Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

LIMITS OF EMPIRICAL VALIDATION: A REVIEW OF ARGUMENTS WITH RESPECT TO SOCIAL SIMULATION

Marko A. Hofmann

ITIS
University of the Federal Armed Forces Munich
Werner-Heisenberg-Weg 39, Neubiberg 85577, Germany

ABSTRACT

Output comparison between simulation model and real world reference system is commonly regarded to be the acid test of model credibility. As sound as the comparison-based approach may seem, serious epistemological and methodological qualifications have been made concerning the foundations of the concept, its applicability, and its dependence from the chosen philosophical perspective. The article reviews and reassesses technical and philosophical arguments on the limits of empirical validation with respect to social simulation. The paper is intended to reposition empirical validation for social simulations that are theory-free and non-predictive. The proposed shift is inspired by the recent critical reassessment of significance tests in applied statistics. According to this shift, it is transparency which becomes paramount for the single social simulation project, whereas empirical validation on the macro level is crucial only after meta-analysis of rival simulation models has shown robust findings despite different sets of assumptions.

1 INTRODUCTION

In order to ensure that simulations are problem adequate representations of real-world systems, validation is a constant major concern of the simulation community (see Helmer and Rescher (1959), Hermann (1967), van Horn (1971), Landry and Oral (1993), Kleindorfer, O'Neill, and Ganeshan (1998), Pace (2004), Bair and Tolk (2013) for an 50 year-overview). "Validity" is generally thought of as the degree to which a model faithfully represents its system counterpart with respect to a special modeling purpose. The most straightforward approach for validation is surely to compare input/output samples from the reference system and the the model. Simulation-specific formal frameworks for such a comparison have been elaborated, for example, from Zeigler, Kim, and Praehofer (2000) (*experimental frames*) and Olsen and Raunak (2013) (*simulation validation coverage*). One might ask whether there is a difference in the validation of mathematical models, in general, and simulation models, in particular. From an epistemological perspective the major distinction concerns simulation models that are not based on explicit theories of the macro phenomenon of interest, but on its imitation. Whereas simulation in the natural sciences is almost exclusively based on theoretical findings (often differential equations) which are "unfolded" by simulation (weather simulation, astrophysical simulation etc.), most social simulations are founded on some assumptions of individual human behavior (micro level) which are used to generate the macro phenomenon of interest. The major difference with respect to validation is, that the latter simulation models cannot rely on the corroboration of theory (thermodynamics, quantum theory, relativity etc.) in other applications (since they are not based on such theories). In order to have any scientific value, theory-free simulations have to be validated using empirical data (directly) on the macro level. It is even not necessary that all the assumption on the micro level are correct. What matters is only the correspondence of simulation output and empirical evidence (The value of such counter-factual models for prediction is well-established in economics (Friedman 1953, Musgrave 1981, Maki 2000)). This is completely contrary to a model of,

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

for example, a supernovae which is based on the already corroborated laws of physics, and for which we do not have a currently observable example (micro level validated, macro level unobservable).

After reflecting arguments voiced against empirical validation, in general, this article argues that for theory-free, non-predictive social simulations (which includes the majority of agent based social simulations) it is generally impossible to judge their value independently from other (simulation) models. The most important argument for this restrictions is the inevitable subjectivity of any experimental design in social research. An alternative for the focus on validation within a single simulation project is a meta-analytic approach, similar to the recent shift in applied statistics: The value of a single study is always limited, knowledge is established on the basis of independent replication.

2 LIMITS OF EMPIRICAL CORROBORATION: TECHNICAL ARGUMENTS

This section gives an overview of technical arguments against the empirical corroboration of theory-based findings. Its main purpose is to reflect whether these arguments are important for social simulation or not.

2.1 Lack of Empirical Data

One reason to built simulations is to explore phenomena for which no or only insufficient empirical data exists (van Horn 1971, p. 253). So called *data-poor environments* (Zeigler, Kim, and Praehofer (2000), p. 26)) are the standard in "soft sciences" (see Troitzsch (2004)) like sociology or applied sciences like military operations research (Hofmann 2002), but they occur frequently even in "hard science" like astrophysics or earth sciences (Ruphy 2011, Oreskes et al. 1994), too. In most cases the lack of information is not a matter of effort, but a fundamental problem. In these applications simulation (Kliemt (1996) introduced the term "thin simulation") is mainly an extended "thought experiment" (Paolo, Noble, and Bullock 2000) done with the support of a computer. The computer simulation enables the scientist to master the complexity of the plethora of (possible) derivations from a possible set of modeling assumptions. In general, these simulations cannot be judged on the basis of comparisons with real data, because data from reality is scarce, uncertain or completely missing. Simulation is used in these cases to detect contradictions, shortcomings or anomalies of the model according to general laws, basic assumptions, common sense, and background knowledge which substitute explicit domain knowledge. The whole range of explorative ("What if...?") models belongs to this class of models.

Assessment with respect to social simulation: Although such models (e.g. on urban development in the next 100 years) defy empirical validation on the macro level, at least in the strong sense of validity within an experimental frame, many assumptions made in such models on the micro level can be addressed empirically (water resources available in the urban environment, for example). However, it is the reasoning with such models (Kliemt (1996): "disciplined speculation within thin simulations") which is sophisticated since such models lack predictive validity. Any theory-free, empirically inaccessible simulation and its results are, at least within the epistemological framework of positivism (see section 3.4), pure speculation.

2.2 Insufficient Statistical Significance

Closely related to the problem of scarce empirical data is the challenge of insufficient statistical significance. Ioannidis (2005) unsettled the whole scientific community explaining the statistical reasons "Why most published research finding are false". Ioannidis analyzed the shortcomings of the currently favored statistical evaluation in medicine which is based on *p*-values and power. Among his many findings, he discovered that "large-scale evidence is impossible to obtain for all of the millions and trillions of research questions posed in current research." It is simply infeasible to get sufficient samples size. In light of these results it is questionable whether simulation studies in data-poor environments will ever contribute to the commonly accepted lore of scientific findings. The scarce empirical data available in such environments will be regarded as insufficient for convincing results.

Assessment with respect to social simulation: For the majority of social simulations in data-poor environments it will be indeed infeasible to gather sufficient data for significant validation results. However, science is not limited to the corroboration or refutation of hypotheses by significance test, which are, anyway, overrated (Lambdin (2012)). A major additional task is to generate such hypotheses via explorative research, which is not focused on statistical significance. In addition, it is often not necessarily the single, huge study which is most convincing. A series of non-significant studies can easily uncover even a strong effect in meta-analysis (Thompson 2007, 429)).

2.3 High Sensitivity of Output to Input Data in Nonlinear Systems

Measurement of empiric processes is limited in precision. Hence, the reference data from reality is always a bit fuzzy. This problem is obvious in all sciences dealing with human factors, but it is also inevitable in hard sciences. An illustrative example from the earth sciences can be found in (Nearing, Govers, and Norton 99). Unfortunately, minor differences in initial conditions can cause completely different system state trajectories in nonlinear systems. If the sensitivity of the output data is high within the range of the input measurement uncertainty, the correspondence of specific system and model trajectories is of little value. Goldspink (2002) writes: "Sensitivity to initial conditions with many complex systems means that there may be little validity in directly comparing the response of one system with that of another. This is so because a system's response to perturbation is dependent on its structure and it's history and no two systems will be identical in structure or in their history." In other words, if the real world process is (mathematically) chaotic and (computationally) complex (or at least nonlinear), "how can we ever set up the initial parameters of any model or simulation with sufficient precision so that the simulation output has any correspondence with what might really happen (Byrne 1997)"? This reasoning is likewise valid if the fluctuations are caused by other factors as direct measurement, for example, if some of the input data are abstract, not directly measurable variables (like "motivation" or "risk behavior" which have to be modeled in many social simulations), since both estimations and abstractions enlarge the range of input uncertainty.

Assessment with respect to social simulation: Fortunately, not all systems are chaotic. Many aspects of human behavior seem relatively robust to minor changes. However, punctuational change in social systems (revolution) seems closely related to non-linearity and its high sensitivity to initial conditions. The prediction of discontinuous change in social systems might be an area beyond the reach of science in general. Hence, it depends on the specific research question whether high sensitivity of output to input data is an insurmountable obstacle for empirical validation or not.

2.4 Overfitting

An ubiquitous problem in practice is that a strong comparison between output trajectories of system and model can easily lead to "overfitted" models, which reflect idiosyncracies of special samples that are not attributable to a general rule. Overfitting is sometimes labeled the "curse of predictive modeling". It refers to the phenomenon in which a predictive model may extremely well describe the data from the past used to develop and validate the model, but may subsequently fail to provide valid predictions. The general remedy for this problem are multiple, independent data sets collected over an extended period of time. Arbitrary selections of special data sets and special time intervals are thereby reduced. As a consequence of the inevitable noise in every data set it is also advisable to avoid perfect fitting and to seek for approximations. However, overfitting cannot be completely eliminated. The fitting of a model to any amount of data is always founded on a special selection and truncation of past trajectories.

Assessment with respect to social simulation: Overfitting is a nuisance for the simulation practitioner during model calibration, but problems are generally solved via iterative model adaptations.

2.5 The Triviality and NP-complexity of Calibration

Model calibration is the task of adjusting internal parameters of an already existing model to empirical data from the reference system. Calibration should never be confused with empirical validation, which presupposes a calibration-independent set of data (Trucano et al. 2006). If calibration and empirical validation are confused, validation is trivial, since, mathematically, successful calibration is self-evident if the degrees of freedom of the data generator (model) are greater than the number of data points in the real world reference sample. Most social simulations comply with this criteria.

On the other hand, it can be shown (within a formal framework) that the computational complexity of model calibration is NP-complete (Hofmann 2005). At present, all known algorithms for NP-complete problems require time that is exponential in the problem size. For huge simulations or simulation federations it is impossible to guarantee that the adjustments of the simulation model to the given system input/output are achievable in reasonable time. This view has been first expressed by Carley (1996), without giving a proof: "There is no guarantee that a sufficient large set of procedure and heuristics, that often interact in complex and non-linear ways, can be altered so that they will generate the observed data". In a nutshell, empirical validation presupposes calibration. If calibration is not feasible, validation is, generally, impossible too.

Assessment with respect to social simulation: The confusion of calibration and empirical validation must be avoided. The reputation of social simulation depends on it. The current tendency to include more and more aspects of social reality into simulation models will, sooner or later, reach the complexity limits of calibration, already known in practice via military simulation federations. Extreme fidelity in details is therefore not advised for social simulation. Both problems are, however, avoidable (see section 4.1).

2.6 Trade-offs between Fidelity, Robustness, and Confidence

Also attributed to calibration is a problem Hemez (2004) has discussed in a technical article. He demonstrated formally that there are irrevocable trade-offs between fidelity-to-data, robustness-to-uncertainty, and confidence in prediction. Fidelity-to-data is assured by model comparisons with reference system samples and subsequent calibration. A perfect match between model and empirical data (a "validated model") implies that the robustness of the model to data uncertainty is low. Hemez (2004) even maintains that: "One consequence that cannot be emphasized enough is that the calibration of numerical models - which focuses solely on the fidelity-to-data aspect - is not a sound strategy for selecting models capable of making accurate predictions (p. 39]"

Assessment with respect to social simulation: Any preoccupation with exact fidelity-to-data is misguided for predictive or explorative purposes.

3 LIMITS OF EMPIRICAL CORROBORATION: PHILOSOPHICAL ARGUMENTS

This section gives an overview of philosophical arguments against empirical validation. Its main purpose is to reflect whether these arguments are pertinent to social simulation or not.

3.1 The Theory-ladenness of Observation and Measurements

One of the most important and influential results from the philosophy of science is the theory-ladenness of observation and measurement. It manifests itself in two forms: either as a psychological principle pertaining to human perception (whether scientific or not) or as conceptual insight concerning the nature and functioning of scientific language and its meaning . According to the psychological principle (first noted by James (1890)), perceptions of scientists, as perceptions of all humans, are subject to prior beliefs and expectations. In its conceptual form ((Brewer and Lambert 2001)) it states that observations rest on the theories they accept and that the meaning of the observational terms involved depends upon the theoretical context in which they occur. If simulations are regarded as means to unfold theories the implications of theory-ladenness for the validity of simulation models are straightforward: If observations are theory-laden and if experimentation involves observation, then experimental data has to be theory-laden, too. Since experiments, according to this view, make sense only in relation to some theoretical background they cannot play a role that is theory-independent. That means that experimental data can make sense only on the basis of some prior theory. Consequently, observations are not “bed rock elements”(Balzer, Moulines, and Sneed 1987) on which theories can safely rely. Thus, the validation of a simulation as an extension of theoretical considerations by experimental data may easily become a self fulfilling prophecy. Ahrweiler and Gilbert (2005), 2.7) even maintain that “At the base of theory is again theory. The attempt to validate our theories by ’pure’ theory-neutral observational concepts is misled from the beginning...Not only can you not verify a theory by empirical observation, but you cannot even be certain about falsifying a theory. A theory is not validated by ’observations’ but by other theories (observational theories). Because of this reference to other theories, in fact a nested structure, the theory-ladenness of each observation has negative consequences for the completeness and self-sufficiency of scientific theories (Carrier 1994, p. 1-19). These problems apply equally to simulations, which are just theories in process.”

Assessment with respect to social simulation: Theory-ladenness in both variants is undeniable, and insurmountable within every single scientific project. In social science, theory and data are always closely related. Thus, the quality of a group of rival social simulations addressing the same phenomenon can rarely be assessed satisfyingly by independent data. Scientific progress, however, is made within communities, and over time. The close relation between theory and experimental data is in itself subject to the specific perception and cognition of individual scientists. Thus, the best each research group can do is to gather as much corroborating data as possible, and wait until other groups counter the conceptual mixture of theories and data with their own (see section 3.4).

3.2 Underdetermination, Nonuniqueness or Equifinality

The concept of validation within experimental frames is logically adequate for descriptive and predictive simulations. As long as input/output samples of reality and model correspond, the model fulfills its purpose. However, most simulation models are also intended to be *explicative*. They assume “causal relationships” that should explain why something happened. Whereas an input/output fit is sufficient for description and prediction it is insufficient for explanation for the following reason: Let us regard each simulation as a Turing machine (a generalized computer, *the* universal tool of computation (Hopcroft and Ullman (1979) that produces a certain output from an initial tape configuration (input). There exists an infinite number of different Turing machines that can produce a given finite output from a given finite input (The proof is trivial; note that the inputs and outputs are finite). Each of them can be seen as a different abstract representation of the causal relationships in the real system and all of them fulfill the validation criterion. It is therefore impossible to surely infer causal relationships from output comparisons only. This means that any model is limited in its validity because of “underdetermination”, “nonuniqueness” or “equifinality” (Quine 1977, Oreskes et al. 1994, Refsgaard and Henriksen 2004). Underdetermination indicates that for any finite amount of evidence, there are infinitely many rival models which equally fit with the data. In other words, “the evidence cannot by itself determine that some one of the host of competing theories is the correct one (Klee 1997)” (further information can be found in (Carrier 1994, Harding 1976). This view can even be extended to scientific theories, as Maxwell (1997) states it in his conception of science: “Any scientific theory, however well it has been verified empirically, will always have infinitely many rival theories that fit the available evidence just as well but that make different predictions in an arbitrary way, for yet unobserved phenomena” (for additional information see Richardson (2003) and Fraassen (1980)). In logical terms a model and even a theory can only be regarded as sufficient but not necessary. Oreskes et al. (1994) conclude: “Two ore more constructions that produce the same results may be said to be empirically equivalent. If two theories (or model realizations) are empirically equivalent, then there is no way to choose between them other than to invoke extra-evidential considerations like symmetry, simplicity, and elegance or personal, political or metaphysical preferences.”

Assessment with respect to social simulation: Underdetermination is inescapable, but attenuated by the continuous process of falsification: Most of the theories and models that can reproduce past data are incapable to generate or explain new findings.

3.3 Unpredictability of Social Systems

In the military domain, actually a subbranch of social simulation, the term "fog of war", introduced by von Clausewitz (1991) describes the uncertainty and ambiguity of military operations. War is commonly seen as inherently volatile, uncertain, complex and ambiguous. Although ubiquitous at the strategic and operational level, the practical importance of military uncertainty is most vividly demonstrated at the tactical level. It includes military commanders' imperfect intelligence regarding their enemies' numbers, disposition, capabilities, current locations, and especially, intents, regarding features of terrain and environment, and even including inaccurate knowledge of the (physical and mental) state of their own forces. During actual high intensity combat the uncertainty of war and the chaos of the battlefield even increase, since chance and imponderables like fear, hate and despair gain importance. Taking all these factors into account it is commonly agreed among military experts that a perfect prediction or control of military situations is impossible. Similar lines of reasoning can be found for most social systems. Is empirical validation therefore limited to the past and useless for the future?

Assessment with respect to social simulation: The intensive use of decision supporting simulation in the military domain demonstrates that although military experts are skeptical about exact prediction they are optimistic about beneficial exploration (Hofmann 2013), for at least 2500 years: "It is not a matter of predicting the future, but of being prepared for it (Pericles, 495 BC - 429 BC)." Empirical validation of the elementary military micro processes (movement, attrition, reconnaissance etc.) is considered to be indispensable for such models, and such limited empirical validation is possible in many other social domains, too.

3.4 In Search for an Adequate Epistemology

Objective, empirically proven validity is a notion that is applicable only within the epistemological framework of positivism, which is, as demonstrated below, an ill-suited epistemology for social science, in general, and social simulation, in particular. The issue of simulation validity with respect to different epistemological perspectives has been addressed first by Naylor and Finger (1967) and received thorough investigation by Barlas and Carpenter (1990), Landry and Oral (1993), and Kleindorfer, O'Neill, and Ganeshan (1998). Here, it would go much too far to discuss all of the positions in the philosophy of science. Fortunately, the point I want to make, can be discussed using only three views: Positivism, positive economics and constructivism.

Positivism is based on the belief that reality is independent from the human observer's perception and is totally governed by laws of nature. The positivistic epistemology is founded on the notion, that humans can fully understand reality, and that experiments can reveal the "true" (in the sense of "observer independent") nature of a phenomenon. The methodology is completely constrained to empiric experiments: All open questions are formulated as hypothesis which are corroborated or refuted on the basis of crucial experiments that follow strict rules in design and documentation. Knowledge is consequently the correspondence between reality and its mental or formal representation. The logical and linguistic foundations of positivism have been completely put into question by two of the most important philosophers of the 20th century: Quine (1977) and Wittgenstein (1953). Therefore, today, this position is often attenuated to a kind of "pragmatic realism", which means that scientist have the aim of developing and using models that are as "realistic as possible“, given the constraints of current knowledge, skills, language, computing power and available time (see (Beven 2002) for a critical discussion of this philosophy). The crucial idea of positivism with respect to the issue of validation is that scientist when addressing the same research question with the same method should get the same results, even if they do not communicate. This is an indispensable precondition of validation in positivism which is obviously not true for social (*theory-free*) simulation: Without explicit prearrangement different research groups will develop diverging social simulation models, based on different, often contradicting sets of assumptions, producing different results. Thus, positivism is inappropriate for the validation of social simulations.

Friedman (1953) tried to liberate economics from the straightjacket of positivism by claiming that "truly important and significant hypotheses will be found to have 'assumptions' that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions (in this sense) (p. 14)."

This conception seems to save the approach of social (theory-free) simulation: Different assumptions, even contradicting or counter-factual assumptions seem to be justifiable. However, Friedman connected the whole idea of positive economics to *successful prediction*. He claimed that a hypothesis (or model) must be "able to predict at least as much as an alternate theory" and that it must be "fruitful in the precision and scope of its predictions and in its ability to generate additional research lines (p. 10)." Consequently, social simulation models that cannot generate useful predictions have to be refuted in the philosophical context of positive economics.

An epistemology which is adequate for the validation of non-predictive, theory-free social simulations is constructivism (The only alternative I can think of is Feyerabend's "Anything goes".): In contrast to most other epistemologies constructivism is based on idealism: Different subjective realities coexist as mental constructs. The observer and his or her cognitive apparatus is not neutral. The "raw data" is never perceived raw but always as already interpreted. With other words, each observation is the result of an interaction between observer and observed situation, thus *the results are strongly influenced by the observers knowledge, attitudes, and values* (v. Glasersfeld 1997)).

According to the constructivist view, the validation of simulation results against empirical data sets "is not about comparing the real world and the simulation output; it is comparing what you observe as the real world with what you observe as the output. Both are constructions of an observer and his views concerning relevant agents and their attributes. Constructing reality and constructing simulation are just two ways of an observer seeing the world (Ahrweiler and Gilbert 2005)."

It is essential to realize that constructivism does *not only* call for developer-independent validation of simulation models, but for the independent development of simulation models. This demand follows from the idea that each research group is limited by a specific construction of reality based on a particular set of assumptions. With other words, the experimental frames considered to be appropriate for a research question differ between research groups.

Clearly, such an observer-oriented view of the world is unsatisfying to most scientist, and in order to avoid both solipsism and indiscriminate relativism it is indeed necessary to explain, how individual perceptions and constructions of the world converge to common pictures of reality, that are shared and trusted. *Ultimately, this convergence is nothing more than a consensus about the reality observed by the members of a special community.* This consensus is traditionally generated by correct useful predictions of models based on scientific theories. Without either successful prediction nor already established theory, there is only the chance to establish such a consensus via "interaction that creates an area of shared meanings and expectations (Ahrweiler and Gilbert 2005)." With other words: Consensus can be reached if independently developed models come to the same conclusions, despite different assumptions and different experimental frames.

Within constructivism, single experimental frames (which are the foundation of each empirical validation!) are not guarantors of truth, but heavily biased approaches to reality. Take for example the input trajectories of any real world queuing system (in front of a till). At least, one has to select a specific starting and end point of the reference trajectory. In order to ensure validation within experimental frames it would be necessary to find objective criteria for such selections. Unfortunately, in general, there is an infinite amount of possible criteria for that selections, none of which can be excluded on formal grounds. A subjective element of choice is inevitable. In fact, it is the purposeful selection of input trajectories, control conditions and output summaries which is both at the root of many success stories of sciences as well as of many scientific scandals. But even if no fraud is consciously committed, the scientist cannot universally ensure that his selection is appropriate for an objective evaluation of his model. It seems that this calamity is much more severe for social science than for natural science (which is theory-based and predictive): The micro level of physical simulations is composed of already corroborated laws, and/or their predictions on the macro level are regularly proven correct. Hence, the dependence from subjective conceptions and biased experimental frames in the natural sciences is less important than in social science.

Assessment with respect to social simulation: The notion of a context- and group-dependent truth in constructivism implies that it cannot be the task of the individual social research group to use empirical validation to render their results bullet-proof, *since each single group is limited by their perception, language, and world-view*. They should confine their efforts in order to make all their assumptions and unavoidable biases clear to the reader, and to demonstrate that the empirical data they have gathered is in line with *their* cognitive setting. The transparency of their modeling is the indispensable prerequisite for other research groups for the comparison with their own cognitive setting.

4 IMPLICATIONS FOR PRACTICAL WORK

4.1 Direct Implications

Most of the technical arguments that try to set limits for empirical validation do not limit empirical validation in itself but the range of pertinent research questions. A purely speculative model of a fictitious society is a thought experiment, not a hypothesis to be refuted by data. The lack of data is here simply irrelevant. A chaotic (in the technically sense of the word) social system is, admittedly, a hard limit for simulation, too. If the evolution of any social system is as dependent from initial conditions as the three-body problem in physics, empirical model validation would be impossible. For scientific research, however, it is reasonable to assume that most social systems are governed by non-chaotic rules that allow understanding, and sometimes maybe even prediction. In any case, sensitivity to initial conditions is, a priori, not a killer criterion for empirical validation.

Due to underdetermination any empirically corroborated model is only a working hypothesis. There is an infinite number of (mostly unknown) rival models that equally fit the data. *The preference of a special model has to be motivated with additional reasoning.* Such reasoning can include parsimony, symmetry, resemblance to already known models from similar research fields, “generally accepted” cause-effect relations, and even plausibility and intuition.

The inevitable subjectivity of theory-free, non-predictive modeling not only calls for developer-independent validation. It also calls for meta-analysis between independently developed models since their is no established single experimental frame all research groups can accept. For that aim the modeling purpose, the conceptual model, the program itself, the experimental design, and the used data for validation have to be documented as transparent as possible. The critical peer review of the results of a social simulation model must mandatory include all the assumptions made. In addition, no single research group can effectively ensure the validity of its own simulation model, regardless of how much real world comparisons they have made. They are limited by their view of the world. Trustworthy scientific knowledge cannot be based on single theory-free, non-predictive social simulations but only on the similar results of many of them. In that sense, extremely complex models of single research groups are a threat for scientific veracity. It is, for example, almost impossible for external experts to assess the overall quality of huge military simulation federations. Ultimately, in practice, the results of such models have to be believed or “instinctively” rejected. The amount of time, personal and money necessary to independently access such models is in the majority of cases disproportional to the achievable results. Thus, simulation models for decision support should not be more complicated than absolutely necessary for the given purpose. (For further arguments for simple models see, for example Ward (1989), Salt (1993), Pedgren, Shannon, and Sadowski (1995), Barretto, Chwif, and Paul (2000)).

Consequently, given a problem to be solved via modeling, and a certain amount of money, it seems advisable not to spent the whole fund into one simulation project. A multitude of independently developed and "micro-validated" models (not necessarily simulations) will probably provide more insights than a single model ever can, especially after comparing and discussing different results and subsequently different modeling assumptions, methods and raw data.

4.2 A Shift Towards Meta-analysis

For decades null hypothesis significance testing has been the standard procedure to ensure that effects (differences between samples) are unlikely caused by randomness. However, the use of significance testing in the analysis of research data has been thoroughly discredited, both logically and conceptually, from numerous top statisticians – continuously for almost 100 years (Boring 1919, Berkson 1938, Bakan 1966, Greenwald 1975, Tukey 1991, Cohen 1994, Schmidt and Hunter 1997, Ioannidis 2005, Armstrong 2007, Lambdin 2012). It is beyond the scope of this paper to discuss all these criticism, and what has been said in support of significance tests (Mulaik et al. 1997, Hagen 1997, Senn 2001), but the social simulation practitioner should know that the scientific value of a single significant study is much lower than what has been expected for a long time (Schmidt and Hunter 1997). Today, meta-analysis (re-analyzing the results from different research groups) is considered to be the only way to establish a solid scientific foundation based on statistical findings. A similar shift seems appropriate for the empirical validation of social simulations. Due to underdetermination, inevitable subjectivity, and dependence of the epistemological perspective any single social simulation study should be regarded to be an exploratory endeavor with a focus on the validation of empirically accessible underlying micro processes. Only if several research groups, based on own social simulation models, own experiential frames, and own assumptions have come to similar results, ambitious empirical data validation with the aim of exact evaluation on the macro level should be started.

A positive example of such an meta-analysis is summarized by Arnold (2014) for variants of Schelling' neighborhood segregation model (Schelling 1971). A critical example can be found in (Arnold 2013). He demonstrates that the results of Axelrod's reiterated prisoner dilemma model (Axelrod 1984) are not robust to variations in basic assumptions.

After such a shift, it is *transparency* (on all levels of the model building process, e.g. assumptions, model, code, input and output data, experimental design and frame, etc.) which becomes paramount for the single social simulation project, whereas empirical validation on the macro level becomes crucial only after meta-analysis.

5 SUMMARY AND CONCLUSION

It is commonly accepted that many natural and almost all social systems are "epistemologically open" in the sense that not everything pertaining to these systems or contributing to their behavior can be modeled within a single model. Hence, such models are always purpose-driven abstractions of reality. Comparisons between outputs trajectories of reference system samples and model results are therefore only feasible within a framework taking into consideration these abstractions imposed by the purpose of the modeling. A general approach to such a framework has been established by Zeigler et al.(Zeigler, Kim, and Praehofer 2000). Within such an experimental frame objective validity seems to be a reachable and therefore compulsory goal. This impression is dangerous. It puts far too much trust into the power of formalized and objective approaches, and also in the certainty of empirical comparisons under a specific mindset. Experimental frames are adequate tools for the validation of a huge range of (technical) simulations (representing epistemologically closed because fully specified systems), but they a not unshakable ground for social simulations. No formal framework can guarantee the validity of such models. Expert intuition, common sense (even common "Weltanschauung", see (Ackoff 1979), which discusses the topic in the context of Operations Research), and open discussion are the supreme ingredients for assessing computational representations of social reality. Objective, irrefutable validation of models of natural and social phenomena (extracted from epistemologically open systems) is both logically and practically impossible. Therefore a single experimental frame cannot guarantee the objective "validity" of simulation results. Actually, the fit of a model and empirical data can only be regarded to be a confirmation of a possible model and its specific experimental frame ("history matching", (Konikov and Bredehoeft 1992)), which is necessarily based on personal selections. The data used for validation, the experimental frame, and the simulation model itself are interdependent. Thus, even the most "significant" result of a single social simulation study is seldom as convincing as independent confirmation among some exploratory models. For each single social simulation study it is therefore recommended to put the focus on transparency with respect to assumptions, model and data. Empirical validation should be the focus when comparisons with similar models are possible.

REFERENCES

Ackoff, R. L. 1979. "Resurrecting the future of Operational Research". *Journal of the Operational Research Society* 30 (3): 189–199.

Ahrweiler, P., and N. Gilbert. 2005. "Cafe Nero: the evaluation of social simulation". *Journal of Artificial Societies and Social Simulation* 8 (4): 14.

Armstrong, J. S. 2007. "Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries". *Internation Journal of Forecasting* 23:335–336.

Arnold, E. 2013. "Simulation models of the evolution of cooperation as proofs of logical possibilities. How useful are they?". *Ethics and Politics* 2 (XV): 101–138.

Arnold, E. 2014. "What's wrong with social simulations?". *The Monist* 97 (3): 359–377.

Axelrod, R. 1984. *The Evolution of Cooperation*. Basic Books.

Bair, L., and A. Tolk. 2013. "Towards a unified theory of validation". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S. Kim, A. Tolk, R. Hill, and M. Kuhl. Institute of Electrical and Electronics Engineers, Inc.

Bakan, D. 1966. "The test of significance in psychological research". *Psychological Bulletin* 66:423–437.

Balzer, W., C. Moulines, and J. Sneed. 1987. *An Architectonic for Science. The Structuralist Program*. Reidel.

Barlas, Y., and S. Carpenter. 1990. "Philosophical roots of model validation: Two paradigms". *Systems Dynamics Review* 6:48–166.

Barretto, M., L. Chwif, and R. J. Paul. 2000. "On simulation model complexity". In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. Joines, R. Barton, K. Kang, and P. Fishwick, 449–455. Piscataway, New Jersey: Insitute of Electrical and Electronics Engineers.

Berkson, J. 1938. "Some significance of interpretation encountered in the application of the Chisquare test". *Journal of the American Statistical Association* 33:526–536.

Beven, K. 2002. "Towards a coherent philosophy for modelling the environment". In *Proceedings of the Royal Society London*, Volume 458, 1–20. Royal Society London.

Boring, E. 1919. "Mathematical vs. scientific significance". *Psychological Bulletin* 16:335–338.

Brewer, W. F., and B. L. Lambert. 2001. "The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process". *Philosophy of Science* 68 (3): 176 –186.

Byrne, D. 1997. "Simulation - a way forward?". *Sociological Research Online* 2 (2): 4.

Carley, K. 1996. "Validating computational models". Technical report, Carnegie Mellon University.

Carrier, M. 1994. *The Completeness of Scientific Theories*. Dordrecht: Kluwer.

Cohen, J. 1994. "The earth is round ($p < 0.5$)". *American Psychologist* 12:997–1003.

Fraassen, B. C. V. 1980. *The Scientific Image*. Oxford University Press.

Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.

Goldspink, C. 2002. "Methodological implications of complex systems approaches to sociality: simulation as a foundation for knowledge". *Journal of Artificial Societies and Social Simulation, JASSS* 5 (1): 3.

Greenwald, A. 1975. “Consequences of predjudice agains the null hypothesis”. *Psychological Bulletin* 82:1–20.

Hagen, R. 1997. “In praise of the null hypothesis test”. *American Psychologist* 52:15–24.

Harding, S. 1976. *Can Theories be Refuted? Essays on the Duhem-Quine Thesis*. Reidel.

Helmer, O., and N. Rescher. 1959. “On the epistemology of the inexact sciences”. *Management Science* 6:25–52.

Hemez, F. M. 2004. “The myth of science-based predictive modeling”. In *Foundations 2004 Workshop for Verification, Validation, and Accreditation in the 21st Century*. Arizona State University, Tempe, Arizona: Los Alamos National Laboratory.

Hermann, C. F. 1967. “Validation problems in games and simulations with special reference to models of international politics”. *Behavioral Science* 12:216–231.

Hofmann, M. 2002. “Validation: real world system knowledge, types of validity and credibility levels”. In *Proceedings of the 16. European simulation Mulitconference*. Darmstadt, DE: SCS Europe.

Hofmann, M. 2005. “On the complexity of parameter calibration in simulation models”. *Journal of Defense Modeling and Simulation* 2 (4): 217–226.

Hofmann, M. 2013. “Simulation-based exploratory data generation and analysis (data farming): a critical reflection on its validity and methodology”. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 10 (4): 381–393.

Hopcroft, J. E., and J. D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley.

Ioannidis, J. 2005. “Why most puplished research findings are false”. *PLoS Medicine* 2 (8): e124.

James, W. 1890. *Principles of Psychology*, Volume 1. New York: Henry Hold and Company.

Klee, R. 1997. *Introduction to the Philosophy of Sciende: Cutting Nature at Its Seams*. Oxford University Press.

Kleindorfer, G. B., L. O’Neill, and R. Ganeshan. 1998. “Validation in simulation: various positions in the philosophy of science”. *Management Science* 44 (8): 1087–1099.

Kliemt, H. 1996. *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, Chapter Simulation and Rational Practice, 13–26. Dordrecht: Kluwer.

Konikov, L., and J. Bredehoeft. 1992. “Ground-water models cannot be validated”. *Adv. Water Resour.* 15:75–83.

Lambdin, C. 2012. “Significance tests as sorcery: Science is empirical - significance tests are not”. *Theory and Psychology* 22 (1): 67–90.

Landry, M., and M. Oral. 1993. “In search for a valid view of model validation for Operations Research”. *European Journal of Operational Research; Special Issue on Model Validation* 66:161–167.

Maki, U. 2000. “Kinds of Assumptions and their Truth: Shaking an Untwisted F-Twist”. *Kyklos* 53 (3): 317–336.

Maxwell, N. 1997, August. “A new conception of science”. *Physics World* 13 (8): 17–18.

Mulaik, S., N. Raju, and R. Harshman. 1997. “There is a time and a place for significance testsng”. In *What if there were no significance tests?*, edited by L. Harlowand, S. Mulaik, and J. Steiger, 65–115. Erlbaum.

Musgrave, A. 1981. “Unreal Assumptions in Economic Theory”. *Kyklos* 34:377–387.

Naylor, T. H., and J. M. Finger. 1967. “Verification of computer simulation models”. *Management Science* 14 (2): 92–102.

Nearing, M., G. Govers, and L. Norton. 99. “Variability in soil erosion data from replicated plots”. *Soil Science Society of America Journal* 63:1829–1832.

Olsen, M., and M. Raunak. 2013. “A framework for simulation validation coverage”. In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S. Kim, A. Tolk, R. Hill, and M. Kuhl. Institute of Electrical and Electronics Engineers, Inc.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994, February. "Verification, validation, and confirmation of numerical models in the earth sciences". *Science* 263 (4): 641–646.

Pace, D. 2004. "Modeling and simulation verifcation and validation challenges". *Johns Hopkins APL Technical Digest* 25 (2): 163–172.

Paolo, E. D., J. Noble, and S. Bullock. 2000. "Simulation models as opaque thought experiments". In *Articial Life VII: Proceedings of the Seventh International Conference on Articial Life*, edited by M. A. Bedau, J. S. McCaskill, N. H. Packard, and S. Rasmussen, 497–506. Cambridge, MA: MIT: MIT Press.

Pedgren, C. D., R. E. Shannon, and R. P. Sadowski. 1995. *Introduction to simulation using SIMAN*. 2nd ed. McGraw-Hill.

Quine, W. 1977. *Ontological Relativity*. Columbia University Press.

Refsgaard, J., and H. Henriksen. 2004. "Modelling guidelines–terminology and guiding principles". *Advances in Water Resources* 27:71–82.

Richardson, K. A. 2003. "On the limits of bottom-Upp computer simulation: towards a nonlinear modeling culture". In *Proceedings of the 36th Hawaii International Conference on System Science*. Institute of Electrical and Electronics Engineers, Inc.

Ruphy, S. 2011. "Limits to modeling: balancing ambition and outcome in astrophysics and cosmology". *Simulation and Gaming* 42 (2): 177–194.

Salt, J. D. 1993. "Keynote Address: Simulation should be easy and fun!". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 1–5. Piscataway, N. J.: Institute of Electrical and Electronics Engineers, Inc.

Schelling, T. 1971. "Dynamic models of segregation". *The Journal of Mathematical Sociology* 1 (2): 143–186.

Schmidt, F., and J. Hunter. 1997. "Eight common but false objections to the discontinuation of significance testing in the analysis of research datat". In *What if there were no significance tests?*, edited by L. L. Harlow, S. A. Mulaik, and J. H. Steiger, 37–64. Erlbaum.

Senn, S. 2001. "Two cheers for P-values?". *Journal of Epidemiology and Biostatistics* 6 (2): 193–204.

Thompson, B. 2007. "Effect sizes, confidence intervals, and confidence intervals for effect sizes". *Psychology in the Schools* 44 (5): 423–432.

Troitzsch, K. 2004. *Networked Simulation and Simulated Networks*, Chapter Validating simulation models, 265–270. SCS Publishing House.

Trucano, T., L. Swiler, T. Igusa, W. Oberkampf, and M. Pilch. 2006. "Calibration, validation, and sensitivity analysis: Whats what". *Reliability Engineering and System Safety* 91:1331–1357.

Tukey, J. 1991. "The philosophy of multiple comparison". *Statistical Science* 6:100–116.

v. Glasersfeld, E. 1997. *The Construction of Knowledge: Contributions to Conceptual Semantics*. Intersystems Publications.

van Horn, R. L. 1971. "Validation of Simulation Results". *Management Science* 17 (5): 247–258.

von Clausewitz, C. 1934 (1991). *Vom Kriege (On War; 1832)*. F. Dmmlers Verlag.

Ward, S. C. 1989. "Arguments for constructively simple models". *Journal of the Operational Research Society* 40 (2): 141–153.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Zeigler, B., T. Kim, and H. Praehofer. 2000. *Theory of Modeling and Simulation, 2nd ed*. New York: Academic Press.

AUTHOR BIOGRAPHY

MARKO HOFMANN is Chief Scientist at ITIS GmbH in Neubiberg, Germany since 2000, and adjunct Professor at the University of the Federal Armed Forces in Munich, Germany since 2010. He holds a M.S., a Ph.D. and the venia legendi in Computer Science. His email address is marko.hofmann@unibw.de.