activity". However, this has neither been easy nor relevant for most models. In addition to high level of abstraction in typical cellular automata models, which would render micro-validation difficult, path dependencies and the stochastic nature of human behavior models (like other multi-agent models) render point-predictions impossible (Pahl-Wöstl, 1995). These, combined with scarcity of data, emergence, which is outside the specifications, and large parameter space render validation (and verification) exercises tenuous, at best.

The situation for cognitively detailed agent based models is not significantly different from that of simple agent model. For specialized, purely cognitive tasks and physically based applications (e.g., training on battlefield tactics, cockpits), greater validation has been achieved than that has been possible in the psychological based models. Not surprisingly, even in this decade, many an evaluation of synthetic agents have been based on the concept of ‘believability” (Gratch and Marsella, 2004, Pew and Mavor, 1998).

Alongside problems, there have also been solutions, or at least debate. Moss and Edmonds have advanced such concepts as micro-macro and cross validation with case studies (Moss and Edmonds 2005). For example, Moss and Edmonds (2005) point out that by relying entirely on statistical validation methods (e.g. econometrics), “many statistical models fail to validate their analysis by any means other than statistically even though other [non-statistical] means are available”. With advocacy for qualitative or subject evaluation, some researchers prefer to use the term “Evaluation of the Models” instead of validation (Schreiber, 2002).

The cognitively detailed human behavior models are relatively closer to homeomorph (Harré 1970) (although no such claim is made) than other abstract agent models, at least at the individual or societal level, which lends itself to the possibility of being validated in multiple dimensions/ aspects. Majority of the simple agent models only reach correspondence at very high level of abstraction. On the other hand, a cognitively detailed model can provide multiple levels of correspondence at the level of observable behaviors, measurable parameters. They could also be evaluated at higher levels of abstractions where aggregated and abstract states of the world can be compared.

Shannon (1987) suggests that since no model is absolutely correct in the sense of a one-to-one correspondence between itself and real life, especially the agent-based and human behavior varieties, one should not expect a black-and-white answer from modeling in general, and complex models in particular. Instead, the modeling should be treated as an iterative process of bringing about a (preferably) qualitative jump in the understanding. This seldom comes through answers the model gives, but, rather, through systematic participation in the exercise of modeling and the transparency it brings about, with stakeholders engaging in dialogue as the result of modeling or witnessing the model outcomes.

Forecasting is also closely related to V&V in the sense that a validated model is a prerequisite for forecasting. The most widely regarded research to date on prediction or forecasting in complex systems lies in the weather forecasting community. WCRP (2009) defines the term forecast to mean a “prediction of the future state (of the weather, stock market prices, or socio-political system)”, and forecast verification as “the process of assessing the quality of a forecast”.

In the following sections, we introduce some of the validation cases arising out of our social system model work. The purpose of our paper is not to demonstrate a generic validation methodology. Validation literature is extensive. For example, researchers such as Balci (1998), Petty (2009), MSCO (2006) discuss verification, validation and testing in general modeling and simulation. Balci outline 15 principles and 75 techniques in his 1998 paper that include a number of techniques that we employ. Instead, we hope to spur research in this direction by providing some cases of validation we had carried out in our social systems modeling work. In doing so, we also highlight some formal as well as informal techniques employed in our modeling and simulation.

We begin with a caveat, that in social systems models, the ultimate objective of the model should not be for prediction, but for exploration and learning. However, the model will be evaluated as though it would be employed in a predictive setting, for without predictions, one cannot really evaluate a model quantitatively. That having been said, a model must be validated in commensurate with the end uses. For instance, if the purpose of the model is to be used in entertainment such as standard game engine, one does not need to validate it extensively. On the other hand, if the purpose is to be used as exploratory test bed to design policies, extensive verification and validation are required. The case of "serious game" would be somewhere in-between on that continuum.

2 VALIDATION CASES

2.1 Description of the Model and some Challenges

We have constructed, using a framework named CountrySim/ FactionSim, a country based on multi-resolution agent based approach. This model has a virtual recreation of the significant agents (leaders, followers, and agency ministers), factions, institutions, and resource constraints affecting a given country and its instabilities.

These agents are cognitively deep and come equipped with values (goals, standards, preferences, cultural and ethical values, personality). The agents belong to factions, which have resources, hierarchies of leadership, followers. The factions that agents belong to, as well as the agents themselves, maintain dynamic relationships with each other. The relationships evolve, or get modified, based on the events that unfold, blames that are attributed etc. As in the real world, institutions in the virtual world provide public goods services, albeit imperfectly owing to being burdened with institutional corruption and discrimination. FactionSim is an environment that captures a globally recurring socio-cultural "game" that focuses upon inter-group competition for control of resources (e.g, Security/Economic/Political Tanks). This environment facilitates the codification of alternative theories of factional interaction and the evaluation of policy alternatives. It is a tool that allows conflict scenarios to be established in which the factional leader and follower agents all run autonomously and are free to employ their micro-decision making as the situation requires. A diagrammatic representation of an example CountrySim model is given in Figure 1.

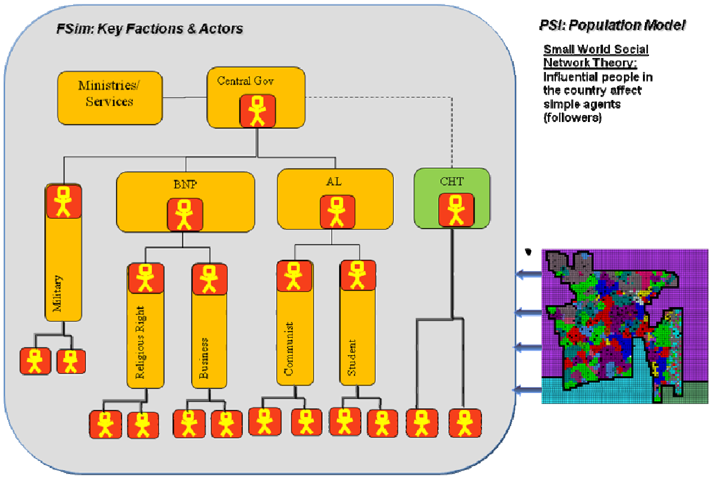
FSim: Key Factions & Actors
Ministries/ Services
Central Gov
BNP
AL
CHT
Military
Religious Right
Business
Communist
Student
PSI: Population Model
Small World Social Network Theory:
Influential people in the country affect simple agents (followers)

Figure 1: Overview of the Components of a CountrySim Model: Bangladesh (Courtesy: Silverman et.al. 2010)

For additional details, see Silverman et.al. (2010, 2007a, 2008a). For a given state being modeled, CountrySim uses FactionSim (and PMFserv) typically to profile 10s of significant ethno-political groups and a few dozen named leader agents, ministers, and follower archetypes. These cognitively detailed agents, factions, and institutions may be used alone or atop of another agent model that includes 10,000s of lightly detailed agents in population automata.

Social system models, by definition, are complex, with imprecise, incomplete and inconsistent theories (Silverman et.al. 2009a). In addition, these models have very large feature spaces (on the order of 1000), giving rise to "curse of dimensionality" (De Marchi 2005). Frequently, an argument is made in favor of reducing dimensionality by building simple, yet elegant toy models adhering to Keep It Simple, Stupid (KISS) principles. However, given the complexity of social systems, many researchers argue otherwise. For example, Edmonds & Moss (2004) argue that we need to move away from the concept of KISS to a new mantra – Keep It Descriptive, Stupid (KIDS). This does not mean the models have to be unnecessarily complex either. Complexity of the model must be driven by purpose. Both have merit under different circumstances. For descriptive modeling intended to provide learning, exploration and immersive training through social systems, "curse of dimensionality" is a given.

Multi-dimensional approach to validation combined with emphasis on out-of-sample comparisons (common knowledge) as well as qualitative insights, including extensive use of domain knowledge in the model construction process (De Marchi, 2005), increase confidence in such complex models.

2.2 Internal Validity

In order to ensure internal validity, we attempt to build our models rich in causal factors that can be examined to see what leads to particular outcomes. We also try to base our models on best currently available scientific theories of social systems and the other types of systems involved. This is because our goal is to see if models based on theory will advance our understanding of a social system, and if not, what is missing. Instead of relying on law of parsimony, we build data-rich, descriptive approaches. We seek approaches that drag in as much exoteric detail as possible about the actual stakeholders and personalities in the scenarios being studied – their issues, dilemmas, conflicts, beliefs, mis-perceptions, etc. We also synthesize domain knowledge from multiple subject matter experts in a broader domain where there is high level of fragmentation. At the level of components, we assess the completeness, clarity, coherence, and robustness (Component Validity Assessment). Such assessment of internal validity is often referred to as Verification. Verification determines whether the model, as a whole, adheres to the specifications. Verification is not only a pre-requisite for validation, but can also be construed as contributing to one dimension in the triangulation process employed to validate the system. By increasing the number of constraints on the system, verification strengthens the triangulation strategy. Specifically, the verification process attempts to re-create the very empirical evidence used for constructing the model. For this, we also carry out verification process hierarchically, at levels such as: internal model, individual agent (involving a single agent in a minimalist world), as well as that of a more realistic and whole scenario. The activity is carried out by establishing conformity of agent behavior against specifications/ expectations through systematic inspection.

In almost all models, we have employed "Degenerate Tests" by interrupting some components of the model and noting the impact on overall results, as well as using "Traces" testing to look at individual agents (time dimension) as they work through the modeling environment (Sargent 1987, Balci 1998). The degenerate tests were combined with extreme bound analysis, where we have started to determine whether the model continues to make sense at the boundary conditions or extreme values (Banks, 1996, Leamer 1985). The examples of these include running the model with and without different actors, institutions or population models or running the models with extreme values (very high value of risk aversion in population). Any suspicious [absurd or against common sense], low level behavior (showing high propensity for violence when population is deliberately and excessively risk averse) displayed in the extreme condition is used to identify any potential errors in the model. Extreme value analysis is appropriate for validating that there is internal consistency (i.e., the parameters are relatively sensible with respect to each other).

When carrying out the degenerate test, we treated the models as white boxes, and worked on actual parameters. These steps are in the right direction.

Another issue related to internal validity is assessing the Ontological Adequacy. It is important to determine whether the combined set of theories and models implemented work well together and what gaps need to be filled in. However, owing to gaps in social science theories, qualitative nature of theories and large feature spaces, it is difficult to assess the ontological validity of model collections. This is a sober reality, but systematic validation process inches in continuous improvement in the right direction. Further research is needed to fill these gaps adequately.

2.3 Methodological Validity

For us, model should be viewed as a System with entire cycle of model conception, data collection, model building, testing, verification, validation, exploration, as well as learning and continuous evolution of models. Accordingly, model evaluation is a gradual, systematic and iterative process of continuous evolution.

In assessing methodological validity, it is desirable to assess two key aspects, namely modeling process adequacy and software process adequacy. Software Process Adequacy is more sober part of the assessment, as software development verification and validation are well established practices, and we will not be concerned with this in this paper.

The key questions in the assessment of Modeling Process Adequacy are: Is there a systematic process for model construction? Is the process by which model is constructed defensible? Are steps being taken at the process level to control errors, cognitive biases, especially confirmation bias?

In recent years, modeling methodologies have been developed that help to construct models, integrate heterogeneous models, elicit knowledge from diverse sources, and also test, verify, and validate models. The details of the process are beyond the scope of this paper, but can be found elsewhere (see for example, Silverman and Bharathy 2005; and Silverman, Bharathy and Kim 2009). We recap the salient features briefly here.

These models are knowledge based systems, and to a significant extent the modeling activity involves eliciting knowledge from subject matter experts as well as extracting knowledge from other sources such as data bases and event data, consolidating the information to build a model of the social system. We have employed web newsfeeds, country databases, and SME interviewing.

We assembled an integrated index of all the fields available from 45 country and social science databases (eg., CIA Factbook, World Values Survey, Global Barometer Survey, etc.). A number of parameters pertinent to our model (e.g. population level and economic parameters) are available in the databases.

Given that the survey was not custom-designed for our model, we have to select proxy measures for our parameters of interest. For example, we also have to carry out some manipulation of data, as the units of analyses employed by the surveys (frequently national and individual levels) and our model (sub-state units such as factions and individual levels) differ (Silverman, Bharathy and Kim 2009)..

Likewise, web newsfeeds provide ample supplementary material on the events of interest in the target countries, however, there are no automated extraction methods yet available to parse this corpus into the sophisticated type of parameters we need for our multi-resolution cognitive and social layer models. Coverage is also a concern with the databases as well as with the newsfeeds.

Knowing the limitations of the two previously mentioned means of extracting information—namely, country databases and automated data extraction tools—in the short term at least, we might in fact be better off by gathering information directly from the best available country experts, tapping their expertise by means of a survey questionnaire to them, supplemented by conducting open-ended interviews. We largely use databases and web news feeds to for background information and sanity checking what our SME survey produces.

There are three main difficulties associated with using SMEs to elicit the information we need. First, eliciting information from SME incurs significant financial and human resources. Second, SMEs, by virtue of being human is fallible, may sometimes provide us with biased and, occasionally, even blatantly incorrect information: e.g., see Tetlock (2005). Third, SMEs are also difficult to find. Therefore, we verify and triangulate at least a sample of SME information against other sources and other SMEs.

We have authored and assembled a survey that is self-explanatory and has validated questions about each parameter needed in a socio-cognitive agent model. This has been employed for eliciting knowledge from country or leader desk experts. The input data, obtained from multiple sources, tend to be incomplete, inconsistent and noisy. In the course of constructing these models, there is the risk of contamination by cognitive biases and human error. Therefore, a process is required to integrate and bring all the information together. We employ a process centered around differential diagnosis. This design is also based on the fact that directly usable numerical data are limited and one has to work with qualitative, empirical materials.

The burden of this integrative modeling process is to systematically transform empirical evidence, tacit knowledge and expert knowledge from diverse sources into data for modeling; to reduce, if not eliminate, the human errors and cognitive biases (for example, for confirming evidence); to ensure that the uncertainties in the input parameters are addressed; and to verify and validate the model as a whole, and the knowledge base in particular. For lack of a better term, the process has been conveniently referred to as a Knowledge Engineering (KE) process due to extensive involvement of KE techniques and construction of the knowledge models. Such systematic approach increase confidence in the models, even though these are not deemed traditional validation exercises.

2.4 External Validity

A social system built primarily of cognitively detailed agents (such as PMF Serv based StateSim) can provide multiple levels of correspondence. At observable levels, the model might have correspondence in behaviors (decisions agents make) and measurable parameters (e.g. GDP, public goods service levels received). They could also be evaluated at higher levels of abstractions where aggregated and abstract states of the world (developmental metrics, conflict metrics such as rebellion, insurgency) can be compared. During the validation exercise, the model would attempt to create scenarios constructed from a fresh set of empirical evidence hitherto unused in model construction. Several historical correspondence tests indicate that PMFserv mimics decisions of the real actors/population with a correlation of approximately 70-90% (see Silverman et al. 2009b, 2007b, 2010).

2.4.1 Macro-Level and Cross Model Validity

The ensuing section summarizes and illustrates macro-level validation with one of our models (additional details can be found in Silverman et.al., 2010). The direct (or default or base) outputs from the CountrySim model include decisions by agents, levels of emotions, resources, relationships between factions, membership of agents in different factions etc. These parameters are tracked over time and recorded in the database. Since our intention is to model instability in selected countries, we defined aggregate metrics or summary outputs of instability from default model outputs.

These aggregate metrics (summary outputs) are called Events of Interests (EOIs). EOIs reveal a high-level snapshot of the state of the conflict. Specifically, as required by the sponsors, CountrySim generates several EOI scores important to instability including: insurgency, rebellion, domestic political crisis, and inter-group violence.

In order to assess Analytical Adequacy, we ask whether the collection of models assembled and implemented thus far satisfy various types of correspondence tests and historic recreation tests. This will often entail backcasts on a set of historic test data with-held during model training and tuning. And to avoid the problem of over-fitting to a single test sample, we always need to examine if the models work across samples. Here we applied them to models of several States (namely Bangladesh, Sri Lanka, Thailand and Vietnam) and Groups, People and across different types of metrics of interest (different EOIs). The following sample results were drawn from one of our previous paper (Silverman et.al., 2010) and illustrates one EOI for one Country.

The EOI Framework has its theoretical basis in a premise that conflict can be measured through a composition of indicators, which include both behavioral and structural or institutional factors (Covey, Dziedzic and Hawley 2005, Baker, 2003). For example, a framework that was developed when we developed ours is MPICE Framework (Dziedzic, Sotirin, and Agoglia, 2008), which has a framework for measuring progress in conflict environments.

Our EOIs such as Rebellion (separatist conflict), Insurgency (Coup and challenge to power), Domestic Political Crisis (opposition to the government, but not to the level of rebellion or insurgency), Inter-Group Violence (violence between ethnic or religious groups that are not specifically directed against the government), and State Repression (use of government power to suppress sources of domestic opposition) were measuring the level of conflict (Silverman et.al., 2010).

Our EOI framework identifies and organizes a set of indicators hierarchically under a given EOI with weights on the arcs of the tree and the indicators on the nodes. These weights represent the importance of different indicators for a given EOI. During the training period, using the weights on the arcs of the tree, the occurrence of EOIs in the simulated world can be tuned against the occurrence of EOI in the real world. Specifically, the weights are then employed to make out-of-sample predictions in the test period. The weights tend to be invariant across similar countries.

Having constructed high level aggregate EOIs, we compared them to Ground Truths of EOIs coded from real data by subject matter experts. Having tested and verified the model over the period of 1998-2003, we ran the model for subsequent 3 years (of 2004 through 2006) and made predictions. The predictions were benchmarked against the Ground Truth consisting of real world EOI for the same interval. Although we generate multiple futures (from multiple runs), in metrics and calculations, we only employ the mean values across alternative histories for validation. We cast mean likelihood estimates from multiple runs into a binary prediction by employing threshold systems.

Definitions:

- Accuracy = (TP+TN)/(P+N)
- Precision = TP/(TP+FP)
- Recall/TP rate = TP/P
- False Positive (FP) Rate = FP/N

Where T: True, F: False, P: Total Positives and N: Total Negatives, TP: True Positives, TN: True Negative, FP: False Positives and FN: False Negatives. ROC curve shown on the left describe the relationship between Recall and FP Rate as FP Rate is varied.

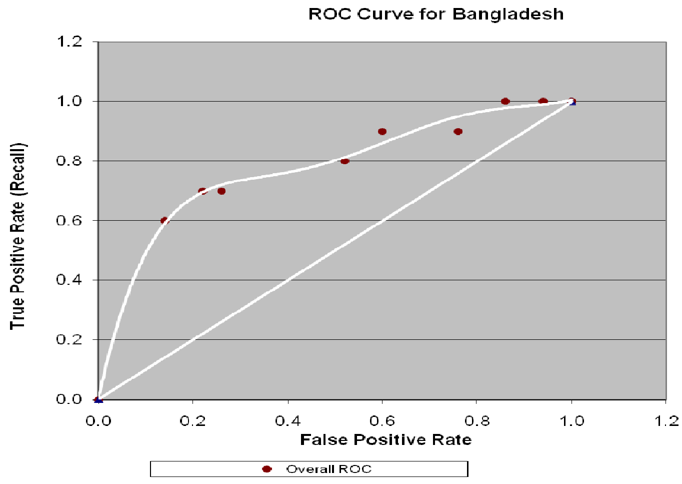

Figure 2: Relative Operating Characteristic (ROC) Curve for CountrySim (Bangladesh) (Courtesy: Our paper - Silverman et.al. 2010)

Table 1: Confusion Matrix

		True Class	
		Positive	Negative
Hypothesized Class	Yes	TP	FP
	No	FN	TN

Table 2: EOI Summary Metrics

Metric for Bangladesh	Accuracy	Precision	Recall
Mean– with two thresholds at 0.65-0.35	87%	66%	81%

In order to get a quantitative relationship between CountrySim and Ground Truth forecasts, we make use of a Relative Operating Characteristic (ROC) curve (example shown in Figure 2). The ROC plots the relationship between the true positive rate (sensitivity or recall) on the vertical and the false positive rate (1-specificity) on the horizontal. Any predictive instrument that performs along the diagonal is no better than chance or coin flipping. The ideal predictive instrument sits along the y-axis.

This curve well above the diagonal shows that CountrySim largely agrees with the Ground Truth. In fact its accuracy measured relative to Ground Truth is 80+%, while its precision and recall were listed in the Table 2. While these would be less than luster results for any physical system, for agent-based models of bottom up social science processes, these are useful results. They are useful both since they significantly beat coin tossing and since these type of models also afford the analyst ways to drill down to try and explore casual factors as we will explain forthcoming subsections. In this case, we employed backcasts with set of data independent of model construction, but eventually, one should move to forecasts and to tracking the actual outcomes to verify the forecast quality.

It is worthwhile mentioning that accuracy does not distinguish between the types of errors it makes (False Positive versus False Negatives). On the other hand, precision and recall do not stand alone and require to be combined with accuracy. Generally speaking, ROC curve is a comprehensive measure. Yet, there are times when ROC Analysis and Precision could yield contradictory results. The implication of all these is that one must understand the data and the domain it pertains to before carrying out analysis. That is, there is no substitute for qualitative domain knowledge.

2.4.2 Micro-Level Validity

In carrying out a micro-level validation process, we primarily aim to create correspondence at the level of agent decisions or other lower level parameters such institutional parameters, socio-economic indicators etc. The intention is to calibrate the model with some training data, and then see if it recreates a test set (actually validation).

In the following case, we describe correspondence in leader decision, although the same thing could be done for follower decisions. We coded and classified the leader actions in the real and simulated worlds same categories (bins). For simplicity, we describe a 3 bin classification, namely positive, neutral and negative. We started with visual correspondences to give an intuitive or face validation, but they neither prove that two distributions are the same, nor give any richer picture.

Specifically, in the test dataset, the real world leader of certain country made 52 decisions affecting the population and that we sorted into positive, neutral, and negative actions. In the simulated world, the same leader made 56 action decisions in this same interval. At this level of classification (positive, neutral, negative), we were able to calculate a mutual information or mutual entropy (M) statistic between the real and simulated base cases. M ranges from 0 to 1.0, with the latter indicating no correlation between two event sets X and Y. M can be expressed by (common knowledge): $M\ (X{:}\ Y) = H(X) – H(X|Y)$ where X and Y are the simulation and historic sources, respectively, and H (.) is the entropy function, defined by: $H(X) = - \Sigma\ p(x)_i \log p(x)$. Applying this metric, the mutual entropy values were found to be less than

0.05 (at least an order of magnitude smaller than the mutual entropy of 1.0), indicating reasonable degree of correlation between real and simulated data. With an M metric, one cannot make statements about the confidence interval of the correlation, however, the Leader in this model seems faithful to his real world counterpart. Additional details may be found in Silverman et.al. (2007b, 2009b).

2.4.3 Qualitative, Causal and Narrative Validity

Any large model of models system will have issues of hyper-confluence, autoregression, grey box impenetrability, and related concerns that cannot be captured by statistics alone. For example, in predicting insurgency (*coup d'état*), another EOI from our model of Bangladesh, our model does not get the timing of the insurgency right; however, the leading indicators for a potential insurgency is well captured.

Some useful techniques for resolving the ambiguity and adding richness to model are tracing the results back to their origins, developing a story from the model outputs and obtaining qualitative feedback from subject matter experts. These are work in progress, more than other aspects of the model. Tracing causal networks manually in a complex system can be difficult, if not impossible, to be carried out manually on a regular basis. If the model library permits the installation of intermediate data capture, drilldown and traceback instruments, an automated system can be created to help in determining the validity of suspicious results. An embedded results-capture-drilldown-traceback system can be important in developmental and periodic testing and may be critical in triggered testing and model use evaluation. An important aspect is our approach is aimed at bringing about end-to-end transparency and drill-down capability – from the front end model elicitation (web interview, database scraping) to the backend EOI views and drill down through indicators to events and even to the ability to query the agents involved in the events. Demonstrating such end to end transparency and narrative validity would go beyond the scope (and page limit) of this paper. One of our crude attempts was creation of preliminary conversational agents. One may interrogate them about the theory behind models or parameter settings of any of their values, (goals, standards, preferences, personality), resources, social relations, group dynamics, decision making history, opinions about other agents and the actions they have done, opinions about institutions, how they feel about grievances and transgressors, etc. Visual descriptions (e.g. movie clips and Tech Reports) of some of these previous applications are available at: <http://www.seas.upenn.edu/~barryg/HBMR.html>

Typical human face validation exercises, are either very descriptive and allow limited statistical treatment or go through checklists and offer limited description. However, it is possible to design subject matter expert involvement in validation to be both holistic and scientific. For holistic treatment, outputs of the model can be weaved into a narrative by the modeler, and then evaluated by subject matter experts. In order to make the test rigorous and counter the human confirmation bias (for descriptions of biases, see (Kahneman and Tversky 1984), experts can attempt to reject, rather than support the stories or hypotheses generated by the model.

A modified Turing's test that we propose here could help bring in the strengths of face validation while also offering the possibility of using human judgment and rich description. The problem of determining whether the outputs of the human behavior model are sufficiently representative of reality is analogous to that of determining whether computer behavior sufficiently resembles human cognition. In our simplest version of the Turing's test, a statistically designed experiment would be conducted wherein a group of experts would be asked to tell the difference between the trace output generated from the model and that generated by the actual occurrences in the historical events. The experts could weigh in both qualitative and quantitative information that they observe and can give a combination of both descriptive assessment and numerical ratings.

In both cases, there are no universally accepted criteria for assessing similarity, and only limited data available. We would also want the experts to discuss the validity of their estimates. It also allows for experts using subtle clues to separate real versus simulation. As such, in a well validated model, an expert will be not be able to distinguish the sequences of moves and outcomes of the trace from those of the real ones. Another important aspect of achieving qualitative validation is extending the knowledge elicitation sessions beyond subject matter experts, and include all key stakeholders. Value of including stakeholders has been demonstrated by Companion Modeling (Barreteau et al. 2003).

3 CONCLUSIONS

In this paper, we summarized some validation dimensions and techniques that we had employed in the past, as well as some key issues in validation of social system models, which are predominantly occupied by cognitively rich agents. We take a life cycle based approach model validation. We have assessed model construction methodology, internal (and ontological) validity and external and qualitative validity. Methodological validity addresses the issue of obtaining input data while external validity test output of the model against independent set of data. The internal validity deals with theoretical and ontological adequacy as well as adherence to specifications. Qualitative, causal and narrative validity are deemed important to capture the richness of the social systems.

We briefly described our model construction process: Our primary inputs come from SME, for this we have designed extensive web questionnaire. We largely use these databases and web news feeds to for background information and sanity checking what our SME survey produces. We integrate multiple inputs through differential diagnosis and hypothesis testing. As knowledge based systems, eliciting knowledge from subject matter experts as well as extracting knowledge from other sources such as data bases and event data, consolidating the information to build a model of the social system, are important. The existing country databases, event data from news feeds and subject matter experts are great assets for those of us in the Modeling and Simulation (M&S) community who are committed to using realistic agent types to populate our simulated world. However, as noted previously, using these sources at this stage of their development requires efforts to take into account their strengths, weaknesses, terminology, and idiosyncrasies. By employing a combination of sources through a triangulation process (e.g. a Bayesian based differential diagnosis), we were able to increase the confidence in our model outputs.

Even though we employ statistical validation, they are not as rigorous as formal validation techniques employed by statistical models. The validation of likes of social system models are, however, carried out in multiple dimensions, conceptual, temporal and spatial. These models can tell a story and provide explanation. These are their strength and complementarities to statistical models. Therefore, one should not expect a black-and-white answer from validating models in general, and complex models in particular. We subscribe to the view that no model can faithfully represent the reality, but detailed, mechanism based models are useful in learning about the system and bringing about a qualitative jump in understanding of the system it tries to model. We advocate that best use of models, especially complex social systems models is exploration rather than prediction. However, we also advocate validating the models well. The verification and validation processes can be either qualitative or quantitative, but for the activity to be worthwhile, it should inform us about the nature of errors and how one could control them.

As a social system built primarily of cognitively detailed agents, our model is amenable to providing multiple levels of correspondence (micro, macro). At observable (micro) levels, we showed correspondence in behaviors (e.g. decisions agents make). The same could be extended to other measurable, observable parameters such as GDP, services provided etc. At higher levels of abstractions, aggregated and abstract states of the world (in this case, conflict metrics such as rebellion) were compared. Included in the validity are equal parts about the data used and the generative mechanisms inside the agents. Both of these are finally more important than whether any particular predictions turn out to be accurate.

In general, adequate and multi-dimensional validation is an expensive, but necessary, proposition for a complex social system model. For example, use of domain knowledge helps reduce the dimensionality curse by structuring limited and available data; extending knowledge elicitation beyond subject matter experts to include all key stakeholders provides further insight into the domain; external out-of-sample validation provides statistical confidence; internal validity ensures adherence of structure and functions to specifications; end-to-end transparency enables drilling down to see the broader narrative not just as a validation exercise, but also as a learning opportunity; and arguably most importantly, synthesizing models and identifying gaps within and between models and commissioning social science research studies to bridge such gaps form a loop of iterative and continuous improvement, thereby furthering the theoretical foundations of evolving field of social system (Silverman 2010).

All these come at a price (literally), however. For example, dialogue between modeler, expert and stakeholders can prove to be expensive. Companion modelers that we mentioned earlier were only able to employ this methods for selected parameters or aspects of the model. We ourselves are limited by using fewer subject matter experts than we would like to. Rarely, new social science research is commissioned to fill-in the gaps. These reinforce the point that commitment and support of the policy makers and sponsors are almost always essential for carrying out not only validation, but also advance the science. The need for more concerted research, to quote Bob Dylan, "is blowin' in the wind"!

ACKNOWLEDGMENTS

This research was supported by Becks Fellowship, various government agencies, including an advanced research branch of DoD, a COCOM, and the US Army. The authors alone are responsible for any opinions, unsubstantiated claims, or errors made in this manuscript.

REFERENCES

Axtell, R., Axelrod, R., Epstein J., and Cohen, M. D. 1996. "Aligning Simulation Models: A Case Study and Results," Computational and Mathematical Organization Theory, Vol. 1, No. 1, pp. 123-141.

Baker, P. H. 2003. "Conflict Resolution: A Methodology for Assessing Internal Collapse and Recovery," by in Armed Conflict in Africa, Carolyn Pumphrey and Rye Schwartz-Barcott, eds. (Triangle Institute for Strategic Studies, Lanham, MD and Oxford: The Scarecrow Press, 2003).

Balci, O. 1998. "Verification, Validation, and Testing", In The Handbook of Simulation, J. Banks, Editor, John Wiley & Sons, New York, NY, August, Chapter 10, pp. 335-393.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2000. Discrete-event system simulation. 3rd ed. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Barreteau, O., and others 2003. Our companion modelling approach, Journal of Artificial Societies and Social Simulation, 6(1), <http://jasss.soc.surrey.ac.uk/6/2/1.html> [accessed 2006/07/30]

Carley, K. M. and Gasser, L. 1999. "Computational and Organization Theory," in Weiss, G. (Ed.), Multiagent Systems - Modern Approach to Distributed Artificial Intelligence, MIT Press, pp. 299-330.

Covey, Jock, Michael Dziedzic, and Leonard Hawley, Editors, The Quest for Viable Peace: International Intervention and Strategies for Conflict Transformation, United States Institute of Peace, Washington DC, 2005; http://bookstore.usip.org/books/BookDetail.aspx?productID=120589.

De Marchi, S. 2005. Computational and Mathematical Modeling in the Social Sciences, Cambridge

Dziedzic, M., Sotirin, B., and J. Agoglia, Editors, Measuring Progress in conflict Environments (MPICE) – A Metrics Framework for Assessing Conflict Transformation and Stabilization," Defense Technical Information Catalog, Aug 2008; or United States Institute of Peace.

Edmonds, B. and S.J. Moss 2004. From Kiss to KIDS - an 'antisimplistic' modelling approach; in: P. Davidsson et al. (eds.): Multi Agent Based Simulation; Springer, Lecture Notes in Artificial Intelligence, 3415: p.130-144. http://bruce.edmonds.name/kiss2kids/kiss2kids.pdf, (accessed 2006/07/30).

Edmonds, B. and Chattoe, E. 2005. When Simple Measures Fail: Characterising Social Networks Using Simulation. Social Network Analysis: Advances and Empirical Applications Forum, Oxford.

Fagiolo, G., Windrum, P., and Moneta, A. 2006. Empirical validation of agent based models: a critical survey, LEM Working Paper 2006/14, Sant'Anna School of Advanced Studies, Pisa, Italy, May. http://www.lem.sssup.it/WPLem/files/2006-14.pdf (accessed 2008/06/15)

Gilbert, N. 2004. "Open problems in using agent-based models in industrial and labor dynamics," In R. Leombruni and M. Richiardi (Eds.), Industry and Labor Dynamics: the agent-based computational approach, World Scientific, pp. 401-405.

Gluck, K., Bello, P. and Busemeyer, J. 2008. Introduction to the Special Issue [on model comparison]. Cognitive Science 32(8):1245-1424.

Goutte, C. 1997. Note on free lunches and cross-validation, Neural Computation, 9, 1211-1215 ftp://eivind.imm.dtu.dk/dist/1997/goutte.nflcv.ps.gz (accessed 2009/04/20)

Hartley, D.S. 1997. Verification and Validation in Military Simulation. In *Proceedings of the 1997 Winter Simulation Conference,* eds. S.G. Henderson, B. Biller, M. Hsieh, J. Shortle, J. D. Tew, R. R. Barton, . Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hales, D., Rouchier J., and Edmonds, B. 2003. "Model-to-Model Analysis," J. of Artificial Societies and Social Simulation, Vol. 6, No. 4, 5 http://jasss.soc.surrey.ac.uk/6/4/5.html (accessed 2009/06/22)

Harré, Rom 1970. The Principles of Scientific Thinking. London: Macmillan.

Kahneman, D. and Tversky, A. 1984. Choices, Values and Frames, American Psychologist, Vol. 39, No. 4: 341-50.

Leombruni, R., Richiardi, M., Saam, N.J., and Sonnessa, M. 2006. A common protocol for agent-based social simulation, Journal of Artificial Societies and Social Simulation, 9(1). http://jasss.soc.surrey.ac.uk/9/1/15.html (Accessed 2009/09/22)

Midgley, D.F., Marks, R.E., and Kunchamwar, D. 2007. The building and assurance of agent-based models: an example and challenge to the field, Journal of Business Research, 60: 884–893. At: http://www.agsm.edu.au/~bobm/papers/Midgley-Marks-Kunchamwar.pdf (Accessed 2010/03/08)

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecasting, 8, 281-293

Modeling and Simulation Coordination Office. Verification, Validation, and Accreditation (VV&A) Recommended Practice Guide. September 15, 2006. At: http://vva.msco.mil. (Accessed June 10. 2010).

Moss, S. and Davidsson, P. 2001, Multi-Agent-Based Simulation, Lecture Notes in Artificial Intelligence, Vol. 1979, Springer-Verlag.

Moss, S. and B. Edmonds 2005. Sociology and Simulation: Statistical and Qualitative Cross-Validation. Manchester, American Journal of Sociology 110(4): 1095-1131.

Pahl-Wöstl, C. 1995. The dynamic natural of ecosystems: Chaos and order entwined. Chichester: Wiley.

Schreiber, D. 2002. Validating agent-based models: From metaphysics to applications. Annual Conference of the Midwestern Political Science Association, Chicago, IL: April 2002.

Silverman, B.G., 2010. Systems social science: a design inquiry approach for stabilization and reconstruction of social systems, J. of Intelligent Decision Technologies. 4 (1): 51-74

Silverman, B.G. Bharathy, G.K., and G. J. Kim 2009a, "The New Frontier of Agent-Based Modeling and Simulation of Social Systems with Country Databases, Newsfeeds, and Expert Surveys " in Agents, Simulation and Applications, A. Uhrmacher and D. Weyns, (eds.) Taylor and Francis.

Silverman, B., Bharathy, G., and B. Nye 2009b, " Gaming and Simulating Sub-National Conflicts ", In Computational Methods for Counter-Terrorism, Shlomo Argamon and Newton Howard (ed.) Springer-Verlag (Berlin, Hiedelberg and New York).

Silverman, BG., Bharathy, GK, Nye, B., Kim, G. J., Roddy, Poe, M. 2010. "Simulating State and Sub-State Actors with CountrySim: Synthesizing Theories across the Social Sciences", In Sokolowski, J. and Banks. C. (ed.): Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains, M. Wiley STM

Silverman, B.G., Bharathy, G.K., Eidelson, R. and B. Nye 2007a. "Modeling Factions for 'Effects Based Operations': Part I –Leaders and Followers ", Journal of Computational and Mathematical Organization Theory. 13:379-406.

Silverman, B.G., Bharathy, G.K., Nye B. and T. Smith 2008a. "Modeling Factions for 'Effects Based Operations': Part II – Behavioral Game Theory ", Journal of Computational and Mathematical Organization Theory, 14(2):120-155, 2008.

Silverman, B.G., Bharathy, G.K., Smith, T. and Eidelson, R., and M. Johns 2007b. " Socio-Cultural Games for Training and Analysis: The Evolution of Dangerous Ideas ", IEEE Systems, Man and Cybernetics. Nov 2007. 37:6: 1113-1130, 2007.

Silverman, B. G., and G. K. Bharathy. 2005. "Modeling the Personality & Cognition of Leaders," in 14th Conference on Behavioral Representations In Modeling and Simulation, SISO, May.

Tetlock, P. 2005. Expert Political Judgment: How Good is It? How Can We Know?, Princeton, NJ: Princeton University Press.

WCRP 2009. Forecast Verification - Issues, Methods and FAQ, World Climate Research Programme, World Meteorological Organization. WWRP/WGNE, WG on Verification (accessed 2010/06/22), http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

AUTHOR BIOGRAPHIES

GNANA K BHARATHY is a researcher and project manager at the University of Pennsylvania (Penn). His areas of research include broadly risk management, analytics, and modeling and simulation, particularly of social systems. E-mail : <bharathy@seas.upenn.edu>.

BARRY G SILVERMAN is a Professor of Systems Sciences and Engineering and Director of Ackoff Collaboratory for Advancement of Systems Approach (ACASA) at the University of Pennsylvania. Among other honors, he has pioneered the synthesis of best of breed models to construct and study social system model. E-mail : <basil@seas.upenn.edu>.