Accepted and In-Print: Journal of Simulation (2012)

HOLISTICALLY EVALUATING AGENT BASED SOCIAL SYSTEMS MODELS: A Case Study

Gnana K. Bharathy* and Barry Silverman
Ackoff Collaboratory for Advancement of Systems Approach (ACASA),
University of Pennsylvania, 120 Hayden Hall, 3340 South 33rd Street,
University of Pennsylvania, Philadelphia, PA 19104-6316.
Emails: bharathy@seas.upenn.edu | basil@seas.upenn.edu | Ph: +610-400-8586
* Corresponding Author ||| Received: 04-Sep-2010 | Revisions: 16-Mar-2011, 17-Nov-2011 | Accepted: 13-Jan-2012

ABSTRACT

The philosophical perspectives on model evaluation can be broadly classified into reductionist/ logical positivist and relativist/holistic. In this paper, we outline some of our past efforts in, and challenges faced during, evaluating models of social systems with cognitively detailed agents. Owing to richness in the model, we argue that the holistic approach and consequent continuous improvement are essential to evaluating complex social system models such as these.

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. We subscribe to the view that no model can faithfully represent reality, but detailed, descriptive models are useful in learning about the system and bringing about a qualitative jump in understanding of the system it attempts to model -- provided they are properly validated.

Our own approach to model evaluation is to consider the entire life cycle and assess the validity under two broad dimensions of (1) internally focused validity/ quality achieved through structural, methodological, and ontological evaluations; and (2) external validity consisting of micro validity, macro validity, and qualitative, causal and narrative validity. In this paper, we also elaborate on selected validation techniques that we have employed in the past. We recommend a triangulation of multiple validation techniques, including methodological soundness, qualitative validation techniques such as face validation by experts and narrative validation, and formal validation tests including correspondence testing.

1 INTRODUCTION AND PREVIOUS WORK

The National Research Council (NRC)'s commissioned study on behavioral modeling and simulation [1] summarized the key concerns of the social system modeling as: *(1) Modeling strategy—matching the problem to the real world; (2) Verification, validation, and accreditation; (3) Modeling tactics—designing the internal structure of a model; (4)* [understanding the] *Differences between modeling physical phenomena and human behavior; (5) Combining components and federating models*. According to NRC, pervasive throughout all these threads are the lack of appreciation that social systems are different from, and more complex than, physical systems [and unrealistic expectation based on dealing with or modeling physical systems].

In this paper, as practitioners of social systems modeling with cognitively detailed agents, we will provide a summary of how we address each of these five issues with particular emphasis on model evaluation, especially validation.

Models are frequently evaluated by their ability to estimate an observed phenomenon over a specified range. In terms of traditional modeling and simulation parlance, the process involved is called validation. Obtaining valid inputs and validating outputs are critical steps in any modeling and simulation endeavor [2] [3].

As crucial as these steps are, there is neither a single established definition nor a process for model evaluation. 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 [4] [5]. (For example, the domains that these models are frequently employed in are plagued by scarcity of consolidated data sources). Many a researchers have made clarion calls for systematizing and improving validation [6] [7] [8].

Several macro or abstract level validation approaches have been proposed for social system models, including [9]: (1) theoretical verification or internal validation by subject matter expert determining conceptual validity [10]; (2) external validation against real world comparing the results from the model to observations in the real world [10]; and (3) model docking cross-model validation that compares different models (e.g. [2] [11]).

Useful as these approaches are, these approaches have been employed at the macro-level, ignoring the micro-level details. While relevance and significance of macro-level phenomena may justify staying at an abstract level under some circumstances, such an approach would not contribute to exploration and understanding of the virtual world.

Using the typology of Schreiber [12] and Harré [13], most agent-based models are significantly abstracted, and do not produce the exact same outputs as their target. As Schreiber suggests most agent-based models could be classified as paramorphic analogues, “producing output that is similar or analogous, but not exactly the same”. Validation of such 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. For example, Fagiolo et.al. [7] report that statistical correspondence test often involves attempting to tell a story relating “stylized facts drawn from empirical research” and model output. 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 (subjective) analogy.

Gilbert, in his recent book [14], largely conforms to this view, generally expressing distrust of empirical techniques. Gilbert’s choice of validation criteria vary depending on the level of abstraction of the model. He recommends focusing on: (1) replicating macro configurations that the modeler wants to explain (in the case of “abstract models” which are primary concerns of Gilbert’s thesis); (2) seeking qualitative resemblance (for "middle-range models"); or (3) comparing with empirical data (in the case of a "facsimile models", here again with considerable skepticism). Gilbert also recommends demonstrating robustness against sensitivity analysis. However, according to Gilbert [15] (and Troitzsch et.al. [16]), "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".

Validating macro relationships and micro level behaviors has not been particularly easy. In addition to high level of abstraction in typical cellular automata models, which would render micro-validation difficult (and less relevant), path dependencies and the stochastic nature of human behavior models (like other multi-agent models) render point-predictions impossible [17]. These, combined with scarcity of data, emergence, which is outside the specifications, and large parameter space render validation exercises tenuous, at best. Yet, replication of models is an important aspect of validation [18].

The situation for cognitively detailed agent based models is not significantly different from that of simple agent models. For specialized, purely cognitive tasks and physically based applications (e.g., training on battlefield tactics, cockpits), greater validation has been achieved. Not surprisingly, even in this decade, many an evaluation of synthetic agents have been based on the concept of ‘believability” [19] [20].

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 [21]. For example, Moss and Edmonds [21] 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”.

The cognitively detailed human behavior models are relatively closer to homeomorph [13] (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 evaluated 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 through measurable parameters at the level of observable behaviors. They could also be evaluated at higher levels of abstractions where aggregated and abstract states of the world can be compared.

Shannon [22] 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.

In the following sections, we introduce some of the evaluation cases arising out of our social system model work. The purpose of our paper is not to demonstrate a generic evaluation methodology, but instead add a case study to the literature. Readers looking for generic methodologies might consult researchers such as Balci [23], Petty [24], MSCO [25]. Readers interested in how diverse methodologies work in a specific case should read on.

2 DESCRIPTION OF THE MODEL AND SOME CHALLENGES

2.1 Model Descriptiveness

We have adopted a descriptive approach to modeling. Before we introduce our model, a few words must be said about the level of model descriptiveness.

Frequently, an argument is made in favor of building simple, yet elegant toy models, thereby keeping the dimensionality very low. This principle is often referred to as “Keep It Simple, Stupid” (KISS) [26]. This principle requires the modeler to have serious justification for increasing the complexity or dimensionality. An alternative and emerging paradigm known as “Keep It Descriptive, Stupid” (KIDS), is to build descriptive model that has a closer correspondence with the reality or target. The model is built as realistic as that is permitted by evidence and resources [27]. Accordingly, the model may be simplified, based on subsequent understanding and experience. Both approaches have merit under different circumstances.

Well-built simple models, adhering to KISS principles, tend to focus on a single aspect or phenomenon of the system in study, but are naturally not very suitable for representing a multi-dimensional, complex systems for which no simple descriptive patterns have been found. Nor do these simple models, by their structure, take into account all available evidence and domain knowledge, but rely instead on assumptions (some contradicting reality) to carry the weight of the model and bear the responsibility for deviating from reality. For a model intending to provide learning, exploration and immersive training in social systems, descriptiveness is required to be high and “curse of dimensionality” is a given. Besides, data for social systems tend to be scarce, and data collection for such a model requires innovative approach. Owing to the significance of data collection and modeling activity to validation, we have a separate subsection on data and modeling methodology.

Finally, dichotomous discussion of KISS versus KIDS as mutually exclusive paradigms imply a binary decision waiting for the modeler (see footnote 4). In reality, what we should be concerned about is the level of descriptiveness in the model, which must be selected in commensurate with the purpose of the model, complexity of the target and the resources available for modeling. Besides, we may be deviating from KISS principles, but a model of models such as ours does have many intertwined sub-models, each of which might be deemed relatively simpler.

2.2 Model Summary

We have built a framework, named StateSim, to model countries based on a 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. StateSim is an environment that captures a globally recurring socio-cultural "game" that focuses upon inter-group competition for control of resources.

Our multi-resolution modeling framework and the software were developed over the past ten years at the University of Pennsylvania as an architecture to synthesize many best-of-breed models and best practice theories of human behavior modeling. This environment facilitates the codification of alternative theories of factional interaction and the evaluation of policy alternatives.

The agents in this framework are cognitively deep and come equipped with values (short term goals, long terms preferences, standards of behavior including cultural and ethical values, and personality). The environment provides contexts. These contexts carry and make decisions available for consideration. These agents make decisions based on a minimum of two set of factors, (i.e. Decision Utility as a function of):

- Values: The system of values that an agent employs to evaluate the decision choices, and
- Contexts: The contexts that are associated with choices.

The values guide decision choices, and in our case, have been arranged hierarchically or as a network. The contexts sway the agent decisions by providing additional and context specific utility to the decisions evaluated. The contexts are broken up into micro-contexts. Each micro-context just deals with one dimension of the context (for example, relationship between the perceiver and target and so on). With a given set of values, an agent (or person) evaluates the perceived state of the world and the choices it offers under a number of micro-contexts, and appraises which of its importance weighted values are satisfied or violated. This in turn activates emotional arousals, which finally are summed up as utility for decisions.

The agents belong to factions, which have resources, hierarchies of leadership, and 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 and services, albeit imperfectly owing to being burdened with institutional corruption and discrimination. The public goods themselves are tied to the amount of resources for the institutional factions, including the level of inefficiency and corruption. The economic system currently in StateSim is a mixture of neoclassical and institutional political economy theories.

The cognitively detailed agents provides in-depth profiles of actual leaders and follower archetypes. This approach utilizes performance moderator functions (PMF) which are micro-models covering how human performance (e.g., perception, memory, or decision-making) might vary as a function of a single factor (e.g., event stress, time pressure, grievance, and so on.). "PMFserv", our cognitive modeling approach, synthesizes dozens of best-of-breed PMFs within a unifying mind-body framework and thereby offers a family of models where micro-decisions lead to the emergence of macro-behaviors within an individual. Further details of these PMFserv models are beyond the scope of this paper. For additional details, see [28] [29] [30] [31] .

For a given state being modeled, StateSim typically profiles 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 a population automata [32]. Diagrammatic representations of example StateSim models are given in Figures 1a and 1b, which show the politically salient factions in the countries made up of cognitively detailed agents and population models consisting of simpler agents. Not shown on the diagram are institutional framework and economy.

Most country models have a central government with state apparatus (e.g. military, Bureaucracy and executive branches). Frequently, cleavage of the country along social, ethnic, economic or religious or po-

litical lines might be significant and salient to the model. In Bangladesh, for example, military is the most salient state apparatus. Among majority population, two opposing political parties, namely BNP and AL, dominate the scene. AL, or Awami League, is a centre-left political party (led by Sheikh Hasina). BNP, or Bangladesh Nationalist Party, is a a center-right political party (lead by Khaleda Zia). Both BNP and AL have relevant sub-factions (in addition to main party body). The communist/ leftist and students constitute the sub-factions of AL while religious right and business interests form the sub-faction of BNP. CHT, the people of the Chittagong Hill Tracts (lead by Shantu Larma, the leader of the Regional Council), are the primary out-group or minority group relevant to the model. They were once the separatists, but have been brought together under fold of the country a treaty.

In Thailand (Figure 1b), the set up is similar with minor variations. A balance of power between Military, Bureaucracy, and Royal Family, has been key to Thailand's precarious political stability. The Thai Buddhist Majority and Southern Thai are the key population factions in Thailand. Thai Buddhist Population can be further divided into Rural Poor and Urban Elite based on their socio-economic status, and these factions frequently act with opposing political interests[1]. The Southern Thai Muslims are the minority or out-group, who can be further divided into Conservative (i.e. Moderate) and Radical factions. Radical factions essentially represent the Southern Thai Muslim rebels who are promoting separatist agenda.

Figure 1a – Overview of the Components of a StateSim Model: Bangladesh

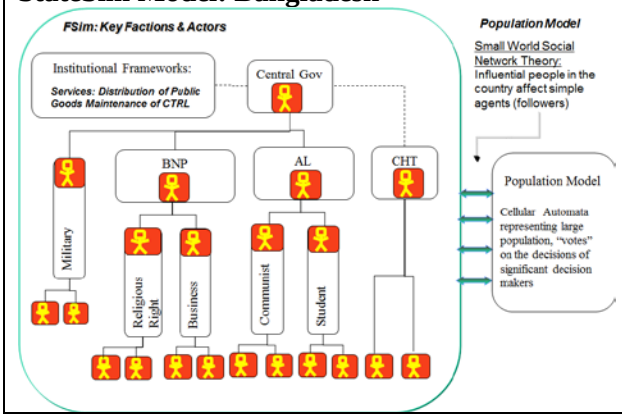
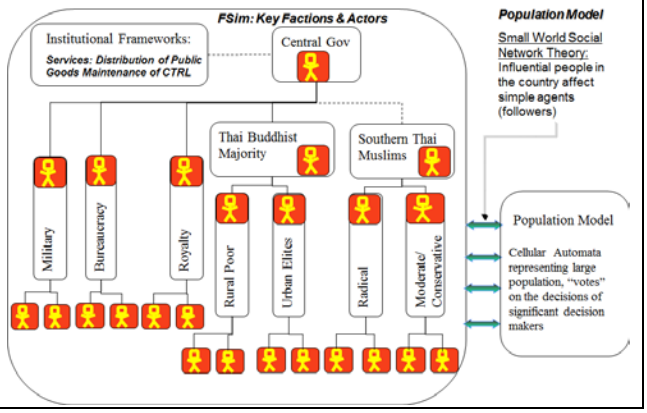


Figure 1b – Overview of the Components of a StateSim Model: Thailand



Social system models like these are complex, with imprecise, incomplete and inconsistent theories [33]. In addition, according to De Marchi, these models have very large feature spaces (on the order of 1000), giving rise to "curse of dimensionality" [34]. Evaluating such models is challenging.

A 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 [34], are needed to increase confidence in such complex models. This is a topic we turn to next.

3 A FRAMEWORK FOR MODEL EVALUATION

This section presents a conception of evaluation that is aligned with the modeling life cycle. Model should be viewed as a System with an entire cycle of model conception, data collection, model building, testing, verification, validation, exploration, as well as learning and continuous evolution. This approach is also in alignment with the prevalent literature. Accordingly, model evaluation is a gradual, systematic and iterative process of continuous evolution.

[1] Some variations to the above configurations were also modeled, but, for our purposes, it suffices to limit our very brief descriptions to the above base cases.

This view is not without controversy, and therefore, we would like to briefly explore the existing paradigms in validation. Barlas and Carpenter [35] group the extant philosophical schools with respect to validation into two diametrically opposing camps. Barlas and Carpenter's discussion has been summarized in Figure 2 as:

Reductionist/ logical positivist school (including subscription to empiricism, rationalism, verificationism and the "strong" falsificationism): • Models are seen as objective representation (one or more aspect) of reality. • Validation is an objective exercise of determining the model validity as binary outcome (as either "correct" or "incorrect") based on model's ability to correspond with real world data. The validity is about accuracy, rather than fitness (/ usefulness) to purpose.	←→	Relativist/holistic school (including subscription to relativistic, holistic philosophies): • Model is essentially one of many subjective representations of reality, for every model carries in it the modeler's worldview. • Outcome of validation is not a binary process of making "true/ false" or "accept/reject" decisions, "but lies on a continuum of usefulness" and is to be seen as an exercise in "confidence building".

Figure 2: A Typology of Validation Philosophies. Summarized from [35]

It is true that in developing the dichotomous typology of philosophies, Barlas and Carpenter may have polarized the perspectives on validation into two extremes. However, such caricatures are useful in developing the subsequent argument.

Drawing on seminal articles dealing with validity in system dynamics modeling [36] [37] [38], Barlas and Carpenter show that the relativist/holistic philosophy is more appropriate for systems dynamics school, owing to its descriptive nature. This position might be even more true for agent based modeling with socio-cognitive agents.

Thus, the holistic approach to validation includes consideration of the life cycle of modeling activities at the onset, proactive management of validation, and continuous improvement (rather than treating modeling as a one off activity). Even in the conventional validation articles, researchers have espoused the life cycle view of validation, verification, and testing (VV&T) and state the applicability of VV&T for each stage of the modeling and simulation life cycle. E.g. [39] [23] [40]. According to Balci, "the VV&T is not a phase or step in the life cycle, but a continuous activity throughout the entire life cycle." Carley has also described a similar Model-Test-Model paradigm. They clearly state the applicability of VV&T. The biggest advantage of this iterative paradigm is the opportunity for continuous evolution of the model, instead of selecting the model to be "in" or "out" in a binary fashion.

3.1 Frameworks of Model Validation in Literature

Typologies for validating simulation models is detailed in a number of articles, including by Knepell and Arangno [41], Sargent [42], Balci [39] [23], and Yücel and van Daalen [43]. While they all agree in spirit and convey the same overall meaning about validation, they have not always been consistent in their use of the language. Yücel and van Daalen describe a typology with: (1) structural representation assessment and (2) behavioral representation assessment. In structural representation assessment, which is variously known as structural validation [44], theoretical verification [45], structure assessment [46], conceptual validation [47], interactions, processes and elements in the model are validated. Behavioral representation assessment, which is also known as operational validity, external validity [45], behavioral validity [40] [46], replicative validity [44] [16], determines how closely the model replicates the real system.

According Knepell and Arangno [41], one or more of the following six types are often discussed in relation to validation: "conceptual, internal, external, cross-model, data, and security". Adequacy of the underlying conceptual model in representing the real world is given by conceptual validity (also known as theoretical validity), while correctness of model construction and transformation from one form to another are referred to by internal validity. The transformation of models (e.g. code constructed from design is free of errors) is more generally called "verification", a topic we are not covering. Silverman et.al. (we)

[30] used the term "internal validity" to refer to the micro-validity of some of the cognitive sub-process models and how they correlate with real world behavior. External validity (also referred to as operational validity), on the other hand, refers to the adequacy and accuracy of the model in replicating real world data. Cross model validity is similar to external validity, except in this case, one compares two models as opposed to comparing a model with reality. Data and security validities respectively refer to accuracy and adequacy of the data, and safeguards in place to protect data integrity.

Thus, it is clear that multiple categorizations exist in validation. An increasing number of modelers, including rationalists, support the idea of using model evaluation (as opposed to model validation) to describe the gradual, iterative process of achieving confidence in the model. These researchers prefer to use the term "Evaluation of the Models" instead of validation [12] to emphasize an encompassing meaning of not only qualitative or subjective nature, but also the positive relationship with credibility, appropriateness and use of the model. Rykiel [47] and Oreskes [48], for example, give some additional insights into the subtle differences. According to Rykiel, validation is about determining fitness for purpose by demonstrating that the model satisfies "prescribed performance criteria". According to Oreskes [48], "evaluation implies an assessment in which both positive and negative results are possible, and where the grounds on which a model is declared good enough are clearly articulated". For Oreskes, "validation implies an exercise in legitimization". Besides, there are no equivalent and elegant words to "valid"(the state of being valid) and "validity", in the evaluative paradigm. The distinction is blurred; historically, most literature refer to validation rather than evaluation; contemporarily, a number of researchers continue to use the terms interchangeably, using "evaluation" in the title, but "validation" inside the body of the articles.

In their defense, not only has the term "evaluation" not taken root in the literature (as much as validation has), but also the vocabulary for evaluation paradigm is not complete yet (e.g. there are no equivalent terms associated with "evaluation paradigm" for "validity").

For us, evaluation refers to the overall process that includes not just model performance but also establishing credibility for the model.

It can be noted that for traditional models the validation/ evaluation falls into two broad categories, namely: (1) Interior Focused Evaluation: Conceptual, Structural or Internal Validation (or includes and is largely Verification, especially for those building models in the hard sciences areas), and (2) Exterior Focused Evaluation or External Validation. Socio-cognitive agent based modeling differs from other forms of models in one significant aspect in that one has to make judgments about values and contexts in the agent based world by looking at the evidence available.

We also introduce two sub-categories of evaluation, namely methodological evaluation and narrative validation. These have been made subservient to and be a sub-section in, conceptual and external validity assessments respectively. Thus, we will describe the evaluation process through the two internal and external focused assessments.

However, a note must be made of the difference between verification and validation, because this is fundamental and is often mis-interpreted too. The distinction between the model validation and verification, according to Balci, is respectively building the right model and building the model right. Balci outlines 15 principles and 75 techniques in his 1998 paper [23], including those that are precursors to a number of techniques that we employ. However, Balci or other researchers such as Rykiel do not point to any inherent differences between Verification and Validation (or even Testing) techniques themselves.

Instead, it is the context of applicability, namely whether one is assessing the performance against the specification on which the model is constructed or against the external observations (equivalently whether the data used is training data employed for model construction or is an independent set of test data), that distinguishes between verification and validation respectively.

Since this paper is predominantly about techniques and there are no serious differences between techniques applied for verification and validation, we report only the salient evaluation activities, leaving behind various unit tests, system tests and many verification and validation activities carried out throughout the modeling life cycle.

3.2 Purpose and Scope of Validation

A model must be evaluated in the context of, and in commensurate with, its intended purpose and uses. In our case, the models are developed to answer questions about key behavioral factors contributing to selected set of instability. This takes particular importance because, as we will see later, we could not assume ontological truth about the models. 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. [As noted elsewhere, we do not advocate using models to make point predictions in social systems].

That said, we begin our description of model evaluation with a caveat, that in social systems models, the ultimate objective of the model should not generally 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.

In a quote variously attributed to Arturo Rosenblueth Stearns and Norbert Wiener, "The best model of a cat is another cat..., specially the same cat". Even complex and elaborate models should not be equated to the reality that they attempt to represent, and therefore, are not suitable for answering questions outside the scope of the models. Yet, it is not always easy to define the scope of the models. Part of the scope is determined at the onset, by defining the purpose of the model, the key questions model would answer as well as the level of abstraction in the model. The remainder is constrained by the assumptions, both implicit and explicit, made at the time of model construction. Availability and quality of data and the quality of the modeling process would also limit the scope of the model. Finally, the phenomenon or system under study, its variability and predictability would influence the scope. It could be argued that even the act of external validity assessment can be construed as evaluating the scope of the model outside the input space, thereby, giving a validation based perspective to model scope.

Accordingly, 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 attempts to model, albeit under limited scope. For example, how much of mechanism that a model can elucidate of course depends on the level of abstraction in the model. [The level of abstraction also guides the level at which validation is carried out]. Models are developed at a given level of detail and scope. It is impossible (and is not the objective) to validate a model at a smaller granularity than at which the model is constructed. For example, if a model is constructed of social system, one cannot expect to replicate various molecular processes in reality. In mechanism based models, we would argue that it is best to statistically validate (and claim validity for) processes at least one level coarser than the model is built at. Correspondence may be obtained for the existence of processes at the same level as the model building. We build our models which attempts to replicate the decision making processes of the agents. In this, agent values, their perception of contexts including historic, socio-economic and physiological stressors, for example, might be constraints on decision making capability. We expect statistical correspondence at the level of individual (actors and institutions etc) decisions and societal decisions and not at the level of sub-processes that lead to them.

Without implying 1-to1 correspondence between model and reality, we believe that a social system model should be able to point in the direction of underlying processes or events than a high level correspondence. Arguably (and arbitrarily), at least one level deeper than the macro-level correspondence is desirable not only for explanation, but also to control the risk of being beguiled by the phenomenon of equifinality[2]. If agent based models would not provide this additional detail, they are no different from statistical models, but at a higher price.

In the following sections, we will present a number of evaluation techniques. The list of techniques are neither prescriptive nor minimalists, but have been presented as cases. Each technique has been se-

[2] Equifinality is defined by Von Bertalanffy as "the tendency (to reach) towards a characteristic final state from different initial states and in different ways" [90].

lected to illustrate one or more specific aspect(s) of the validation challenges. The considerations for selecting the validation techniques broadly includes, but not limited to: (1) the purpose of the model (2) validation criteria adopted, (3) availability of data, (4) potential errors anticipated and (5) nature of the system (e.g. specific challenges such as stochastic nature and presence of counterfactuals).

We hope to spur research in these directions by providing some cases of evaluation 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.

4 CONCEPTUAL OR INTERNALLY FOCUSED EVALUATION

In order to ensure conceptual evaluation, 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, fill in what is missing. Instead of relying on law of parsimony, we build data-rich, descriptive approaches. We seek approaches that drag in necessary, but exoteric detail (which are often left behind in the name of parsimony) 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 a high level of fragmentation. In validating the model, we assess the model construction methodology, the structure, processes and mechanisms. At the level of components, we assess completeness, clarity, coherence, and robustness. It is also worth noting that conceptual evaluation predominantly consists of verification activities for it always compares the existing model against the intended specification (equivalent of training data) and not against an independent set of data from external reality. [For a simulation model, the specification is the training data. The verification process attempts to re-create the very empirical evidence used for constructing the model].

4.1 Structural Evaluation

Structural evaluation is not only a pre-requisite for model 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, this (be verification or validation) strengthens the triangulation strategy.

Congruence Inspections: Structural congruence is a key part of conceptual validity. A model is deemed structurally valid, if the model structure (e.g. actors, factions, parameters, qualitatively underlying relationships between parameters) does not contradict the structure expected in the real system. Most tests against the structure tends to take the form of verification than validation. Here, one tends to verify the knowledge that has already gone into building the model and hence cannot be deemed independent of model construction. Therefore, such structural or conceptual validation exercises are frequently internal validation activities, best called verification.

We also carry out structural or conceptual 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 structure and behavior against specifications/ expectations through systematic inspection.

Degenerate and Extreme Value Tests: Even in conceptual or internal evaluation, there are some tests that might be regarded as validation. In a multi-level model, high level or emergent patterns that are not deliberately designed for can be correlated with reality and will qualify for validation activity in the orthodox sense. In addition, in almost every model, 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 [23]. The degenerate tests were combined with extreme bound analysis, where we determine whether the

model continues to make sense at the boundary conditions or extreme values [49]. 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. In rare cases, when thorough investigation following an "anomaly" fails to reveal any errors, there are possibilities of models correlating with existing theories or phenomenon that are not coded into the model. This would validate the model. Generally, 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.

4.2 Theoretical Adequacy

The next level of issue in conceptual evaluation is assessing the epistemological and ontological adequacy. The social sciences often tend to be fractured into disciplinary silos, yet social dilemmas and cultural concerns cut across such silos. Stakeholder issues often include aspects from all disciplines at once. As a result, we do not want a single model of a narrow behavior, but the ability to synthesize a range of the relevant models. If done well, this synthesis might also help to identify gaps in the science and suggest new research directions.

Accordingly, as we mentioned earlier, our framework synthesizes dozens of best-of-breed theories as PMFs or performance moderator functions. None of these PMFs are "home-grown"; instead they are culled from the literature of the behavioral sciences. These PMFs are synthesized according to the interrelationships between the parts and with each subsystem treated as a system in itself. Elsewhere we have discussed how the unifying architecture and how different subsystems are connected [50] [29].

We believe that it is important to determine whether individual theories and models have been faithfully represented, whether the combined set of theories and models implemented work well together and what gaps need to be filled in. When evaluating the collections such as sub-models as well as the theories from which they are derived, considerations must be given to purpose, context, gaps in theories and the assumptions required for operationalizing theories into models. Such an assessment has not been a trivial proposition even with simpler models. In social sciences especially, this often tends to bring out a sober reality of less than perfectly valid theories and sub-models. For example, not all models are quantitative and most models have gaps.

However, there is every reason to be optimistic, albeit without expectations of strict timelines. In recent years, "the mechanism-centered explanation and analysis" has become part of the mainstream discourse in social sciences [51] [52] [53]. This is a welcome break from the dominant tradition of "correlational school" (aka neo-Humean positivist school) that had once monopolized the quantitative research landscape. According to "positivistic" view, causality is interpreted as a probabilistic association (including "constant conjunction") between variables. Mechanism-centered school (and its variants such as "causal reconstruction") advocate(s) white-box models and drilling down to explanation using mechanism [54] [55].

Causal reconstruction, which, at least in theory, would deal with processes (not correlations), intends to produce historical narratives using causal mechanisms. In time, such approaches would result in robust models. Depending on their scope and coverage, one or more of these models can be synthesized and be employed in social system models.

According to Balci, valid sub-models or theories alone do not guarantee valid models (i.e. integrated models). That would mean synthesis of sub-models, where the models only provide partial explanation or mechanism. Frequently, the sub-models may be developed with different purpose, context or assumptions unaccounted for in the integrated model. This is another place where agent based models are particularly useful.

Firstly, a modeling framework with capability to synthesize and implement theories would assist the progress of science by providing a testbed where theories could be tested in a transparent and concrete fashion. Theoretical (especially Ontological) adequacy is determined by examining the model for necessity and sufficiency of existing theories, including under-specifications in theory, the gaps in theories and their potential influence on the results.

Secondly, it becomes necessary to address various (sub)model compositional issues. According to Szabo and Teo [56], the validation process must address model "compositional" issues. In model construction, we also look for logical consistency or contradictions, temporal or behavioral issues over time, and formal aspects such as metrics. Specifically, we might be interested in the appropriateness of operational variables, mappings between models, and ensuring that plumbing work is carried out in a satisfactory manner. Thus, internal consistency (within sub-models), and external consistency (in the neighborhood) and behavior need to be ensured.

As an example, one interesting case is the development of bridging models to integrate models of different levels of aggregations. Among other considerations, constructing a country model requires use of agents at different levels of details, as "thinking, feeling" cognitive agents are computationally expensive to construct an entire country with. Salient agents are constructed using cognitively detailed agent framework, but the lower level population that does have significant contribution to the course of decision making (in other than supporting or opposing different decisions made by leaders of various salient groups) are constructed with simpler agents. Merging the population model with cognitively detailed StateSim agent model is a case in point, as the two models operated at different levels of temporal and spatial dimensions, albeit describing the same region. Specifically, both the models had different time periods of simulation and the number of agents in the cognitively detailed and cellular automata/ network were also different. Therefore, we created an intermediate model based on small world theory that specifically handled the integration. Use of the small world theory as a cementing, meso-model enabled integration of the two models. Use of such intermediate models, which complement and cement the existing models by selecting theories from social sciences in order to create a more complete view of the behavior of the system at intermediate levels of analysis, is frequently helpful in integrating existing models. When making cross-level linkages such as this, it is important to make sure that the micro- and macro-level variables are within the same time frame, and are referring to the same variables or dimensions. Names alone are not sufficient or even necessary indicators of the same dimensionality and time frame.

The mode of our examination is by human inspections, although we have various automatic tests checking for obvious inconsistencies, particularly deadlocks, safety etc that can easily covered by automated test scripts. However, these are not as rigorous and exhaustive tests as what semantic composability researchers (such as Petty and Weisel in [24] and Szabo and Teo in [56]) prescribe. We rely on human inspections to complement automatic testing and pick up the slack. In one of their recent papers, Szabo and Teo [56] have shown a proof of concept for composing sub-components, it is our hope that such techniques would mature enough to encompass descriptive models from their coverage of logical models today.

In summary, any model of models collection, such as StateSim, and particularly so in the social sciences, suffers from the fact that each of its component theories (models) is less than perfectly valid, and the interstitial programmatics of the ensemble tend to be even less reliant on theory (since it generally straddles a gap between two or more subfields). This, combined with large feature spaces, means that there can be no truly objective standard for internal validity assessment of such collections. Nevertheless, the better than 70-80% (as we will see later in external assessment) statistical correlations of the trials done to date between simulated and real world agents suggests that it is important to struggle with these issues and attempt to use best of breed component theories and invest in social scientists to help with the interstitials (as has been done for StateSim).

Besides, our framework is open and is based on synthesis of best-of-breed social and behavioral theories. This is pivotal as the social science is inherently fragmented into sub-disciplines and is still evolving; Facilitating continuous learning and improvement in our framework is essential to bridging the gaps be-

tween theories and in turn improving the model (and vice-versa). In our framework, as the science of social systems advance with time, these sub-models (or PMFs in our case), which are replaceable, will be either refined or replaced with new best of breed models. Thus the entire system of models is amenable to continuous improvement, ultimately bringing about an evolutionary growth in virtual social systems.

4.3 Methodological Evaluation: Of Data and the Process of Building Model

Owing to complexity of socio-cognitive agent based models, gaps in theories, and quality and extent of data available, it is not appropriate to solely rely on selected tests and inspections after the construction of the model. For example, it is also difficult to carry out thorough verification and conceptual evaluation in an end-of-process fashion because the model runs deep in many levels and layers. Other researchers (such as Bankes in [49], and Bankes and Gillogly in [57]) do recognize the difficulty of validating complex system based exploratory models and point to the need to validate the modeling process. Long before the question of validation arises, consideration must be given to understanding how the model is constructed. Many researchers have advocated a life cycle based, cradle-to-grave approach to modeling (e.g. Banks in [49]). Accordingly, our methodology includes an iterative and complete life cycle of steps to manage the modeling - including conception (framing the problem and developing conceptual model), construction (synthesizing theories, developing structural models, gathering data), evaluation (testing, verification, validation and sensitivity analysis), exploration, and retirement or renewal. In line with holistic paradigm, the modeling process should also be concerned with fitness to containing system and usefulness to the stakeholders [58], and the analytical aspects should be embedded inside one or more systems inquiry paradigms such as Soft Systems Methodology [59], Critical Systems Thinking [60] and Systemic Intervention [61], Dialectical Paradigm [62] [63], Action Research [64] to name a few (and we are not the first to suggest this). However, for the purpose of describing evaluation (and for the sake of brevity), we will limit our discussion to things of immediate relevance and value to validity.

In this regard, special consideration for data, management of cognitive biases, continuous improvement, replicability and transparency are essential considerations in the management of validation. The data issues take particular importance even outside the holistic validation paradigm that we have adopted (e.g. [41]) and are emphasized appropriately in the ensuing section on methodological evaluation.

In assessing methodological validity, it is desirable to assess two key aspects, namely modeling process adequacy and software process adequacy. Software Process Adequacy is the more well-understood part of the assessment, as software development verification and validation are well established practices. We will not be concerned with that 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 the model is constructed defensible? Are steps being taken at the process level to control errors, cognitive biases, and especially the confirmation bias? Are there provision for incorporating appropriate stakeholder inputs and continuing interaction with the model? 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, [65] [33]). We briefly recap the salient features relevant to model validity here.

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. The goals are 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. Use of qualitative or empirical evidence is not unique to us (see e.g. [66] [67]). To do this, we have employed web newsfeeds, country databases, and SME interviewing.

These models are a built on a decision framework that uses both decision theory and structured knowledge. 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, consolidat-

ing the information to build a model of the social system. 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 a systematic approach increases confidence in the models.

We assembled an integrated index of all the parameters 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 these data sets were not custom-designed for our model, we have to select proxy measures for our parameters of interest. 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 [33]. For example, the discordance in the unit of analysis can be resolved by cross tabulating and sorting these survey databases according to properties that categorize survey respondents into specific groups that match our interests. The surveys are sufficiently detailed to allow us, for example, to infer information about whether an average supporter of a particular political party has a more or less materialistic vision of life than another average supporter of another political party or a different faction.

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. There are some difficulties with the use of these materials. Coverage is a concern with the databases as well as with the newsfeeds. Another challenge in using automated content analysis tools lies in building the catalogue that contains the necessary categories of key words and their combinations, both (or all) of which represent our model parameters. In addition to having proper keyword synonyms, it also may be that a schema or model of a given parameter has to be constructed to accommodate the interpretation and transformation of proxy variables. One also needs to test the error rates of all the extraction tools. This implies assembling a test corpus in addition to a training data set where all the ground truth is known. Instead, we largely use these web news feeds for background information and sanity checking what our Subject Matter Expert (SME) survey produces.

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 or by conducting open-ended interviews. For our purposes, administering a structured, self-explanatory web survey tailored to elicit exactly the information we need would in most cases be preferable to conducting unstructured, open-ended interviews (partly because these interviews would elicit a wealth of information that would then need to be sorted and coded).

There are three main difficulties associated with using SMEs to elicit the information we need. First, eliciting information from SMEs incurs significant financial and human resources. Second, SMEs, by virtue of being human are fallible, may sometimes provide us with biased and, occasionally, even blatantly incorrect information: e.g., see [68]. Being a country expert does not mean that one has complete and comprehensive knowledge, instead, one may need to seek out multiple experts just to get full coverage. For these reasons of limiting bias, we would want to consult more than one SME on any particular country or topic (which means expenses and availability). Third and finally, simply finding SMEs for a particular country of interest, especially those which are not widely studied, may by itself pose a significant challenge. In summary, this seemingly direct route of eliciting parameters information from SMEs is also beset with difficulties. [Therefore, we vet the SMEs ahead of time and then we verify and triangulate at least a sample of SME information against other sources and other SMEs.]

To assist in this process, we have authored and assembled a survey that is self-explanatory and has a validated set of questions about each parameter needed in a socio-cognitive agent model. This has been employed for eliciting knowledge from country or leader desk experts. This web interview, designed, implemented and tested during previous projects is used by SMEs to initially fill in these parameters in

about 12 hours time for any given country of interest. This interview follows the acronym of FAIREST (faction-by-faction actors, institutions, resources, economics, supra-system, and timelines). We also use the FAIREST scheme to stimulate conversation about new idealized designs (and metrics) that seek to mitigate negative factors and promulgate positive ones. FAIREST factors were elicited for each major factional group in a region. A summary of the parameters is given in Appendix 1.

Example: Eliciting Expert Input and Differential Diagnosis

The following example illustrates the use of surveys and differential diagnosis. SMEs use these surveys to provide parameter estimates as well as reliability which is their confidence in their estimates. Then using these estimates as hypotheses, we conduct differential diagnosis to spot-check the estimates by the SME against other multiple sources, including other SMEs.

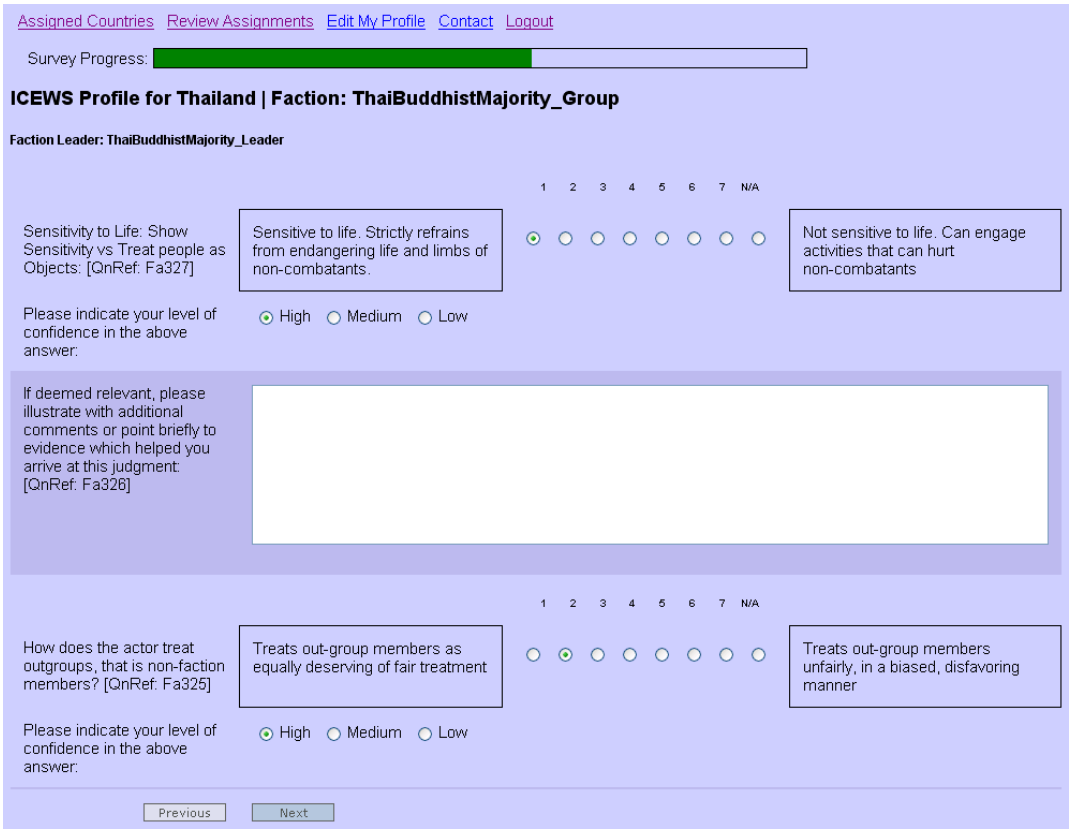


Figure 3a Example Snapshot from the Questionnaire that we had designed

The SME questionnaire that we have designed (see Figure 3a for a snapshot) directly elicits the parameter by presenting them on a linear scale with diametrically opposite traits and asking for their assessment on a Likert scale that would match the personae's profile. This is a straight forward process, but the design is worth mentioning. The value parameters, whose values are being elicited, are structured hierarchically in a tree with values in the nodes and importance weights on the arcs. The weights of the arcs were semi-quantitatively assessed against the competing arc at the same level, a direct comparison process. The assessment questions were designed as a structured survey by using the arcs to be pitted against its pair. This system for assessment of the weight permits a reduction of consolidated complex evaluations of alternatives to a long series of pair-wise comparisons, in which the accumulating results are stored for later calculation while the user (in this case, the subject matter expert) can focus serially, distinguishing only two qualities or quantities at any one time. This design reduces judgment errors.

The Figure 3a shows a small section of the questionnaire employed, where just two nodes in a given level are being considered at a time. Similar questions are posed for other pairs of attributes. Figure 3b shows how one estimates the weights from the survey data. Not shown are the additional applications and details of the pairwise comparison process that also supports elicitation of relative balance between pairs of traits.

Let us look at the leadership's trait (see Figure 3b), namely Humanitarianism (also referred to as Sensitivity to Life), which is described by the pair of diametrically opposing nodes: Show Sensitivity vs Treat people as Objects.

Figure 3b Sensitivity to Life: Show Sensitivity vs Treat people as Objects:

Sensitive to life. Strictly refrains from endangering life and limbs of non-combatants.	1	2	3	4	5	6	7	Not sensitive to life. Can engage activities that can hurt non-combatants
	○	○	○ x	○	○	○	○	

Score (S) from 7 pt Survey = S = 3

Extent of Sensitity to Life = Normalized Ratio Score = (S – 1)/ (7 – 1) = 1/3

Wt (Sensitivity to Life) = 0.33 and Wt (Not Sensitive to Life) = 0.67

The survey has proved to be of significant value. However, for the reasons mentioned earlier, we do not solely rely on SMEs, but combine it with other sources of information as mentioned above. The input data obtained from multiple sources, tend to be incomplete, inconsistent and noisy. 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. Therefore, in the course of constructing these models, there is the risk of contamination by cognitive biases and human error.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
6			K-ve		1									
7			K-ve		10									
8			**Evidence (Ei) for CountryT Leader (TS)**											
9			**Evidence (Ei)**	**Relevance**	**Reliability (Ri)**	H1: Sensitivity to Life	H2: Lack of Sensitivity to Life	H3: Wanting to Protect the Country & Resources	H4: Taks Focussed Behavior	H5: Catering to Popular Support	H6: Being Closed	H7: Loss of Control - Querky/ One off event	H8: Entirely due to External Circumstances	**Frequency of sighting this types of event: P'[E|c] (Typical?) Qualitative**
10														
11		**Theme**	How to account for TS's Record on Human Rights - Could it be sufficiently explained by internal factors? Or Would it require some external factor to be considered?											
50		Discriminat	Prime Minister announces financial aid to "green" villages but not to "red" villages	1.00	1.00	-1.00	1.00			-0.50	1.00	0.50		Medium
51		Mil	Soldiers clash with insurgents to slow increased violence	1.00	0.70							0.50		High
52		Mil	Government tries to manage increased violent acts at train stations/tracks	1.00	1.00							1.00		Medium
53		Assist	Prime Minister declares light criteria for compensation to Southern families of victims	1.00	1.00	1.00	-0.50	0.50	-0.50	1.00	-0.30	-0.50	-0.50	Low
54		Mil	Prime Minister pledges to Southern youth to stop violence	1.00	1.00	0.50		1.00		1.00	-0.20	-1.00	0.50	Medium
55		Counter Te	Government initiates discussions on SIM card regulations due to the fact that a lot of the bombings are detonated through the use of cell phones	1.00	1.00							0.50		Medium
56		Assist	Prime Minister offers monetary rewards for exchanges of guns	1.00	1.00				-0.20	0.50		-1.00		Low
57		Assist/ Neg	Prime Minister order a series of peaceful measures	1.00	1.00									Medium
58		Negotiates	Prime Minister initiates talks on Emergency Law	1.00	1.00	1.00				0.50	-0.60			Medium
59		Media	Censor the media in the country's 3 Muslim dominated southern provinces, have the right to detain suspects without trial for a number of days, tap telephone lines, monitor email exchanges, confiscate suspect's property in Yala, Narathiwat and Pattani and install curfews.	1.00	1.00	-1.00	1.00	-0.50	0.50	1.00	1.00	1.00		Low
60		Mil	Troops raid Islamic establishments in Bangkok	1.00	1.00						1.00			Medium
63			Confidence Index			-0.29	-0.48	-0.29	-0.18	-0.10	-0.22	-0.11	-0.93	
64														

Figure 3c: Excerpts of Differential Diagnosis of Leader [About Humanitarism] [Abridged]

Let us illustrate briefly how triangulation of various, often contradictory sources are resolved through the process of differential diagnosis. Figure 3c shows a small sample of the diverse sources of information collected for assessing "humanitarianism" of one of the leaders being profiled. This evidence table accu-

mulates inputs from many sources as can be seen. These pieces of evidence are organized through a thematic organizer and thematically coded and attributed with reliability, relevance and typicality, as shown in the left columns of Figure 3c. Figure 3c shows a matrix, where the hypotheses are pitted against each other as mediated by the evidence. Recall that the hypotheses in this case are the nodes in the Value Trees (goals, standards and preferences of the characters being assessed) [with weights].

Hypotheses (parameter weight estimates for example) are pitted against each other by enabling disconfirming (to a lesser extent confirming) evidence to play on them. Using this process, the hypotheses are weeded. For example, when constructing models of behavior from evidence, the differential diagnoser attempts to determine the motives of someone's behavior, and ascribe locus of control for the behavior (internal, external, or chance). Specifically, it assesses evidence and attributes behavioral traits by evaluating competing hypotheses and various evidence that support (confirm) or oppose (disconfirm) these hypotheses. In these cases, confirmation bias is eliminated by giving higher weight to disconfirming evidence; and mirroring bias is reduced by rendering the explanation transparent and subject to scrutiny.

The differential diagnose involves estimating/ ascribing frequency (or likelihood), reliability, relevance and typicality for each piece of evidence. Relevance identifies which items are most helpful in judging the relative likelihood of the hypotheses, and help control the time spent on what would seem like irrelevant evidence. In pitting evidence against hypotheses, if a reliable piece of evidence rejects a hypothesis, then the likelihood of that hypothesis is diminished significantly. Likewise, atypical but relevant events (given they have occurred) are quite informative and will also be given higher weight [rare events weigh higher when they do occur as evidence]. In order to favor rejection of hypotheses, disconfirming evidence are weighted heavily (an order of magnitude more than confirming evidence). Combining these factors, differential diagnosis estimates an index (similar to posterior probability) of the likelihood of the hypothesis. [We find it better to work with a confirmation index than probabilities]. Specifically, let us assume that the evidence (E_i), with a reliability Re_i, rejects (or supports) a hypothesis (H_j) with a strength (C_{ij}), where $C_{ij} \in (-1, +1)$. C_{ij} value of +1 implies full support, while –1 implies complete rejection, as assessed by the expert or knowledge engineer. Then, the simplest form the confidence index can be expressed as:

$$CI_{Avg}(H_j) = \frac{1}{n} \times \sum_{i=1}^{n} K \times C_{ij} \times Re_i \times U_i \times f_i \qquad \text{------ (Eq 1)}$$

where f_i is the frequency of evidence (if similar evidence is clubbed together); U_i is the extent of untypicality of evidence, so that it is informative and not mundane; and K = { w_1 when $C_{ij} >= 0$, and w_2 when $C_{ij} < 0$ }. Essentially, K is used to assign a higher weight (say, an order of magnitude) to disconfirming evidence ($w_2 >> w_1$). We have used w_1 value of 1 and w_2 value of 20.

The competing hypothesis that has the highest positive confidence wins only if the hypotheses are mutually exclusive, the difference in CI is significant ($\Delta CI_{Avg} > 1.0$), and the variance is small. For hypotheses which are not mutually exclusive, ordinal ranking might be obtained. When mutually exclusive hypotheses cannot be clearly distinguished by their confidence score, multiple competing hypotheses might have to be entertained during the course of the sensitivity analysis.

The model thus constructed through techniques described above is subsequently tested, verified and validated through techniques described elsewhere in this paper. The modeling methodology also has provisions to maintain, review and improve the knowledge base over time.

A discussion of methodology will not be complete without any mention of the automated tools that support that efforts. We have been attempting to automate the end-to-end workflow of modeling and simulation activities. As it stands, workflow has been automated into chunks with human intervention being required to fill in the gaps. We are currently trying to assemble an automated knowledge extraction workbench (see Appendix 2). As mentioned earlier, our rudimentary ad-hoc testing system, with limited verification capabilities, checks for specifications set a priori. We do not have any external and generic verification/ validation system. Our efforts at validating the model involves carrying out simulation experiments and traversing the decision space. We employ an in-built experimental designer with MonteCarlo

simulation features. All tracers are logged and analyzed externally in statistical software after the experiment is completed. Our generic metric system enable us to build aggregate metrics (Events of Interest or EOIs for example) on top of the model output.

Recently, some researchers have been pushing the boundary of modeling and simulation workflow automation, especially evaluation experiments by designing automated systems or elaborate models that help with the design of simulation experiments, based on the understanding of the real world, specify important constraints that the model should not violate, monitor various parameters specified during the design process, highlight any violations of constraints and generation of any extraordinary values, and log decisions, events and other output parameters for post-simulation data analysis. Examples of these types include JAMES II (JAva-based Multipurpose Environment for Simulation II) [69] [70] and VOMAS (Virtual Overlay Multi-agent System) [71]. In these systems, the experimental design and/or validation design can be specified by SMEs. While VOMAS is focused and specialized on validation, JAMES II, concerned with broader workflow issues to standardize the modeling and simulation processes, appears to have conceptual or preliminary models of entire workflow management system [72].

5 EXTERNAL VALIDITY

During the validation exercise, the model would attempt to correlate with reality by creating scenarios constructed from a fresh set of empirical evidence hitherto unused in model construction. In this section, we present a selection of techniques to illustrate key issues encountered in external validation (As a consequence, the techniques are covered to different lengths, as required). A social system built primarily of cognitively detailed agents (such as PMF Serv based StateSim) can provide multiple levels and forms of correspondence. In the following section, we have described validation techniques relevant to quantitative and qualitative forms. In conformation with Gilbert [15] [see Section 1], we have also described quantitative validation at two different levels of abstraction, namely observable and aggregate levels. At observable levels, the model might typically have correspondence in behaviors (decisions agents make) and other measurable parameters (e.g. Gross Domestic Product (GDP), public goods service levels received). Quantitative correspondence 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. Aggregations are patterns (albeit quantitative) that can be akin to comparison of configurations that Gilbert recommends (see discussion in Section 1).

5.1 Quantitative Macro-Level Validity

The ensuing section summarizes and illustrates macro-level validation with one of our models (additional details can be found in [29]. In illustrating macro validity, we chose to present: (1) longitudinal plot of predicted likelihood versus ground truth, (2) the summary table of accuracy, recall and precision as summary measures, and (3) ROC curve showing comprehensive summary with sensitivity analysis with respect to thresholds. The direct (or default or base) outputs from the StateSim model include decisions by agents, levels of resources, relationships, emotions 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. Unlike EOIs which are more abstract metrics, indicators, by definition, are quantifiable and tangible measurements that reflect the EOIs. Typically they are count up events of that type, or averaged values of the parameter as the case may be, that arise across time in either the real or simulated world.

By definition, the indicators are causally related to the EOI they characterize, which makes them relevant as predictors. For instance, three of the leading indicators of rebellion were: (a) claims of discrimination made by followers (members) of an out-group, (b) low intensity military attacks on an out-group by

the Central Government (or state apparatus), and (c) number of high intensity attacks on the Central Government (or state apparatus) by out -group or vice versa. Likewise, two of the leading indicators of an Insurgency are: (a) the extent of mobilization among dissident in-group against the Central Government (or state apparatus) and (b) the extent of corruption at the highest levels of the government.

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 in Figure 4, illustrating two EOIs for Bangladesh (drawn from one of our previous papers [29]) and Thailand are shown.

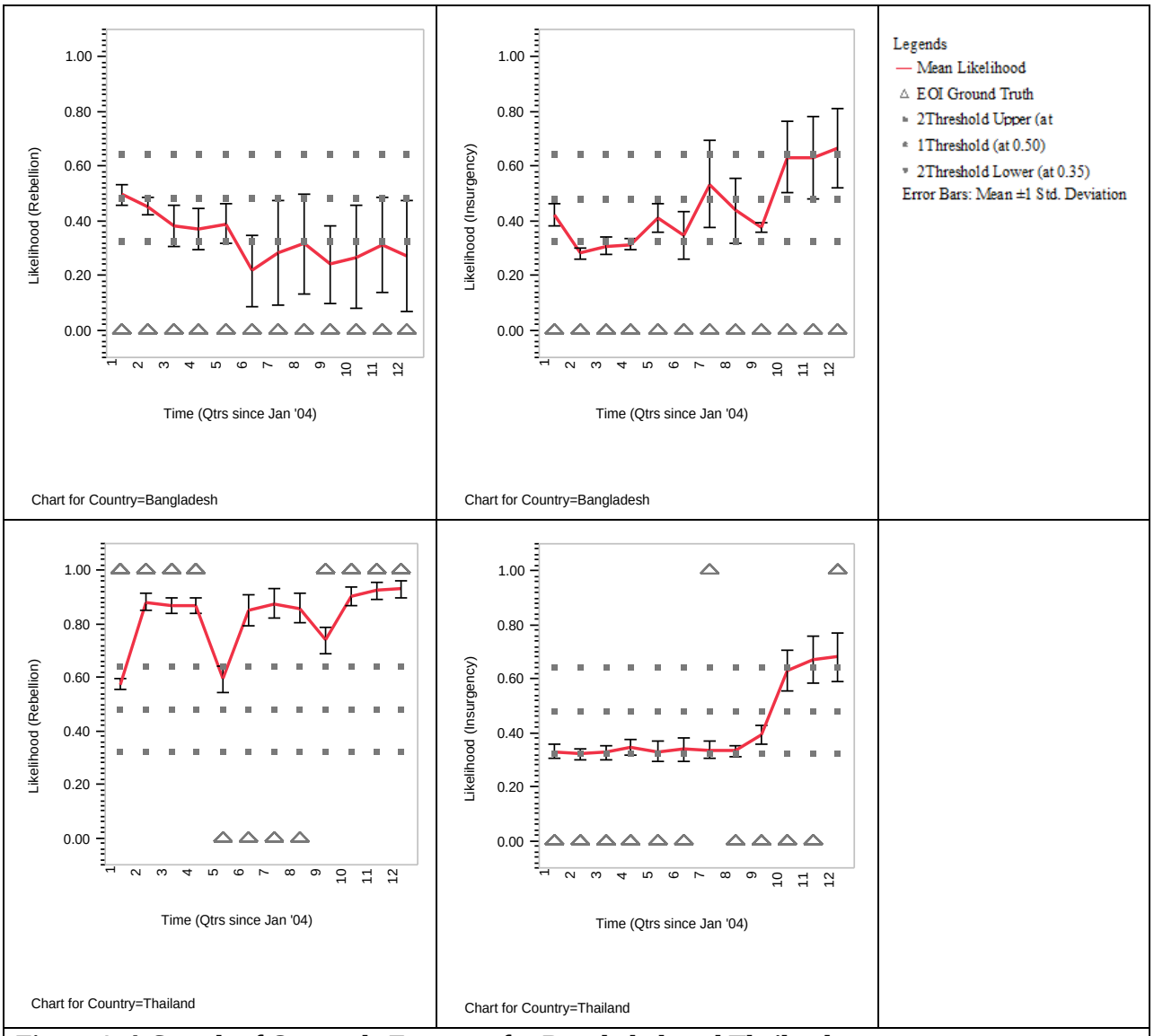


Figure 4: A Sample of Quarterly Forecasts for Bangladesh and Thailand

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 [73] [74]. For example, a framework that was developed when we developed ours is MPICE Framework [75],

which is 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 [29].

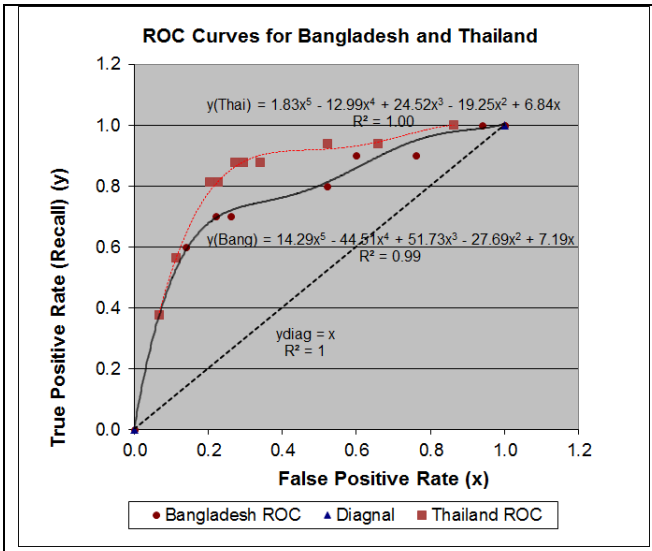


Figure 5 - Relative Operating Characteristic (ROC) Curve for StateSim (Bangladesh & Thailand)

Table 2: EOI Summary Metrics

Metric Mean– with two thresholds at 0.65-0.35	Accuracy	Precision	Recall
Bangladesh	87%	66%	81%
Thailand	89%	68%	75%

Definitions

Accuracy = (TP+TN)/(P+N) --[2a]
Precision = TP/(TP+FP) --[2b]
Recall/TP rate = TP/P --[2c]
False Positive (FP) Rate = FP/N --[2d]

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.

Table 1: Confusion Matrix

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

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.

In a complex, stochastic system (such as a real country), a range of counterfactuals (alternate futures) are possible, whereas observed, manifested values are deterministic, point estimates. In this case, our simulated outputs are likelihood estimates and are shown as a band (one standard deviation around mean) to account for counterfactuals resulting from multiple runs, while Ground Truth values are shown as binary points, showing the EOI statistics as occurring (1) or not (0). The probability/ likelihood estimates cannot be directly compared to a realized or manifested value. Yet, one requires some metrics for measuring the performance of the models. Although we generate and display the multiple futures (from multiple runs), in metrics and calculations, we only employ the mean values across alternative histories for validation. For the sake of practicality, we employ system(s) of threshold to cast the likelihood estimates from the model into a binary value, just like the Ground Truth.

We employ two systems of threshold, consisting of a single threshold line (1Threshold) and a double threshold system with upper and lower bounds (2Threshold). With the double threshold system, the likelihood estimates at or above the upper threshold are classified as 1 while those at or below the lower threshold are classified as 0. (The middle band is turned into uncertainties and ignored from further calculations. This will impose conservative assumptions on our results, as we will see later.) For the sake of simplicity, we refrain from presenting results that involve sensitivity analysis of thresholds and report results with a conservative 1/3-2/3 band of thresholds.

As can be seen in Figure 4 (left charts), there is reasonable degree of visual correlation between our predictions of rebellion and that of the Ground Truth.

The Bangladeshi government has forged a treaty agreement with CHT Tribe, once a separatist group. There is nothing in the way of separatist conflict in Bangladesh today. In consistent with this, EOI for rebellion has a very low likelihood of occurrence in both real as well as simulated outputs in Bangladesh (upper left). After a steep climb in second quarter of 2004, the likelihood of rebellion predicted by the model remains high in Thailand, except in early 2005 (see in Figure 4 lower, left). This is consistent with reality, as the separatist movement of the Muslim south which grew increasingly violent that year as the Buddhist government and police carried out suppression of protest events. This drop in likelihood of rebellion follows tsunami, which was input as an exogenous event into the model (at the end of 2004 (Dec 24, 2004), the real (and simulated) tsunami hit, especially hard in the south).

We see the likelihood of insurgency EOI (coup d'état) for Bangladesh and Thailand (see Figure 4 upper and lower right respectively) are non-zero and rising (we will, however, defer discussing insurgency till later). It should also be noted that the simulation results show increasing dispersion with time. So, the divergence from ground truth also increases with time.

When imposing a 2Threshold System (with a conservative approx. 1/3-2/3 band) to our predictions, nearly a third (30-40%) of our predictions are turned into uncertainties. In resolving uncertainties, we have ignored all cases that might be classified as uncertain and then proceeded to calculate the metrics mentioned earlier in the paper. This could be simply interpreted as limited discriminatory power of the model for some EOIs such as Inter-Group Violence and State Repression. In addition, this naturally affects our precision and recall, which are more susceptible to having smaller number of data points.

Below, we will see that based on this Ground Truth benchmark, our metrics such as precision, recall and accuracy are mostly in the range of 65-95% for multi-country, multi-year study.

In order to get a quantitative relationship between StateSim and Ground Truth forecasts, we make use of a Relative Operating Characteristic (ROC) curve (see Figure 5). 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.

In the figures (Figure 4), we display threshold values of 0.5 for single threshold system and 0.65 and 0.35 for double threshold systems. In reality, these threshold values were varied to generate ROC curves. Based on these Ground Truths and Threshold Systems, we calculated our metrics such as precision, recall and accuracy for multi-country, multi-year study.

This curve well above the diagonal shows that StateSim 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.

5.2 Quantitative 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 ensuing section, we will highlight three mini-cases (pertaining to Thailand) to demonstrate micro-level validation exercises, but not all details (e.g. sample size estimation etc.) are shown for the want of space. Mutual Entropy simply introduces the concept of correspondence while Chi-Square tests take into account of the counterfactuals.

In the following case in Figure 6, 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 three-bin classification, namely positive, neutral and negative with respect to a given target (in this case, a minority population with separatist tendencies). We started with visual correspondences (on the same scale) to give an intuitive or face validation, but they neither prove that two distributions are the same, nor give any richer picture.

Mutual Entropy: 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 (formula is common knowledge and Axtell et.al. 1996): $M(X:Y) = H(X) - H(X|Y)$ -- [3a]

where X and Y are the simulation and historic sources, respectively, and H (.) is the entropy function, defined by: $H(X) = -\Sigma\, p(x)_i \log p(x)$. -- [3b]

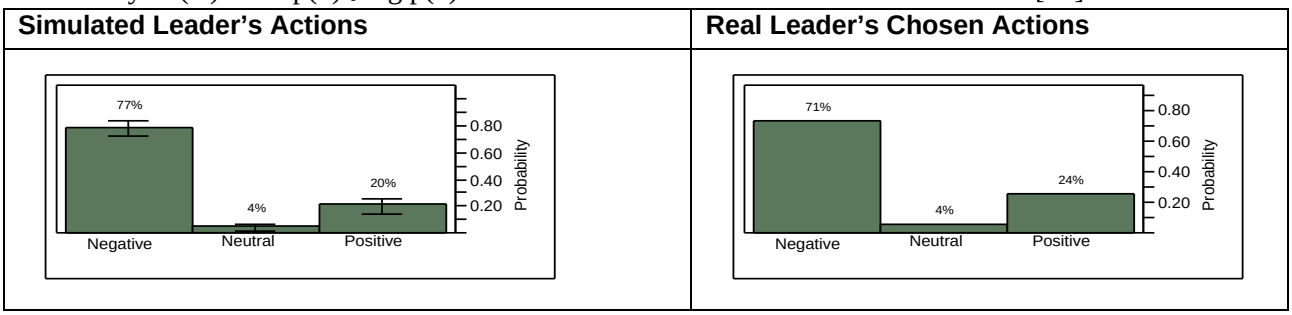


Figure 6: Correlation of Simulated Leader vs. Real Action Decisions [Comparison of distributions to see Mutual Entropy (M). Reject H0 & Accept H1 if M<<0.1]

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. [30] [29].

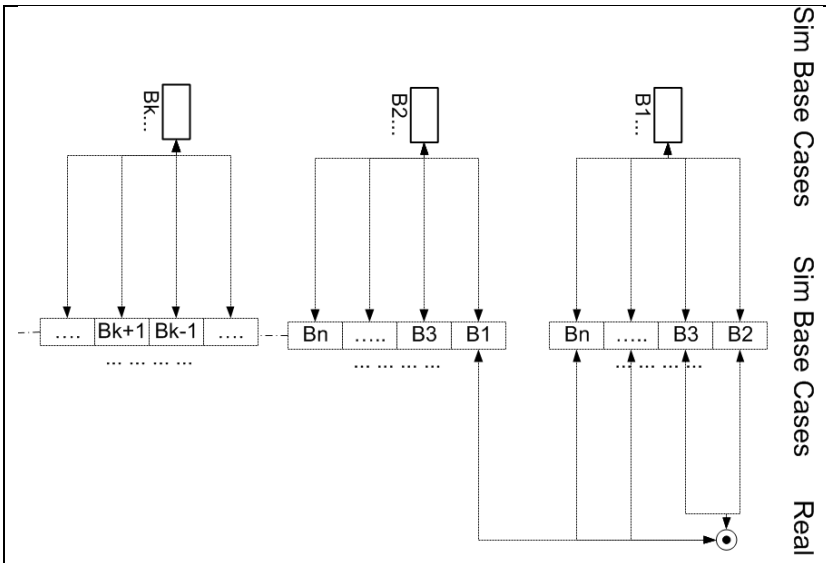


Figure 7a: Pairwise Comparison of Base Case Simulations and Real Data

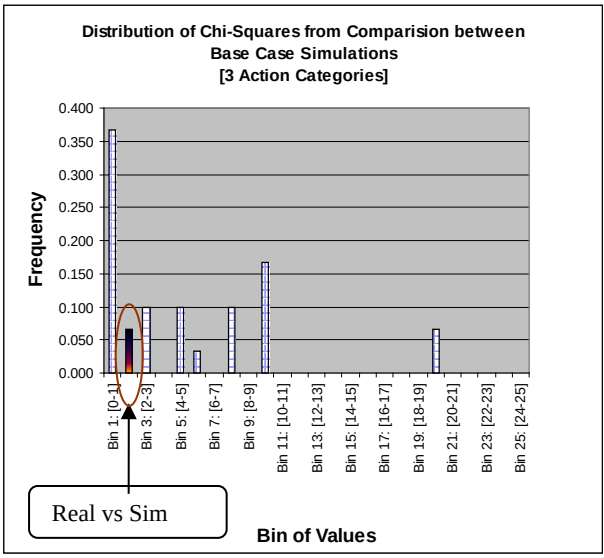


Figure 7b: Distribution of Chi-Square Values

Chi-squares, and Micro Validation: While we have multiple counterfactuals or multiple outcomes from simulations, we have the benefit of only one set of real observations. One could characterize these different outputs from model and reality, as samples obtained from a distribution. If they agreed, they would be as close as possible. Assume for simplicity, the distribution has only single mode. The single reality can be a sample obtained near the central tendency or from the tail. In order to assess this, we have designed the following tests. Since data does not follow the assumption of normality, we start with the non-parametric chi-square test, which are measures of actual divergence of the observed and expected frequencies. We will also make additional modifications.

Therefore, in order to estimate how well these different samples differ, we carried out the equivalence of "leave-one-out" cross-validation [76], as with single reality, we cannot carry out split-sample validation. We compare each of them in a pair wise fashion. We ran N simulations (N=30) of the base cases. One take the 30 simulations as alternative futures or counterfactuals emerging from the simulation model. In the simple Chi-squared calculation, we compared each simulation with the remaining 29 of them, leaving one out, and also compared the real or observed leader decisions with the 30 simulation cases. Figure 7a illustrates the process.

Simple Chi-squared: The simplistic notion is to use a Chi-square goodness fitness test, as-is. For example, we can take the chisquares of real leader decisions by testing against the distribution of decisions generated by the base case simulation. The word of caution is that chisquare is a test of rather low power with limited ability to reject the null hypothesis (not the typical skeptic's null hypothesis; instead, formulated as H0 = E(foj – fej) = 0: The data follow a specified distribution), even when the null hypothesis is patently false.

Comparing Distributions of Chi-squares: A more sophisticated (and perhaps novel) approach that we designed involved creating a plot of the distribution of pairwise chi-squares between simulated base cases, and then examining whether and where the reality versus simulated base cases could be mapped on the same plot. Here, we do not employ Chi-square test in the traditional hypothesis testing context. Instead of treating each observation-simulation data pair independently, we examine the pattern created by the map of the pairwise chi-squares (arguably we could have employed root mean square errors also in the similar fashion, except chi-squares allowed weak hypothesis testing as a bonus, but this is not shown here). We were able to ask: where the reality stands with respect to these counterfactual simulation outputs, and whether the reality is an outlier among the counterfactuals or whether it is closer to the counterfactuals.

We have described the calculation for a classification of decisions into 3-scales, namely positive, neutral and negative (action categories). There were 30 sets of simulated leader decisions from base case simulation outputs and a set of real leader decisions. Chi-square (can be found in such textbooks as [77]) was calculated between the decisions of each set of base case simulations i and 29 other base case simulations across every decisions j in the action categories of size d+1 as follows:

$$< \chi_{i,d} >^2 = \sum_{j=1}^{d+1} \frac{(E_{i,j} - O_{i,j})^2}{E_{i,j}} \qquad \text{--[Eq 4]}$$

Where Eij is the expected value (count) of the decision / action categories type j and Oij is the observed value (in this case the count for the decision of bin j). In an identical fashion, the Chi-square was also calculated between the set of real leader decisions (treat reality as ith simulation) and the base case simulations across every decisions j in the action categories of size d+1. A sample of these results for action categories of size j have been recorded in the following table as e_base_i and e_RealEvents for base case simulations and real events respectively. Please note that this table 3 gives only a sample of the output. Figure 7b displays the results visually as an integrated system.

Table 3: Sample Chi-Squared Values for Different Action Classifications and Base Cases

	Chi-Squared Value with Bins or Action Categories (AC)					
Average Chi-Squared Value taken between:	**8- AC**	**5- AC**	**4- AC**	**3- AC**	**2- AC**	**9- AC**
Simulation 1 Vs each of Simulation 2 to 30 (e_base_01)	7.318	0.530	0.519	0.460	0.423	4.218
Simulation 2 Vs each of Simulation 1 to 30 excluding Simulation 2 (e_base_02)	27.409	11.629	1.681	1.641	1.530	22.905
Simulation i Vs each Simulation 1 to 30 excluding Simulation i (e_base_i)	…	….	….	….	….	…..
Simulation 30 Vs each Simulation 1 to 29 (e_base_30)	7.318	0.530	0.519	0.460	0.423	4.218
Real Data Vs each Simulation 1 to 29 (e_RealEvents)	9.755	7.896	2.831	1.265	0.096	123.654

For example, in Table 3, e_base_02, is a simulation to simulation comparison, containing average of Simulation 2 Vs each of Simulation from 1 to 30 excluding Simulation 2. Also, e_base_02 > e_base_01. That is, compared to Simulation 1, Simulation 2 output is a relative outlier. Of all the chi-square values between base case simulations with 3-action categories (cases: e_base_01 to e_base_30 in Table 3), about 37% fall between 0 and 1. For the 3-action categories, the chi-square values between real leader decisions and base case simulation (cases: e_RealEvents) was 1.26, which has been marked on Figure 7 using dark shade. This shows that the real decisions (darkened and encircled in the following figure) are well within the distribution of the simulated leader decisions, and not very far from the most frequent chi-squares between base case simulations.

Kendal Tau: Similar analyses could be performed with other indicators or variables. For example, consider the case where we compared perceived grievances estimated by the model (Y) and its visible proxies such as conflict related fatalities and injuries (X). The time series data from real and simulated worlds were compared using Kendal Tau statistics. Kendal Tau is computed as the excess of concordant (nc) over discordant (nd) pairs, divided by a term representing the geometric mean between the number of pairs not tied on X (X0) and the number not tied on Y (Y0). Details of Kendal Tau can be found in such textbooks as [77]. As per textbook definitions, Tau (t) is: $\tau = \frac{(n_c - n_d)}{n(n-1)/2}$ --[5a]

For tied observations, however, tb is used:

$$\tau = \frac{(n_c - n_d)}{\sqrt{\left[n(n-1)/2 - \sum_{i=1}^{t} t_i * (t_i - 1)/2\right]} * \sqrt{\left[n(n-1)/2 - \sum_{i=1}^{u} u_i * (u_i - 1)/2\right]}} \quad \text{--[5b]}$$

where ti is the number of observations tied at a particular rank of x and u is the number tied at a rank of u. In the process similar to previous example (Figure 7a), we obtained two sets of Kendall tau (KT) values for numeric parameters, such as grievance, by comparing:

- parameter value (such as grievance) from each base case simulation against every other base case simulation, and
- real value of the parameter's proxy or actual parameter value and the parameter value from every base case simulations.

Table 4: Summary Table of Kendall tau

	Kendall tau b					
Variables:	**Value**	**Distribution**	**Summary Stats**	**P Value**	**Distribu-tion**	**Sum-mary Stats**
Simulation Runs Vs Real Data	-1.00 to -0.25	0%	Average: 0.24 Stdevn: 0.06	0.00 to 0.05	87%	Average: 0.04 Stdevn: 0.07
	-0.25 to 0.00	0%		0.05 to 0.10	0%	
	0.00 to 0.25	83%		0.10 to 0.20	3%	
	0.25 to 0.50	17%		0.20 to 0.25	10%	
	0.50 to 0.75	0%		0.25 to 0.35	0%	
	0.75 to 1.00	0%		0.35 to 1.00	0%	
Simulation Runs Vs Simulation Runs	-1.00 to -0.25	0%	Average: 0.62 Stdevn: 0.38	0.00 to 0.05	95%	Average: 0.01 Stdevn: 0.04
	-0.25 to 0.00	0%		0.05 to 0.10	1.5%	
	0.00 to 0.25	31%		0.10 to 0.20	1.5%	
	0.25 to 0.50	17%		0.20 to 0.25	2%	
	0.50 to 0.75	0%		0.25 to 0.35	0%	
	0.75 to 1.00	52%		0.35 to 1.00	0%	

Note: Given low frequency of data in Negative bins, they are combined to save space in the above table.

In this example, the parameter of interest is grievance. We will be making comparisons between real data (aggregated and smoothened proxy of injury and fatality) and base cases (smoothened moving average output from base case simulation). The non zero correlations are found in bins 3 (0.0-0.25) and 4 (0.25-0.50). This shows that all base cases show positive correlation with real data. A Kendall tau of 0.25 or 0.5 might appear to be a small correlation compared to a KT value of, say, 1.0, but in reality, these numbers indicate a fairly good degree of correlation, especially considering this is a time series and any mismatch would be counted as discordance resulting in negative correlation. This is illustrated by converting Kendall tau to p values, as shown in Figure 8a and 8b. Although the real data is an outlier, it is no more outlier than about half the simulation base cases themselves. This is the case with counterfactuals.

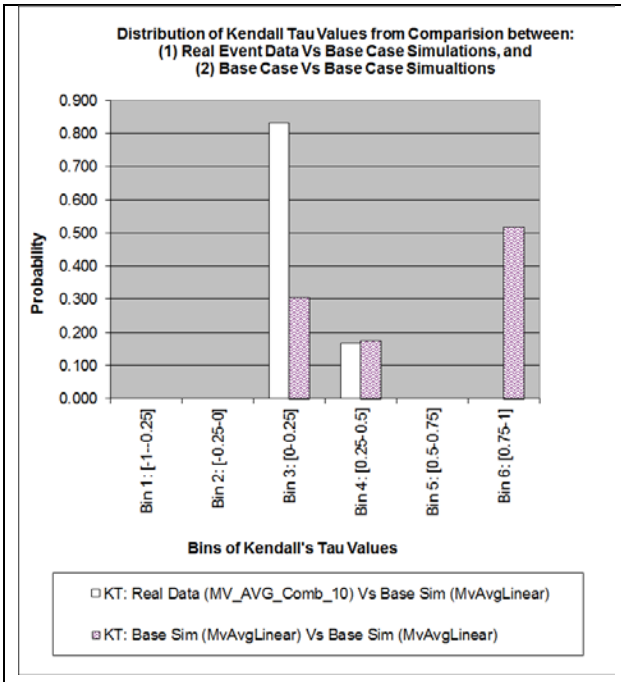


Figure 8a: Distribution of Kendall tau for Real Data and Simulation (Base Case) Grievances

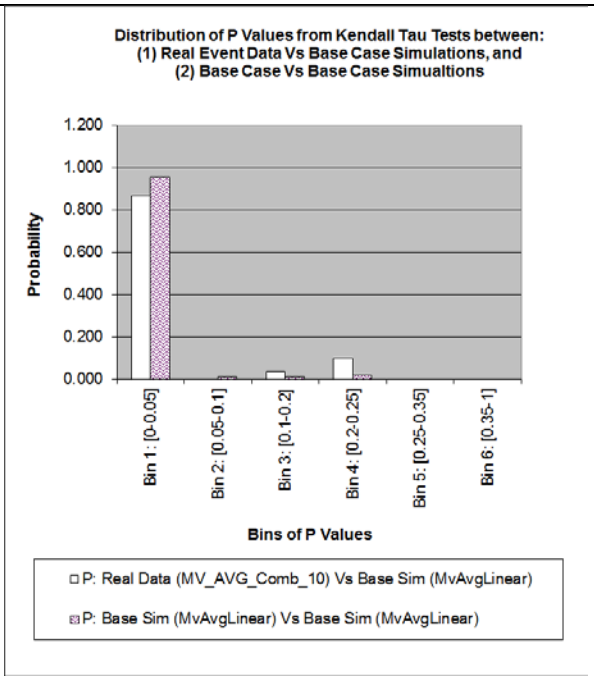


Figure 8b: Distribution of p Values for Real Data and Simulation (Base Case) Grievances

While we recognize that the p values are considered weak in the case of Kendall tau, and therefore exercise caution in the interpretations, it is hard to not notice p values for the same Kendall tau. The range of p values from the models run is in the order of 5% (80% of the data), but is also higher for remainder. Strictly statistically speaking, a null hypothesis of lack of correlation between real and simulated data cannot be rejected with 95% confidence. As an alternative to Kendall tau, we also estimated the local slopes of the real data and base case simulations, and estimated simple Pearson correlations between them. We noticed correlations of about 0.4-0.5 for moving averages of real data and base case simulation data. Arguably, these are reasonable levels of correlation for a time series. Likewise, several historical correspondence tests indicate that our models mimics decisions of the real actors/population with a correlation of approximately 70-90% (see [29] [30] [50]), which fall short of 95% confidence typically demanded in hard sciences.

In both the cases (pairwise Chi-Square and Kendal Tau), we have created a system that: takes into account of counterfactuals, does not depend on any single statistical test, provides visual comparison, and is also amenable to statistical comparisons between each pair. In these cases, a high value of correspondence indicates similarity in two sample distributions, but not necessarily the actual underlying distributions. On the other hand, a very low value of correspondence does show, to the given level of significance, that the distributions are different.

5.3 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. It will be a shame to think of complex agent based models (such as the one described earlier) as a black-box, for this misses the richness behind the model. In line with the social mechanism advocacy [51] [52] [53] [54], agent based model would be a tool to help expose the mechanism behind the models, especially addressing the question of causal equivalence through qualitative and more holistic data. Researchers, such as Griffin [78] and Reisch [79], treat narratives as analytical constructs that might help integrate otherwise disparate,

complex events and trajectories into meaningful whole. In models such as ours, narratives serve to connect the dots and weave a story.

For example, reader might recall that in predicting insurgency (*coup d'état*) (Figure 4, charts on the right), our model seems to show rising potential of insurgency for both Bangladesh and Thailand, while ground truth reports none (which is the official version of the story) for Bangladesh and two cases in Thailand. Those seems to be incorrect predictions, right?

One reason for discrepancy is that we detect likelihood (or potential) that is persistent over a longer period. According to our prediction (see Figure 4), the likelihood of Insurgency has been growing in Thailand since Thaksin became unpopular and corruption charges were leveled against the leader. Specifically, in Thailand, the likelihood of insurgency has been increasing towards the end of 2006, as indicated by the actual insurgency event manifesting in Thailand in 2007.

As for Bangladesh, which seemed outright wrong prediction, there actually was a change of government between Jan 11th -12th, 2007 (just after the quarter 12, just off the chart) following political turmoil. Political analysts[3] (for example [80]) allege that it was a coup by the military, but the military and the country fearing international repercussions (e.g. losing peace keeping role, sanctions, aid withholding), never declared as one. It seems that while the models do not get the timings of the insurgencies right, they do capture the underlying mechanism (as military coup) and provide leading indicators such as perception of escalating corruption, political crisis and declining legitimacy of the government (violent protests) that foretell impending and consequent military take over.

In the above examples, statistical validation, sticking to the rules that are set a priori, was forced to discount the performance of the models as having failed to replicate "reality". With increasing complexity and richness of models, statistical validation efforts get increasingly difficult. It does not necessarily mean validity, especially qualitative, causal and narrative validity, is compromised. If the model complexity increases due to synthesis of theories, its explanatory power and narrative richness often grows as well. This is because the model can explain the causes and effects in a more coherent manner rather than depending on a "hammer" and seeking "nails". One of the authors of this article, Silverman, identifies it as a paradox. Because there are models across the social sciences, the agents can explain their dilemmas in terms of psychology, sociology, politics, economics, and other sciences. They also can explain their situation relative to the range of organizations and networks that impact them – kinship, ethnic, ego, commercial, religious, and so on.

Some useful techniques for resolving such ambiguities and adding richness to model are tracing the results back to their origins (in this case, indicators, but could go deeper into sub-indicators or follow them temporally and causally), interrogating or drilling down into agents, developing a story from the model outputs and obtaining qualitative feedback from subject matter experts.

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 of 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

[3] For the sake of clarity, it is helpful if the reader recognizes that three types of subject matter experts contributed data for the project, namely: (1) Academic country experts/ domain experts who answered detailed surveys about the state conditions in the country for the training period; (2) The providers of official ground truth data, who extracted the data from automatic classifiers; (3) Analysts in the media (e.g. in the Economist) who commented on this particular insurgency in Bangladesh speculating, whether the change of leadership was effected by a military coup when the government of Bangladesh was changed in extra-ordinary fashion. Data in (3) did not play any role in model construction or testing, but has been shown here as an additional information.

(and page limit) of this paper. Instead, we have shown in the Appendix 2, an outline of this traceback including our attempts at creating preliminary conversational agents. One may interrogate these agents (through structured language queries) 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 [80]. These are still work in progress, but we have found that our descriptive approach to agent-based modeling, based on synthesis of models, is amenable to drill down and explanation.

Typical human face validation exercises, are either very descriptive and allow limited statistical treatment or go through checklists and offer limited description. We know that human judgment and decision making are subject to cognitive biases and errors, and it is important to recognize this and arrest any biases and errors while without curtailing the richness, efficiency and creativity that human judgment can offer.

For example, 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 seminal paper by Kahneman and Tversky [81]), experts can attempt to reject, rather than support the stories or hypotheses generated by the model. It is also possible to design the test to take the group input by employing such techniques as Delphi or Normative Group Technique (NGT) while controlling the risk of 'group think'.

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 these cases, there are no universally accepted criteria for assessing similarity, and only limited data is 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.

The evaluation could also be carried out by generating plausible narratives (and structures) from model outputs and evaluating the same. Supplementing experts' reconstruction of events is such situated methods as Heise's [82] event structure analysis (ESA), for defining the logical relations among events based on interactions among events. According to Griffin [78], ESA techniques enable "the analyst to replace temporal order with her or his 'expert judgment or knowledge' about causal connections."

Pivotal to achieving qualitative validation is extending the knowledge elicitation sessions beyond subject matter experts to include all key stakeholders. Value of including stakeholders has been demonstrated by Companion Modeling [67], an iterative participatory approach where multidisciplinary researchers and stakeholders work together continuously throughout a "four-stage cycle: field study and data analysis; role-playing games; agent-based model design and implementation; and intensive computational experiments". Our lab has been taking a similar approach to modeling, especially during the development of immersive training, serious games.

6 DISCUSSION AND CONCLUSIONS

6.1 Summary

In this paper, we have summarized some evaluation dimensions and techniques that we had employed in the past, as well as some key issues in evaluation of social system models that contain cognitively rich agents. We take a life cycle based and multi-dimensional approach to model evaluation. We have carried out conceptual evaluation including structural and ontological evaluation and model construction methodology evaluation, and external validity including qualitative validity. Methodological evaluation addresses the issue of obtaining input data while external validity tests outputs of the model against an independent set of data.

The internal evaluation 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.

In knowledge based systems, it is important to elicit knowledge from subject matter experts as well as extract knowledge from other sources such as data bases and event data. As explained, our primary inputs come from SMEs. For this, we have designed an extensive web questionnaire. We largely use country databases and web news feeds for enriching expert inputs and for background information and sanity checking what our SME survey produces.

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.

Evaluation of methodology is significant because it also increases the credibility of a complex model, by enhancing its transparency and reproducibility, and by controlling biases and errors. It makes it possible to approach the model construction process as an experiment and hypothesis testing. In cases where concepts and methodology themselves are not brought under scrutiny, complex models cannot be reliably evaluated.

This excessive parsimony and sole reliance on external validation have lead to many a model failures. Examples of this can be seen in many complex system modeling exercises, including the failure of game theory to help in asymmetric warfare, to the inability of statistical forecasting models to explain causality and mechanism, and to the kinds of overly simple quantitative risk models that helped cause the financial collapse [83]. These models are intensive on statistical or mathematically tractable model building and external validation, while ignoring conceptual, mechanism based and methodology based evaluations [84].

6.2 Wider Applicability

In this paper, we have discussed a number of techniques that we have applied in the evaluation of social systems that are primarily based on cognitively detailed agent based models.

Although we have limited ourselves to describing the techniques as applied in our modeling framework, these techniques could be applied without any modification to any modeling framework that has decision making agents operating under time steps and that have higher level abstraction mechanism such as events of interest summarizing from lower level outputs. For other systems, modifications may be required before applying these techniques.

However, it is important to keep in mind some of the limitations of the techniques. In the following sections, we present the limitations associated with the model in general as well as those associated with specific evaluation techniques that we had employed.

6.3 Limitations imposed by Modeling Approach

Firstly, we must recognize that any model, even complex models, are constructed within limited scope. Then, there are specific limitations. Having to discretize the world is a limitation that is inherent in computational modeling as a whole. Similarly, we have assumed that knowledge has a structure; this is a requirement of any model dealing with knowledge bases.

Large parametric or feature space is another limitation of our approach. In the model building process, we do depart from the prevailing paradigm of KISS and embrace KIDS, because we believe that blind adherence to a "minimalist" KISS[4] principle may result in elegant models, but not necessarily useful ones. For example, consider theoretical models, that are relatively "simple", "transparent", "mathematically tractable", focusing on a few important mechanisms and requiring a number of assumptions. Such models tend to be too generic to test and validate in actual situations [85], neglect many essential interactions and mechanisms (unsuitable for exploratory purposes) and are tautological. Besides, multiple levels and forms of correspondence from a descriptive model also enables a modeler to control for the phenomenon of equifinality.

That said, we subscribe to the view that level of descriptiveness and associated complexity of the model must be driven by the complexity of the target (real world phenomenon being studied), purpose of the model and availability of resource. A model does not have to be more complex than necessary, but many researchers believe that it takes a complex model to represent a complex system [87] [7].

Most interesting and important social system descriptions are in the nuances and that unfortunately also results in the curse of dimensionality. In psycho-social domains, compared to mechanical or physical domains and in relation to the size of the feature space, the data available also tends to be sparser. In summary, this comes down to a tradeoff decision between degree of realism and degree of mathematical tractability, and is deemed a limitation in every model. If one wants to explore this trade-off further, it is possible to construct multiple models with different level of detail or abstraction, but is beyond the scope of this paper.

6.4 Limitations of Validation Techniques

In addition to limitations of the model, which has been discussed throughout the paper, it is also appropriate to discuss the limitations of the specific techniques discussed in this paper.

Coding, Subjectivity and Model Resolution: To begin with, all tests are directly or indirectly dependent on comparing simulated and empirical data side by side. Given that both real and simulated outputs are unrelated entities except through modelers' intention, some level of coding is required to bring them both in a comparable form. Coding, however, goes hand in hand with some level of abstraction (e.g. categorizations). Coding is a necessary evil, as it helps validate a given model. Coding not only introduces subjectivity and bias, but it also undermines the possibility of distinguishing (resolution) between similar models. While such decisions as number of bins or categories will be dictated by statistical considerations such as sample size, subjective considerations are important in coding. For this reason, we have had all real data coded by external parties, including sponsors, and have applied identical coding schemes across country models for model outputs.

Also of concern along the lines of coding (and data compatibility) is comparison of real and simulated values (e.g. EOIs). We end up with a band of simulated results that must be compared to a single reality. In order to compare them, we have been using mean values. In addition, in the case of EOIs, we have been thresholding likelihood estimates from multiple runs into binary values so that they could be com-

[4] We would like to acknowledge that KISS versus KIDS is an artificial division and is about practice rather than principle. In principle, KISS allows for building complex models, if warranted. In practice, however, this principle is often invoked to justify and demand simplification, prompting the emergence of an alternate camp referred to as "KIDS". With KISS having to come to be associated with over simplification, it helps to dialogue through this dichotomy.

pared with a manifested reality expressed in binary form as whether an event occurred. This is a limitation we tried to remedy via pair-wise comparison-based tests.

In this process, a pair of thresholds, named upper and lower thresholds, were applied to continuous values of EOI likelihood to convert them into a binary value. While the thresholds themselves are subjective and were set at 1/3 and 2/3 in the Figure 4, these thresholds were systematically varied to construct the ROC curve (see Figure 5).

Since ground truth itself was subjective information and was not in the exact same form as the model prediction, alternative forms of ground truth, including the idea of having a group of subject matter experts (SMEs) estimate the likelihood were considered but not undertaken. This requires experts to forecast likelihoods of conflicts.

Eliciting SME opinion about the forecast has multiple problems well documented by researchers such as Tetlock in [68] and Green in [86]. Setting up a large expert network is not only expensive, but prediction is a task in which experts have been known to perform poorly. [Note: While experts were no better than students in the task of forecasting/ estimating likelihoods, they are able to assess and round up the necessary information pertaining to their areas of expertise.]

Projection Length: In the macro-validation section, we have been projecting out to about three years. Longer the projects, more are the opportunities for the two path dependent systems, namely model and reality to deviate from each other. In weather forecasting, for example, the forecasts are carried out at relatively short horizons (short-term) and are revised and updated frequently. However, more frequent update of results would also consume significantly more resources. Such tradeoffs are relevant in the real time applications.

Further, there is the issue mentioned earlier that real-world results are not deterministic. The uncertainty may be due to stochastic effects in the real world or due to our own limitations in understanding the system. Nor do we know what form of probability distribution the ground truth may have been "drawn" from. Thus divergence among output (and deviation from real world data) does not automatically imply lack of validity. Regardless of limitations in data and understanding, white-box models are particularly apt in these cases, as one could explore the causes of the deviation and carry out continuous improvement.

Statistical Tests and Metrics: It is difficult to know which statistical measures of goodness to rely upon. 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 provide a full picture either (they need to be combined with accuracy). Generally speaking, the ROC curve is a comprehensive measure. Yet, there are times when ROC Analysis and Precision could yield contradictory results; e.g. when there is severe class imbalances, giving majority class an unfair advantage [87]).

Mutual Entropy and Chi-Square are two other candidate metrics. Currently, however, there are no benchmarks that could indicate what would be an acceptable limit of mutual entropy for establishing correspondence. Likewise, the value of the chi-square and mutual entropy tests statistics are dependent on how the data is binned (bin definition i.e. how decisions are categorized into bins), how groups are defined as well as the sample size. For a valid chi-square approximations to apply, the chi-square test requires a sufficient sample size. The power of the test, depicting the sensitivity to departures from the null hypothesis, will not only be affected by the sample size, bin categories, but also influenced by the shape of the null and underlying distributions and the number of groups and how they are defined.

The mutual entropy and Chi-squares also depend on classifying actions into categories. More action category bins may mean being more descriptive, but it will also spread the data thin and some bins will end up with sparse data. Ideally, we want to design bins such that each bin is sufficiently filled and is equi-probable. However, that is a difficult proposition with small samples. Instead, we optimize by resorting to rules of thumb found in standard statistical textbooks. For example, rules of thumb for choosing the number of bins include requiring every bin (or more relaxed variant requiring 80% of the bins) to have at least five data points, and using a starting value of $2n^{2/5}$ [88]. We also carried out the analysis with bins varying from 2-9 (quasi-sensitivity study with respect to the number of bin) and found that 3-4 bins are stable, gives sufficient power to the test and meets the requirement of minimum of five data points. We

selected 3 bins, because it is also conceptually easier to categorize actions into three distinct categories of positive, neutral and negative with respect to target. How groups are defined will also affect the number of groups, actor-target definitions as well as have direct statistical implications as it would also influence the power of the test. In our case, the groups are defined along salient socio-political cleavages by subject matter experts. While we also have a metric called "group membership" which is dynamically calculated (the dynamic group membership is employed in estimating political capital and such), for the purpose of attributing actor-target responsibilities, our group definitions are relatively unambiguous.

The generalized Kendall tau is another candidate, yet it is known to be inaccurate when many tied values are present in the data. Besides, when the variables show fluctuations, including serrated curves, it can give a lower score, as there are less likely to be matches than mismatches. We have eliminated the local serration by using moving averages. Similarly, Kendall tau also accounts for the entire rank, not just local concordances and discordances. However, it is the local concordances and discordances that matter when validating the time series. As an alternative to Kendall's tau, we estimated the local slopes of the real data and base case simulations, and estimated simple Pearson correlations between them, and observed a correlation of about 0.4-0.5 for moving averages of real data and grievances. Arguably, this is a reasonable correlation for a time series.

Qualitative Evaluation Techniques: In the holistic paradigm that we have adopted, we discuss conceptual and narrative validities as directly contributing to overall validity. Frequently, face validation (and inspection in more structured cases) techniques are employed in these cases. It is worth noting that reductionist/ logical positivist schools of validation emphasize that external validity (correlation with real world data) is the ultimate contributor to, if not sole arbiter of, final validity. While this seems like a fundamental and philosophical difference, qualitative techniques are embedded throughout the validation process (including in statistical validation techniques) and they are most apparent in coding of data and in carrying out inspections or evaluations during conceptual and narrative validation exercises.

Qualitative techniques, including face validity, have traceability to real subjective experiences of the subject matter experts, but are subject to various cognitive biases and errors. In the methodological validation section, for example, we have proposed some techniques that might be useful in addressing these. Yet, the subjectivity of qualitative techniques still remain. However, one employs face validation in circumstances where it is not easy to find real world data or in circumstances where multitude of dimensions need to be comprehended. Having to tie down face validation exercise solely to real world data not only results in circular argument that is not easily resolvable, but also makes one miss out on the rich, multi-dimensional comprehension that humans can express. In path dependent systems with counter-factuals, simply attempting reproduce out of sample data can also be limiting to the actual validity.

Criteria: We have shown the performance of the models against reality. However, there is no universal standard for setting the level of correspondence that is acceptable. In science, statistical significance tests frequently use an arbitrarily confidence of 95%. Most of our models (as well as most social systems models) would fail, if we blindly applied such a criteria.

On the other hand, the criteria for validation should take into account of the purpose(s) of the model and the validity of the model as accepted by the user. This fitness for purpose is a significant part of the validation criteria. As exploratory models, we propose a liberal criteria of reaching 80% accuracy average at the macro level. If the models were to be used for predictive purpose (which we do not recommend for social systems), stricter criteria could be applied. Likewise, if the models were to be used in serious games, the criteria could be further relaxed. While we have proposed arbitrary criteria, it must really be set in consultation with the stakeholders (including experts in the domain).

6.5 Final Comments

The current and potential future state of the models can be explained using the 2 x2 matrix suggested by Rykiel [47] (but in turn modified from Holling 1978 , Starfield and Bleloch 1986 as reported by Rykiel [47]), where modeling problems are classified based on available data and understanding. When both data and understanding are low, only face validation of an exploratory model is possible. However, when un-

derstanding increases (but not data), then conceptual or internal evaluation is possible. Statistical validation is the domain where data is available, but understanding is limited. Conceptual, data and behavioral (operational) dimensions of validity are possible when both data and understanding are high. For most social systems models, we might be in the low data, low understanding quadrant, but some models such as those for conflict may be faring a little better than the average. In the holistic evaluation paradigm, the goal is to move towards the top right corner of high understanding and high data. However, in order to reach the top-right corner of high understanding and data, more time, effort and resources need to be invested.

In addressing some of the key concerns of the National Research Council [1], we believe that our particular approach to agent-based modeling, highlighted by our recent effort in building StateSim, has achieved a 'reasonable' level of realism and more importantly has been improving over time. The approach also has good theoretical and practical justifications.

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, as we have done in some later projects. 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: (1) internal validity ensures adherence of structure and functions to functions, form and specifications; (2) external out-of-sample validation provides statistical confidence of the model behavior; (3) methodological validity and use of domain knowledge helps reduce the dimensionality curse by structuring limited and available data; (4) extending the knowledge elicitation beyond subject matter experts to include all key stakeholders, brings in additional perspectives and provides further insight into the domain; (5) narrative validity combined with end-to-end transparency enables drilling down to see the broader narrative not just as a validation exercise, but also as a learning opportunity; (6) multi-dimensional validation attempts to control equifinality; and (7) arguably most importantly, synthesizing models and identifying gaps within and between models and commissioning social science research studies to bridge such gaps constitute a loop of iterative and continuous improvement in model quality, thereby furthering the theoretical foundations of the evolving field of social system modeling.

Likewise, in order to control the cost, we have employed model trade-offs in the forms of: (1) selecting a judicial mix of cognitively detailed, "thinking, feeling" agents to deliberate on key decisions and simpler population agents to show support or opposition to various decisions or ideas; and (2) handling multiple levels of abstractions by building models at different resolutions (e.g. Country, District, Village).

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 validation, and also for advancing the science. The need for more concerted research, to quote Bob Dylan, "is blowin' in the wind"!

7 BIBLIOGRAPHY

[1] G. Zacharias, J. MacMillan and S. Van Hemel, Behavioral Modeling and Simulation: From

Individuals to Societies, National Research Council, National Academies Press, 2008.

[2] R. Axtell, R. Axelrod, J. Epstein and M. D. Cohen, "Aligning Simulation Models: A Case Study and Results," *Computational and Mathematical Organization Theory,* vol. 1, no. 1, pp. 123-141, 1996.

[3] S. Moss and P. Davidsson, Multi-Agent-Based Simulation, vol. 1979, Springer-Verlag, 2001.

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

[5] D. Hartley, "Verification and Validation in Military Simulation," in *Proceedings of the 1997 Winter Simulation Conference*, 1997.

[6] R. R. M. Leombruni, N. Saam and M. and Sonnessa, "A common protocol for agent-based social simulation," *A common protocol for agent-based social simulation, Journal of Artificial Societies and Social Simulation,* vol. 9, no. 1, 2006.

[7] G. Fagiolo, A. Moneta and P. Windrum, "FagiolA Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems," *Computational Economics,* 2007.

[8] D. Midgley, R. Marks and D. Kunchamwar, "The building and assurance of agent-based models: an example and challenge to the field," *Journal of Business Research,* vol. 60, p. 884−893, 2007.

[9] K. M. Carley and L. Gasser, "Computational and Organization Theory," in *Mul-tiagent Systems - Modern Approach to Distributed Artificial Intelligence*, Boston, MIT Press, 1999, pp. 299-330.

[10] K. Gluck, P. Bello and J. Busemeyer, "Introduction to the Special Issue [on model comparison]," *Cognitive Science,* vol. 32, no. 8, pp. 1245-1424, 2008.

[11] D. Hales, J. Rouchier and B. Edmonds, "Model-to-Model Analysis," *Journal of Artificial Societies and Social Simulation,* vol. 6, no. 4, p. 5, 2003.

[12] D. Schreiber, "Validating agent-based models: From metaphysics to applications," in *Annual Confer-ence of the Midwestern Political Science Association*, Chicago, IL, 2002.

[13] R. Harré, The Principles of Scientific Thinking, London: Macmillan, 1970.

[14] N. Gilbert, Agent-Based Models (Quantitative Applications in the Social Sciences), Sage Publications, 2007.

[15] N. Gilbert, "Open problems in using agent-based models in industrial and labor dynamics," in *World Scientific*, 2004, pp. 401-405.

[16] K. Troitzsch, "Validating simulation models," in *18th European Simulation Multiconference*, 2004.

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

[18] U. Wilensky and W. Rand, "Making Models Match: Replicating an Agent-Based Model," *Journal of Artificial Societies and Social Simulation,* vol. 10, no. 4, p. 2, 2007.

[19] J. Gratch and S. Marsella, "A Domain-independent Framework for Modeling Emotion," *Journal of Cognitive Systems Research,* vol. 5, no. 4, pp. 269-306, 2004.

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

[21] S. Moss and B. Edmonds, "Sociology and Simulation: Statistical and Qualitative Cross-Validation," *American Journal of Sociology,* vol. 110, no. 4, pp. 1095-1131, 2005.

[22] R. Shannon, Systems Simulation: The Art and Science, Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[23] O. Balci, "Verification, Validation, and Testing," in *The Handbook of Simulation*, New York, NY,

John Wiley & Sons, 1998, pp. 335-393.

[24] M. Petty and W. Weisel, "A Composability Lexicon," in *Proceedings of the Spring Simulation Interop-erability Workshop*, Orlando, 2003.

[25] M. a. S. C. O. MSCO, "Verification, Validation, and Accreditation (VV&A)," 26 September 2006. [Online]. Available: http://vva.msco.mil. [Accessed 10 June 2010].

[26] R. Axelrod, The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration, Princeton, NJ: Princeton University Press, 1997.

[27] B. Edmonds and S. Moss, "From KISS to KIDS - an 'antisimplistic' modelling approach," in *Multi Agent Based Simulation*, Springer, 2004, pp. 130-144.

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

[29] B. Silverman, G. Bharathy, B. Nye, G. Kim, Roddy and M. Poe, "Simulating State and Sub-State Actors with StateSim: Synthesizing Theories across the Social Sciences," in *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*, Wiley STM, 2010.

[30] B. Silverman, G. Bharathy, T. Smith, R. Eidelson and M. Johns, "Socio-Cultural Games for Training and Analysis: The Evolution of Dangerous Ideas," *IEEE Systems, Man and Cybernetics*, vol. 37, no. 6, pp. 1113-1130, 2007.

[31] B. Silverman, G. Bharathy, B. Nye and T. Smith, "Modeling Factions for 'Effects Based Operations': Part II – Behavioral Game Theory," *Journal of Computational and Mathematical Organization Theory*, vol. 14, no. 2, pp. 120-155, 2008.

[32] I. Lustick, B. Alcorn, M. Garces and A. Ruvinsky, "From Theory to Simulation," in *Annual Meeting of the American Political Science Association*, Washington, D.C., 2010 (September).

[33] B. Silverman, G. Bharathy and G. Kim, "The New Frontier of Agent-Based Modeling and Simulation of Social Systems with Country Databases, Newsfeeds, and Expert Surveys," in *Agents, Simulation and Applications*, Taylor and Francis, 2009.

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

[35] Y. Barlas and S. Carpenter, "Philosophical Roots of Model Validation: Two Paradigms," *System Dynamics Review*, vol. 6, no. 2, pp. 148-166, 1990.

[36] J. W. Forrester and P. Senge, "Tests for Building Confidence in System Dynamics Models," in *System Dynamics*, Amsterdam, North-Holland, 1980.

[37] J. W. Forrester, Industrial Dynamics, Portland, OR: Productivity Press, 1961.

[38] J. Forrester, "A Response to Ansoff and Slevin," *Management Science*, vol. 14, pp. 601-618., 1968.

[39] O. Balci, "Validation, Verification and Testing Techniques Throughout the Life Cycle of a Simulation Study," *Annals of Operations Research*, p. 53, 1994.

[40] Y. Barlas, "Formal aspects of model validity and validation in system dynamics," *System Dynamics Review*, vol. 12, no. 3, pp. 183-210, 1996.

[41] P. L. Knepell and D. Arangno, Simulation validation: A confidence assessment methodology, Los Alamitos, California: IEEE Computer Society Press, 1993.

[42] R. Sargent, "Verifying and validating simulation models," *Proceedings of the 1996 Winter Simulation Conference*, pp. 55-64, 1996.

[43] G. Yücel and E. van Daalen, "An Objective-Based Perspective on Assessment of Model-Supported Policy Processes," *Journal of Artificial Societies and Social Simulation*, vol. 12, no. 4, p. 3, 2009.

[44] B. Zeigler, Theory of modelling and simulation, New York: Wiley, 1976.

[45] K. Takadama, T. Kawai and Y. Koyama, "Micro- and macro-level validation in agent-based

simulation: re-production of human-like behavior and thinking in a sequential bargaining game," *Journal of Artificial Societies and Social Simulation,* vol. 11, no. 2, p. 9, 2008.

[46] J. Sterman, Business dynamics: systems thinking and modeling for a complex world, Boston: Irwin/McGraw-Hill, 2000.

[47] E. Rykiel Jr., "Testing ecological models: the meaning of validation," *Ecological Modeling,* vol. 90, no. 3, pp. 229-244, 1996.

[48] N. Oreskes, "Evaluation (Not Validation) of Quantitative Models," *Environ Health Perspect,* vol. 106, no. 6 (Supp), pp. 1453-1460, 1998.

[49] S. Banks, "Exploratory Modelling for Policy Analysis," *Operations Research,* vol. 41, no. 3, 1993.

[50] B. Silverman, G. Bharathy and B. Nye, "Gaming and Simulating Sub-National Conflicts," in *Computational Methods for Counter-Terrorism*, Berlin, Hiedelberg and New York, Springer-Verlag, 2009.

[51] P. Hedstrom and R. Swedberg, "Social Mechanisms: An Introductory Essay," in *Social Mechanisms: An Analytical Approach to Social Theory*, P. Hedstrom and R. Swedberg, Eds., Cambridge, Cambridge University Press, 1998, pp. 172-203.

[52] J. Mahoney, "Beyond correlational analysis: Recent innovations in theory and method," *Sociological Forum,* vol. 16, no. 3, pp. 575-93, 2001.

[53] M. Bunge, "Mechanisms and explanation," *Philosophy of the Social Sciences,* vol. 27, no. 4, pp. 410-65, 1997.

[54] R. Boudon, "Social Mechanisms without Black Boxes," in *Social Mechanisms: An Analytical Approach to Social Theory*, Cambridge, Cambridge University Press, 1998, pp. 172-203.

[55] R. Mayntz, "Mechanisms in the Analysis of Social Macro-Phenomena," *Philosophy of the Social Sciences,* vol. 34, pp. 237-254, 2004.

[56] C. Szabo and Y. Teo, "An Approach for Validation of Semantic Composability in Simulation Models," in *Proceedings of the 23rd ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation*, Lake Placid, NY, USA, 2009.

[57] S. Bankes and J. Gillogly, "Validation of Exploratory Modelling," in *Proceedings of the Conference on High Performance Computing*, San Diego, CA, 1994.

[58] R. Ackoff, Ackoff's Best, New York: John Wiley, 1999.

[59] P. B. Checkland and J. (. Scholes, Soft Systems Methodology in Action, John Wiley & Sons, 1990.

[60] M. C. Jackson, "Critical Systems Thinking and Practice," *European Journal of Operational Research,* vol. 128, pp. 233-244, 2001.

[61] G. Midgley, Systemic Intervention: Philosophy, Methodology, and Practice, NY: Kluwer Acad, 2000.

[62] P. Reason, Human Inquiry in Action: Developments in New Paradigm Research, London: Sage, 1988.

[63] J. Heron, Co-operative Inquiry: Research into the human condition, London: Sage, 1996.

[64] R. McTaggart, "Principles for Participatory Action Research," *Adult Education Quarterly,* vol. 41, no. 3, pp. 168-187, 1991.

[65] B. Silverman and G. Bharathy, "Modeling the Personality & Cognition of Leaders," in *14th Conference on Behavioral Representations in Modeling and Simulation*, 2005.

[66] A. Geller and S. Moss, "Growing Qawm: An Evidence-Driven Declarative Model of Afghan Power Structures," *Advances in Complex Systems,* vol. 11, no. 2, pp. 321-335, 2008.

[67] O. Barreteau and Others, "Our companion modelling approach," *Journal of Artificial Societies and Social Simulation,* vol. 6, no. 1, 2003.

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

[69] S. Leye, J. Himmelspach and A. Uhrmacher, "A Discussion on Experimental Model Validation," in *UKSim 2009: 11th International Conference on Computer Modelling and Simulation*, 2009.

[70] R. Ewald, J. Himmelspach, M. Jeschke, S. Leye and A. Uhrmacher, "Flexible experimentation in the modeling and simulation framework JAMES II—implications for computational systems biology," *Briefings in Bioinformatics,* vol. 11, no. 3, pp. 290-300, 2010.

[71] M. Niazi, A. Hussain and M. Kolberg, "Verification & Validation of Agent Based Simulations using the VOMAS (Virtual Overlay Multi-agent System) Approach," in *MAS&S 09 at Multi-Agent Logics, Languages, and Organisations Federated Workshops*, Torino, Italy, 2009.

[72] S. Rybacki, J. Himmelspach, E. Seib and A. Uhrmacher, "Using workflows in M&S software," in *Proceedings of the 2010 Winter Simulation Conference*, 2010.

[73] J. Covey, M. Dziedzic and L. Hawley, "The Quest for Viable Peace: International Intervention and Strategies for Conflict Transformation," 2005. [Online]. Available: http://bookstore.usip.org/books/BookDetail.aspx?productID=120589. [Accessed 23 10 2008].

[74] P. H. Baker, "Conflict Resolution: A Methodology for Assessing Internal Collapse and Recovery," in *Armed Conflict in Africa*, Triangle Institute for Strategic Studies, Lanham, MD and Oxford: The Scarecrow Press, 2003.

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

[76] C. Goutte, "Note on free lunches and cross-validation," *Neural Computation,* vol. 9, pp. 1211-1215, 1997.

[77] M. Hollander and D. A. Wolfe, Non-parametric statistical methods, New York: Wiley, 1999.

[78] L. Griffin, "Narrative, Event-Structure Analysis, and Causal Interpretation in Historical Sociology," *American Journal of Sociology,* vol. 98, no. 5, pp. 1094-1133, 1993.

[79] G. Reisch, "Chaos, History and Narrative," *History and Theory,* vol. 30, pp. 1-20, 1991.

[80] B. Silverman, "Systems social science: a design inquiry approach for stabilization and reconstruction of social systems," *Journal of Intelligent Decision Technologies,* vol. 4, no. 1, pp. 51-74, 2010.

[81] D. Kahneman and A. Tversky, "Choices, Values and Frames," *American Psychologist,* vol. 39, no. 4, p. 341, 1984.

[82] D. Heise, "Modeling Event Structures," *Journal of Mathematical Sociology,* vol. 14, pp. 139-169, 1989.

[83] F. Salmon, "Recipe for Disaster: The Formula that Killed Wall Street," *Wired,* vol. 17, no. 3, 23 February 2009.

[84] D. H. Freedman, "Why Economic Models Are Always Wrong," *Scientific American,* 26 October 2011.

[85] E. Van Nes and M. Scheffer, "A strategy to improve the contribution of complex simulation models to ecological theory," *Ecological Modelling,* vol. 185, pp. 153-164, 2005.

[86] K. C. Green, "Further evidence on game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts," *Green, K. C. (2005). Further evidence on game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts. International Journal of Forecasting,* vol. 21, p. 463 – 472, 2005.

[87] N. Japkowicz, "Classifier Evaluation: A Need for Better Education and Restructuring," in *ICML-*

2008 Workshop on Evaluation Methods for Machine Learning II, Helsinki, Finland, 2008.

[88] E. S. Keeping, Introduction to Statistical Inference, NY: Dover, 1995.

[89] G. Bharathy, "Designing Events of Interest and Indicators for Country Stability," Unpublished, Philadelphia, 2008.

[90] L. Von Bertalanffy, General System Theory: Foundations, Development, Applications, NY: George Braziller, 1969.

[91] K. Troitzsch, U. Mueller, G. Gilbert and J. Doran, Social Science Microsimulation, Berlin: Springer, 1996.

[92] J. Sterman, "Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models," *Dynamica*, vol. 10, no. 2, pp. 51-66, 1984.

[93] B. Silverman, "Barry Silverman's Website," 2010. [Online]. Available: http://www.seas.upenn.edu/~barryg/HBMR.html. [Accessed 5 10 2010].

[94] R. Sargent, "A tutorial on verification and validation of simulation models," *Proceedings of the 1984 Winter Simulation Conference*, Vols. 84CH2098-2, pp. 115-122, 1984.

[95] A. Murphy, "What is a good forecast? An essay on the nature of goodness in weather forecasting," *Weather Forecasting*, vol. 8, pp. 281-293, 1993.

[96] J. Logan, "In defence of big ugly models," *Transactions of the American Entomological Society*, vol. 40, pp. 202-207, 1994.

[97] J. W. Forrester and D. Freedman, "Why Economic Models Are Always Wrong Scientific American," p. 2011, 26 October 2011.

[98] G. Deffaunt, G. Weisbuch, F. Amblard and T. Faure, "Simple is Beautiful ... and Necessary," *Journal of Artificial Societies and Social Simulation*, vol. 6, no. 1, 2003.

[99] J. Banks, C. J. S., N. B. L. and D. Nicol, Discrete-event system simulation, Upper Saddle River, New Jersey: Prentice-Hall, 2000.

[100] Economist, "The coup that dare not speak its name," *Economist*, 18 January 2007.

[101] WCRP, "Forecast Verification - Issues, Methods and FAQ, World Climate Research Programme," 2009. [Online]. Available: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html. [Accessed 22 June 2010].

[102] G. Bharathy and B. Silverman, "Applications of Social Systems Modeling in Political Risk Management," in *Handbook on Decision Making*, vol. ISRL 33, Berlin Heidelberg, Springer-Verlag, 2012, p. 331–371.

[103] G. Bharathy, L. Yilmaz and A. Tolk, "Agent-directed Simulation for Combat Modeling and Distributed Simulation," in *Engineering Principles of Combat Modeling and Distributed Simulation*, John Wiley & Sons, 2012, pp. 669-713.

[104] G. Bharathy and B. Silverman, "Validating Agent Based Social Systems Models," in *Proceedings of 2010 Winter Simulation Conference*, Baltimore, MD, 2010.

ACKNOWLEDGMENTS

This research was supported by Becks Fellowship, various government agencies, including advanced research branches of DoD (including DARPA's ICEWS program), a COCOM, and the US Army. This document has been approved for public release. Distribution unlimited.

The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors alone are responsible for any opinions, unsubstantiated claims, or errors made in this manuscript.

Several past and present members of the Ackoff Collaboratory for Advancement of Systems Approach (ACASA) at the University of Pennsylvania contributed to development of software framework described in this paper. They, in the chronological order of getting involved, are: Barry Silverman, Kevin O'Brien, Jason Cornwell, Michael Johnes, Ransom Weaver, Gnana Bharathy, Ben Nye, Evan Sandhaus, Mark Roddy, Kevin Knight, Aline Normoyle, Mjumbe Poe, Deepthi Chandasekeran, Nathan Weyer, Dave Pietricola, Ceyhun Eksin and Jeff Kim.

AUTHOR BIOGRAPHIES

GNANA K BHARATHY is a researcher and project manager at ACASA, UPenn, passionately interested in social, environmental and socio-technical systems. Particularly, Gnana studies complex system problems in business and society through modeling, simulation and analysis, employing both quantitative and qualitative techniques. Other areas of his work include risk management, analytic, and systems science and systems thinking.

BARRY G SILVERMAN is a Professor of Systems Sciences and Engineering and Director of AckoffCollaboratory 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.

Appendix 1: Summary of FAIREST Factors

Profiling the parameters includes the following social theory parameters and models:

Factions of the region:

- Philosophy, Sense of Superiority, Distrust, Perceived Injustices/Transgressions
- Leadership, Membership, Other Roles
- Relationship to other groups (ingroups, outgroups, alliances, atonements, etc.)
- Barriers to exit and entry (saliences)
- Institutional infrastructures owned/controlled by the group
- Access to institutional benefits for the group members (Level Available to Group)
- Fiscal, Monetary and Consumption Philosophy
- Various Identities that are important to these Factions

Agents (Decision Making Individual Actors) that fill the roles (leaders, followers, ministers, etc):

- Value System, also known as Goal-Standard-Preference (GSP) Tree: Hierarchically organized values such as short term goals, long term preferences and likes, and standards of behavior including sacred values and cultural norms,
- Ethno-Linguistic-Religious-Economic/Professional Identities
- Level of Education, Level of Health, Physiologic/Stress Levels
- Level of Wealth, Savings Rate, Contribution Rate
- Extent of Influence/ Authority over each Group, Degree of Membership in each Group
- Personality and Cultural Factor sets (conformity, assertiveness, humanitarianism, etc.)
- Population Model: Geographic Prevalence and Influence of Each Faction in region, Geographic distribution of resources and other key factors, and the Local factors that affect transmission of Influence and Identities in a region

Institutions available to Each Group: (Public Works, Protections, Health/Education, Elections, etc.)
- □ Capital Investment, Capacity for Service, # of Jobs
- □ Effectiveness, Level of Service Output, Level of Corruption, Group Influence
- □ Costs of Operation, Depreciation/Damage / Decay

Resources:
- □ Group Level Resources such as Political, Economic and Security Strengths
- □ Disparity, Resource levels, Assets Owned/Controlled

Economy Model (Dual Sector - LRF Model)
- □ Formal Capital Economy (Solow Growth Model)
- □ Undeclared/Black Market (Harrod-Domar Model)

Supra-System
- □ Political Model (loyalty, membership, voting, mobilization, etc.)
- □ Information Propagation/Votes/Small World Theory
- □ External threats

Time Periods (How will the FAIREST items shift across near, mid, and long term under current conditions and with new designs. If the model runs do not produce satisfying answers, new questions arise and must be answered).

Appendix 2: Drilldown into the Model

End to End Panel

Once the country model is assembled, it produces a view as Figure A2.1 shows. This Figure is scrolled down to the Southern Thai provinces and it shows the SME created 3 groups in that region (other groups hidden). Each has a leader agent and archetypical core and fringe follower agents. The parenthetic numbers next to each of the follower agents represents the relative size in millions of that subgroup. On the right side of the screen are panels that show the sense of membership and belonging that internal and external members feel toward each group. These estimates are model-derived and they shift dynamically as the leaders take actions in a run.

Figure A2.1 – A Country Viewer Showing the Southern Portion of Thailand (Agent Interrogation below)

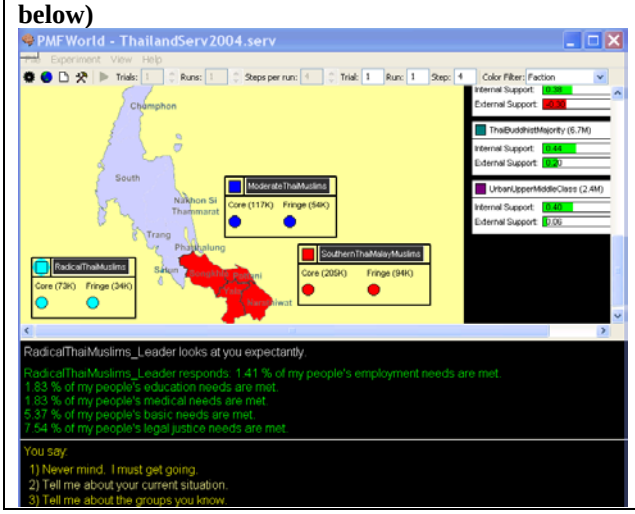
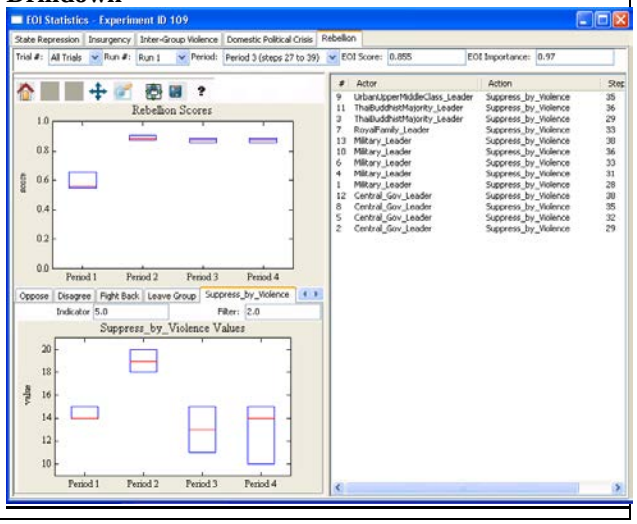


Figure A2.2 – Backend EOI Viewer and Outcomes Drilldown



The top of the viewer shows a toolbar that includes the end to end tools. First is an icon that opens the questionnaire in which one can update any of the model parameters. Second is a world viewer that permits one to open tables that show group resources, institutions, membership, relations, and other properties. Basically, these tables let one browse through the properties of the SME's country model. The next icon (hammer and wrench) opens the experiment designer. This finds every parameter of the SME's model and shows them in hierarchical fashion with their default setting. This tool treats the default as a mean and allows the user to specify probability distributions for any number of the parameters and it elicits the attributes of those PDFs (e.g, mean, standard deviation, etc.). It then computes the number of runs needed for statistically meaningful experiments on each permutation of the parameters being experimented over. The remainder of the toolbar of Figure A2.1 provides some tape recorder style buttons for starting a run and playing or pausing it to see the actions unfolding as you watch. In the body of the map are various icons that signify when one group is launching an attack (!) and/or being a target of an attack (X).

When the runs finish and the user wishes to see the logged results, an EOI viewer may be opened. To explain this viewer we show it in Figure A2.2 with an EOI hierarchy plotted to the left of it. Specifically, this run was setup to track several EOIs important to instability including: insurgency, rebellion, domestic political crisis, and inter-group violence.

Drilling Down in to EOIs

As the viewer shows in A2.2, each EOI is given a tab at the top of the screen, and the Rebellion tab is currently selected. Immediately beneath the EOI tab are some parameters showing the trial and run being displayed and the overall score of that EOI. As the left side diagram shows, the EOIs are calculated from a set of indicators that in turn count up events of that type that were produced in a given month and quarter of the simulation year.

The viewer has 3 main windows that allow one to drill through this layering. Thus the rebellion scores by quarter are shown in the top left, while the many indicators that combine into the rebellion EOI are shown as tabs in the bottom left window. There, the tab labeled "disagree" is shown. This indicator counts up the number of agents that disagree with the government (or rebels or other groups as the indicated by target). Disagree is a step less severe than oppose which is a step before fight back. A document on the many indicators tallied up into each EOI is available as [89]. This document shows that dozens of indicators are combined from all three layers of the StateSim multi-resolution architecture (FactionSim entities, PMFserv agents, and PSI agents). It also shows how they are weighted and tuned.

Interrogating Agents Example

As mentioned earlier, once the user drills down to events of interest in the righthand window of Figure A2.2, they can then locate the relevant agents on the Country Viewer who caused those events and interrogate them as to what they were thinking that lead them to take those actions. As one example of this, the base of earlier Figure A2.1 shows a dialog with the Radical Muslim Leader agent. There the user is asking "tell me about your current situation" and the leader responds with a summary of the conditions and services denied his followers. Many other types of interactive queries and answers are possible to learn the situation that the agents face and what is motivating the various agents toward courses of action in the simulated world.

Tracing Example

From a totally different perspective, one can also trace what is happening to the overall EOIs across time periods and countries. For example, Figure 4 (earlier in the paper) presented the summary EOI scores over a three year period. Now, we would like to use these runs to drill into to some of those EOIs to understand them more fully. Let us examine a rare event, that of insurgency or takeover of the government by another power. Two of our country forecast years do show takeover occurring – Bangladesh and Thailand (see earlier Figure 4). In this section we want to drill in a bit and see what the EOI scores actually look like.

On the top-left quadrant of Figure 4, we see the insurgency EOI for Bangladesh is non-zero in all four quarters and above the cutoff threshold in the 1st quarter. In Thailand (bottom-left quadrant of Figure 4), the EOI starts out moderate but JUST below our cutoff, and then leaps up to well above the cutoff for the last 3 quarters. As discussed in the previous section, these forecasts are correct and in fact the military in both countries suspended the constitution and took power away from corrupt leaders. Both countries' militaries are fairly weak and cannot keep power for long. By 2008 and 2009 respectively, election processes were re-constituted in Thailand and in Bangladesh. It is a different matter that the democracy would remain very fragile in these countries.

Drilling down one layer deeper into why the StateSim for Thailand forecast insurgency in 2006 and rebellion in 2004, we can examine the indicators of the EOI scores as shown in Figure A2.4. Specifically, the reader will recall from the discussion earlier, that EOI scores are computed by combining together a number of indicator variables which in turn count up the kinds of decisions, events and states that occurred in the simulator. Actions and events coming out of the simulated world are hierarchically organized and summarized into Indicators which in turn are used to compute EOIs. In reverse, one can drill down from EOI to Indicator to actions that are being counted. By definition, the indicators are causally related to the EOI they characterize, which makes them relevant as predictors. Specifically, the left side of Figure A2.3 shows that the insurgency EOI indicators while the right side shows indicators for rebellion. We can observe substantial numbers of highly elevated indicators.

For instance, three of the leading indicators of rebellion shown were: (a) claims of discrimination made by followers (members) of an out-group, (b) low intensity military attacks on an out-group by the Central Government (or state apparatus) (suppress by violence), and (c) number of high intensity attacks on the Central Government (or state apparatus) by out -group or vice versa (military attacks), and (d) overall disregard for life, limb and welfare of the members of the out-group. We can also observe such coincident indicators as military attacks and many angry agents opposing and fighting back.

Likewise, two of the leading indicators of an Insurgency are: (a) the extent of mobilization among dissident in-group against the Central Government (or state apparatus) and (b) the extent of corruption at the highest levels of the government. Likewise, indicators such as Take Over action being performed is a coincident indicators pertaining to Insurgency. We see on the left that Insurgency may be explained by noting that these indicators are substantially above zero and one is below. This is a logarithmic scale, so the elevated levels are quite high. This configuration also shows high degrees of disagreeing and angry agents wanting to leave the group, favorable exceeds unfavorable group strength, and VID (vulnerability, distrust, and injustice) is negative. As a result we see that Take Over is non-zero which means that some agent or agents are undertaking this action. In fact it is military take over events that occur Thailand in the very next quarter (and covertly in Bangladesh a few quarters later). In general, we found it useful to use about 2 dozen indicators per EOI to adequately count and track events coming from the simulated world.

Figure A2.3 – Indicators Leading to the EOIs for Thailand: Insurgency 2006 (2Q) and Rebellion 2004 (2Q)