Comput Econ (2007) 30:195–226
DOI 10.1007/s10614-007-9104-4

A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems

Giorgio Fagiolo · Alessio Moneta · Paul Windrum

Received: 10 August 2006 / Accepted: 18 July 2007 / Published online: 4 September 2007

Abstract This paper addresses the methodological problems of empirical validation in agent-based (AB) models in economics and how these are currently being tackled. We first identify a set of issues that are common to all modelers engaged in empirical validation. We then propose a novel taxonomy, which captures the relevant dimensions along which AB economics models differ. We argue that these dimensions affect the way in which empirical validation is carried out by AB modelers and we critically discuss the main alternative approaches to empirical validation being developed in AB economics. We conclude by focusing on a set of (as yet) unresolved issues for empirical validation that require future research.

Keywords Methodology · Agent-based computational economics · Simulation models · Empirical validation · Calibration · History-friendly modeling

JEL Classification B41 · B52 · C63

G. Fagiolo (✉)
Laboratory of Economics and Management, Sant'Anna School of Advanced Studies,
Piazza Martiri della Libertà 33, Pisa 56127, Italy
e-mail: giorgio.fagiolo@sssup.it

A. Moneta
Evolutionary Economics Group, Max Planck Institute of Economics, Jena, Germany
e-mail: moneta@econ.mpg.de

P. Windrum
Manchester Metropolitan University Business School, Manchester, UK
e-mail: p.windrum@mmu.ac.uk

1 Introduction

The aim of this paper is to provide a critical guide to alternative approaches to empirical validation developed in agent-based (AB) economics in recent years. More specifically, the paper explores a set of fundamental methodological problems faced by all researchers engaged in empirically validating economics AB models and discusses alternative solutions within three domains: (i) the relationship between theory and empirical research, (ii) the relationship between models and the real-world systems being modeled, and (iii) the way in which a validation procedure deals with (i) and (ii).

The last two decades have seen a rapid growth in AB modeling in economics. An exhaustive survey of this vast literature is of course beyond the scope of this work.[1] However, before proceeding, it is useful to introduce the main (but by no means all necessary) ingredients that tend to characterize economics AB models.

1. *A bottom-up perspective* A satisfactory account of a decentralised economy is to be addressed using a bottom-up perspective because aggregate properties are the outcome of micro-dynamics involving basic entities (agents) (Tesfatsion 2002). This contrasts with the top-down nature of traditional neoclassical models, where the bottom level typically comprises a representative individual and is constrained by strong consistency requirements associated with equilibrium and hyper-rationality. Conversely, AB models study economic systems that may be persistently out of equilibrium (if any) or fluctuating around some meta-stable states.
2. *Heterogeneity* Agents are (or might be) heterogeneous in almost all their characteristics. These can range from initial endowments and other agents' properties, all the way through to behavioral rules, competencies, rationality, and computational skills.
3. *Bounded rationality* The environment in which real-world economic agents live is too complex for hyper-rationality to be a viable simplifying assumption (Dosi et al. 2005). It is suggested that one can, at most, impute to agents some local and partial (both in time and space) principles of rationality (e.g., myopic optimization rules). More generally, agents are assumed to behave as boundedly rational entities with adaptive expectations. Moreover, since they are not initially endowed with a full understanding of the underlying structure of the environment in which they operate, they engage in open-ended searches wherein the nature of learning is at odds with Bayesian decision rules assumed by neoclassical economics (Dosi et al. 2005).
4. *Networked direct interactions* Interactions among economic agents in AB models are direct and inherently non-linear (Fagiolo 1998; Windrum and Birchenhall 1998; Silverberg et al. 1988). Agents interact directly because current decisions

[1] On the (often subtle) differences which characterise different research schools that has been employing AB models to study market and industry dynamics (e.g., evolutionary economics, agent-based computational economics, neo-Schumpeterian, and history-friendly models), cf. Lane (1993a,b), Dosi and Nelson (1994), Nelson (1995), Silverberg and Verspagen (1995), Tesfatsion (1997, 2002), Windrum (2004), Dawid (2006), and Pyka and Fagiolo (2005). Also see Gilbert and Troitzsch (1999), and Wooldridge and Jennings (1995) for a discussion of AB techniques in other social sciences.

directly depend, through adaptive expectations, on the past choices made by other agents in the population (i.e., a widespread presence of externalities). These may contain structures, such as subgroups of agents or local networks. In such structures, members of the population are in some sense closer to certain individuals in the socio-economic space than others. These interaction structures may themselves endogenously change over time, since agents can strategically decide with whom to interact according to the expected payoffs. When combined with heterogeneity and bounded rationality, it is likely that aggregation processes are non-trivial and, sometimes, generate the emergence of structurally new objects (Lane 1993a,b).

Having defined the main ingredients of AB models, it is clear that profound differences exist between these types of models and neoclassical micro–macro models (e.g., standard models of exogenous or endogenous growth and game-theoretic approaches in industrial economics). In particular, a fundamental difference exist regarding how economists should process information derived from real world observation (e.g., behaviors of the agents and interaction structures) and plug it into their models.

In fact, the interest in AB modeling was stimulated by the breakdown of the general equilibrium approach and the rise of new classical macro models in the 1970s and 1980s. This became the dominant means of representing a dynamic, decentralized market economic in neoclassical economics. AB models reject the aprioristic commitment of neoclassical models to individual hyper-rationality, continuous equilibrium, and representative agents.[2] Everything in the neoclassical world can, in principle, be known and understood. It is often assumed that the entire set of objects in the world (e.g., techniques of production, or products) is known at the outset. The opposite is the case in the AB world. There the set is unknown, and agents must engage in an pen-ended search for new objects. Associated with this distinction are important differences with regards to the types of innovative learning and adaptation that are considered, definitions of bounded rationality, the treatment of heterogeneity amongst individual agents and the interaction between these individuals, and whether the economic system is characterized as being in equilibrium or far-from-equilibrium.

AB researchers have enjoyed significant success over the last 20 years. Indeed, neoclassical economists have (sometimes) recognized the significance of the AB critique, and have reacted by extending their own modeling framework to incorporate (certain) aspects of heterogeneity, bounded rationality, learning, increasing returns, and technological change. Yet orthodox neoclassicals have not been moved to join the AB camp. There are many possible explanations for this but an important aspect, recognised by AB modelers themselves, is a perceived lack of robustness in AB modeling. This threatens the AB research enterprise as a whole. Four key problem areas

[2] An alternative view (though one which we doubt would be shared by AB economists themselves) is that the AB approach is complementary to neoclassical economics. Departures from standard neoclassical assumptions, found in AB models, can be interpreted as ‘what if,’ instrumentalist explorations of the space of initial assumptions. For example, what happens if we do not suppose hyper-rationality on the part of individuals? What if agents decide on the basis of bounded rationality? and so on.

were identified in a recent conference and special workshop attended by the authors.[3] First, the neoclassical community has consistently developed a core set of theoretical models and applied these to a range of research areas. The AB community has not done this. Indeed, the sheer diversity of alternative AB models put forward over the last 20 years is striking (cf. Leombruni et al. 2006). There is little or no understanding of the connection amongst the set of highly heterogeneous models that has been developed.[4]

A second, related set of issues concerns a lack of comparability between the models that have been developed. Not only do the models have different theoretical content but also they seek to explain strikingly different phenomena. Where they do seek to explain similar phenomena, little or no in-depth research has been undertaken to compare and evaluate their relative explanatory performance. Rather, models are viewed in isolation of one another. The problem is compounded by the high degrees of freedom in AB models. Not only do AB models contain highly non-linear, recursive interactions and feedbacks, but they tend to have many dimensions and, hence, degrees of freedom. With many degrees of freedom, a wide range of simulation outputs can be generated by an AB model.

This leads us to a third set of issues. These concern the lack of standard techniques for constructing and analyzing AB models. It has been argued that developing a set of commonly accepted protocols for AB model building would benefit the profession (Leombruni 2002; Richiardi 2003). This would address, for instance, issues such as how and when sensitivity analysis (over the space of initial conditions and parameters) should be conducted, how one should deal with non-ergodicity in underlying stochastic processes, and how one should interpret, in terms of real-world time, the timing and lag structures that are typically built into AB models.

The fourth, and final, set of issues—which is our main concern in this paper—regards the problematic relationship between AB models and empirical data. As well as there being diversity with respect to the process of constructing and analyzing AB models, fundamental differences exist in the ways researchers conduct empirical validation (if any). In what follows empirical validation involves examining the extent to which the output traces generated by a particular model approximates reality, typically described by one or more ‘stylized facts’ drawn from empirical research.

Key areas of debate include the following questions: Is a ‘realist’ methodology appropriate? Why should empirical validation be the primary basis for accepting or rejecting a model? Do other tests of model validation exist than the reproduction of stylized facts? If we do proceed down the path of empirical validation, then how should we relate and calibrate the construction of parameters, initial conditions, and stochastic variability in AB models to the existing empirical data? Which classes of empirically observed objects do we actually want to replicate? How dependable are the micro and

[3] At a special session on ‘Methodological Issues in Empirically-based Simulation Modeling’, hosted by Windrum and Fagiolo at the 4th EMAEE conference, Utrecht, May 2005, and at the ACEPOL 2005 International Workshop on ‘Agent-Based Models for Economic Policy Design,’ Bielefeld, July 2005.

[4] Take, for example, the many types of models that have been put forward to explain technological diffusion. What are the common features among sand pile models, Polya urn models, dynamic learning models such as genetic algorithms, evolutionary games, and network models?

macro stylized facts to be replicated? To what extent can we truly compare simulated output traces with stylized facts or, alternatively, with counterfactuals? What are the consequences, for the explanative power of a model, if the stylized facts are actually ‘unconditional objects’ that only indicate properties of stationary distributions and, hence, do not provide information on the dynamics of the stochastic processes that generated them? More generally, if the empirical phenomena that are specified as stylized facts are themselves rather general in nature, comparisons with stylized facts not only represent a weak test for the validity of individual models, but also fail to provide a strong methodological basis for comparing competing models. Exploring the possible answers to these questions is one of the goals of this paper.

Before moving on, we note that we have implicitly assumed that a high degree of heterogeneity is problematic. Heterogeneity makes it difficult to compare alternative models that seek to explain the same phenomenon, makes it difficult to advance a new, alternative paradigm, and to contrast it with the existing neoclassical paradigm. Having a small set of core models, developed by researchers over time using a (few) commonly accepted protocol(s) for model building and empirical validation would, it is suggested, be better for AB community. It could, however, be argued that a high degree of heterogeneity is, initially at least, commensurate with Kuhn (1962) discussion of the formation of new paradigms. Heterogeneity and flexibility are a consequence of a high degree of scientific debate and dissent on how best to proceed with the development of a new paradigm, and is a necessary prerequisite for the emergence of a more codified paradigm. It should also be remembered that a degree of heterogeneity and dissent also exists within the neoclassical community, and plays a necessary role in maintaining the vitality of that research paradigm. Still, while a degree of heterogeneity is beneficial, this must be balanced against the benefits of establishing a core set of models and modeling techniques that define a research paradigm, and are the focus of a practicing community that adheres to, and develops, a coherent paradigm.

This work focuses on the methodological issues related to the practice of taking AB economics models to the data. We will see that a strongly heterogeneous set of approaches to empirical validation is to be found in the AB literature. We argue that this is also partly due to the high level of heterogeneity that characterizes the process of constructing and analyzing AB models.

The paper is structured as follows. Section 2 discusses the methodological basis of empirical validation, i.e., the issues arising in the comparison of stochastic-dynamic models with empirical data. We identify a set of core issues concerning empirical validation that are faced by all modelers (neoclassical and AB). Section 3 opens the discussion on methodological diversity in empirical validation within the AB community. We suggest that this methodological heterogeneity is due to two factors. The first factor is the set of problems associated with analyzing stochastic, highly nonlinear, disequilibrium models. The second factor is the diverse structural content of AB models, and the very different ways in which AB models are analyzed.

Building on Sect. 3, Sect. 4 provides a detailed survey of three major approaches to AB empirical validation. These are the indirect calibration approach (Sect. 4.1), the Werker–Brenner approach to empirical calibration (Sect. 4.2), and the history-friendly approach (Sect. 4.3). Having highlighted the strength and weaknesses of each approach, Sect. 5 concludes by discussing a set of outstanding, but still open, issues for empirically-oriented AB modelers.

2 The Methodological Basis of Empirical Validation

Models, in economics as in any other scientific discipline, isolate some features of an actual phenomenon, in order to understand it and to predict its future status under novel conditions. These features are usually described in terms of causal relations and it is usually assumed that some causal mechanism (deterministic or stochastic) has generated the data. We call this causal mechanism the 'real-world data generating process' (*rwDGP*). A model approximates portions of the *rwDGP* by means of a 'model data generating process' (*mDGP*). The *mDGP* must be simpler than the *rwDGP* and, in simulation models, generates a set of simulated outputs. The extent to which the *mDGP* is a good representation of the *rwDGP* is evaluated by comparing the simulated outputs of the *mDGP* with the real-world observations of the *rwDGP*. In what follows, we call this procedure *empirical validation*.

Some key methodological problems are involved in this inductive process. The first issue concerns how to deal with the trade-off between 'concretization' and 'isolation.' Faced with the essential complexity of the world, scientific (not only economics) models proceed by simplifying and focusing on the relationships between a very limited number of variables. Is it possible to model all the different elements of the *rwDGP*? And how can we possibly 'know' all the different elements of the *rwDGP*? Leading economists as J. S. Mill and J. M. Keynes have in the past expressed serious doubts about whether we can expect to have models that are fully concretized. In a highly complex world, a fully concretized model would be a one-to-one mapping of the world itself! Thus, economists usually agree that models should isolate some causal mechanisms, by abstracting from certain entities that may have an impact on the phenomenon under examination (Mäki 1992). A series of open questions remains. How can we assess that the mechanisms isolated by the model resemble those operating in the real world? In order to isolate such mechanisms, can we make assumptions that are 'contrary to fact,' that is, assumptions that contradict our knowledge of the situation under discussion?

These dilemmas are strictly related to the trade-off between analytical tractability and descriptive accuracy that is faced by all theoreticians seeking to model markets, industries and other economic systems. Indeed, the more accurate and consistent is our knowledge about reality with respect to assumptions, and the more numerous the number of parameters in a model, the higher is the risk of failing to analytically solve the model (and thus have sharp implications from our set of assumptions). By contrast, the more abstract and simplified the model, the more analytically tractable it is. The neoclassical paradigm comes down strongly on the side of analytical tractability.

This brings us to the second core issue of empirical validation: instrumentalism versus realism. Realism, roughly speaking, claims that theoretical entities 'exist in the reality,' independent of the act of inquiry, representation or measurement (Mäki 1998). By contrast, instrumentalism maintains that theoretical entities are solely instruments for predictions and not true descriptions of the world. A radical instrumentalist is not much concerned with issues of empirical validation, in the sense that she is not much interested in making the model resemble mechanisms operating in the world. Her sole goal is prediction. Indeed, a (consistent) instrumentalist is usually more willing than a realist to ‘play’ with the assumptions and parameters of the model in order to get better predictions. While the neoclassical paradigm has sometimes endorsed instrumentalist statements à la Friedman (1953), it has never allowed a vast range of assumption adjustments in order to get better predictions (e.g., full rationality, equilibrium, etc.). In this sense it has failed to be consistent with its instrumentalist background.

The third issue is related to the choice of a pluralist or apriorist methodology. Methodological pluralism claims that the complexity of the subject studied by economics and the boundedness of our scientific representations implies the possibility of different levels of analysis, different kinds of assumptions to be used in model-building, and legitimacy of different methodological positions. Apriorism is a commitment to a set of a priori assumptions. A certain degree of commitment to a set of a priori assumptions is normal in science. Often these assumptions correspond to what Lakatos (1970) called the ‘hard core’ assumptions of a research program. But strong apriorism is the commitment to a set of a priori (possibly contrary to the facts) assumptions that are never exposed to empirical validation (e.g., general equilibrium and perfect rationality). Theory is considered prior to data and it is denied the possibility of interpreting data without theoretical presuppositions. Typically, strong apriorist positions do not allow a model to be changed in the face of anomalies, and encourages the researcher to produce ad hoc excuses whenever a refutation is encountered. Lakatos (1970) dubbed research programs involved with such positions as ‘degenerating.’

The fourth issue regards the under-determination or identification problem. In other words: What happens when different models are consistent with the data that is used for empirical validation? The issue is known in the philosophy of science as the ‘under-determination of theory by data.’ In econometrics the same idea has been formalized and labeled as ‘the problem of identification.’ As Haavelmo (1944) noted, it is impossible for statistical inference to decide between hypotheses that are observationally equivalent. He suggested specifying an econometric model in such a way that—thanks to restrictions derived from economic theory—the problem of identification does not arise. The under-determination problem is also strictly connected to the so-called Duhem-Quine thesis: it is not possible to test and falsify a single hypothesis in isolation. This is because any hypothesis is inevitably tied to some auxiliary hypotheses. Auxiliary hypotheses typically include background knowledge, rules of inference, and experimental design that cannot be disentangled from the hypothesis we want to test. Thus, if a particular hypothesis is found to be in conflict with the evidence, we cannot reject the hypothesis with certainty, since we do not know if it is the hypothesis under test or one of the auxiliary hypotheses which is at odds with the evidence. As shown by Sawyer et al. (1997), hypothesis testing in economics is further complicated by the approximate nature of theoretical hypotheses. The error in approximation, as well as the less systematic causes disturbing the causal mechanism object of modeling, constitutes an auxiliary hypothesis of typically unknown dimension. For example, in time-series econometric models a distinction is made between ‘signal’ (which captures the causal mechanisms object of interest) and ‘noise’ (accounted by the error terms). But it may be the case, as pointed out by Valente (2005), that noises are stronger than signals, and that the mechanisms involved undergo several or even continuous structural changes. Econometricians have adopted sophisticated tests which are robust to variations in the auxiliary hypotheses (see, for example, Leamer 1978). Nonetheless, the Duhem-Quine thesis still undermines strong apriorist methodologies that do not check the robustness of the empirical results under variations of background assumptions.

3 Empirical Validation and Heterogeneity of AB Models

In the last section, we introduced the main issues that every modeler faces in trying to construct and empirically validate Her model. Let us now turn to discuss the methodological problems related to empirical validation in AB models.

As mentioned in the introduction, there is no consensus at all about how (and if) AB models should be empirically validated (as partially happens also in the neoclassical arena). In this section we argue that this variety depends on two factors. First, AB models invariably contain non-linearities, stochastic dynamics, non-trivial interaction structures among economic agents, and micro–macro feedbacks. Therefore, the resulting macroeconomic dynamics can hardly be studied in equilibrium (e.g., in steady state). This opens up a whole set of methodological problems related to the relationships between *rwDGP* and *mDGP*. Second, heterogeneity in empirical validation procedures might also be due to the lack of standard techniques for constructing and analyzing AB models. This is another key difference to neoclassical modeling, where variety is instead intimately associated with the diverse roles played by statistical inference.

Let us return to the meta-model *rwDGP* and *mDGP* introduced in Sect. 2, and discuss how the output of an AB model can be compared to real-world observations. As illustrated in Fig. 1—see Pyka and Fagiolo (2005) for details—assume that the unknown *rwDGP* has generated a single (observed) instance of a vector of microeconomic time series $\underline{z}_{i,t}$, characterizing the behavior of a population of economic agents labeled by $i \in I = \{1, \ldots\}$. Once aggregated over I, microeconomic time series induce a vector of macroeconomic time series $\underline{Z}_t$.

Let us suppose that we observe the single instance of micro and macro time-series in the sample period $t_0, \ldots, t_1$. Assume that the goal of the AB modeler is to proxy the *rwDGP* with the *mDGP* generated by Her AB model and suppose that each agent i in the model is fully described by a vector of microeconomic parameters $\underline{\theta}_i$ (governing, e.g., its behaviors and interactions) and a vector of microeconomic variables $\underline{x}_i$. Let us further suppose that the environment (i.e., the economy) is completely characterized by a vector of macro-economic parameters Θ. For any particular choice of initial conditions $\underline{x}_{i,0}$, and micro and macro parameters, the AB model will output a vector of micro time series $\underline{x}_{i,t}$ for each agent and, upon aggregation, a vector of macroeconomic time series $\underline{X}_t$.

To begin with, the extent to which the *mGDP* accurately represents the *rwDGP* depends on many preliminary, model-related factors. These range from the quality of micro and macro parameters that are specified, to set of initial micro and macro conditions that are taken to proxy initial real-world conditions.

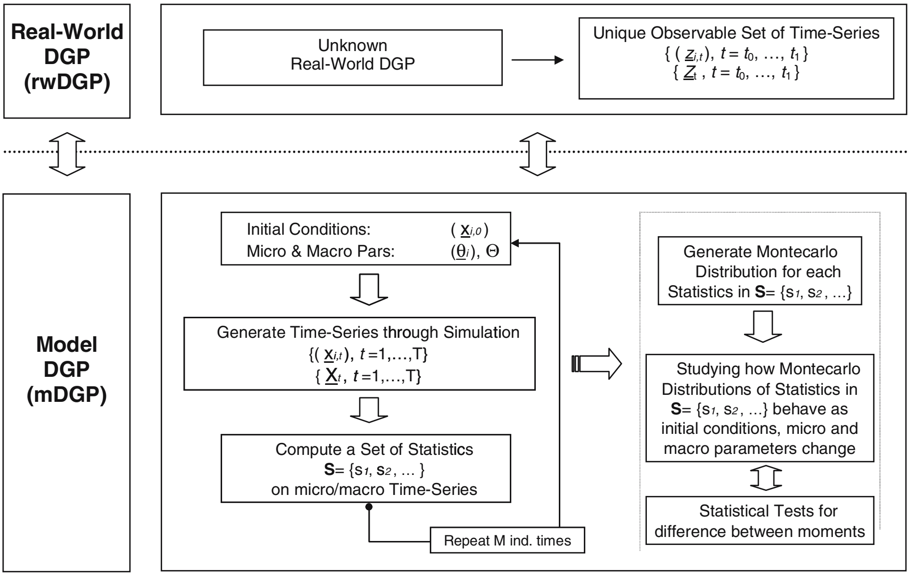
Real-World DGP (rwDGP)
Unknown Real-World DGP
Unique Observable Set of Time-Series
$\{ (\underline{z}_{i,t}),\ t = t_0, \ldots, t_1 \}$
$\{ \underline{Z}_t,\ t = t_0, \ldots, t_1 \}$
Model DGP (mDGP)
Initial Conditions: $(\underline{x}_{i,0})$
Micro & Macro Pars: $(\underline{\theta}_i)$, Θ
Generate Time-Series through Simulation
$\{(\underline{x}_{i,t}),\ t=1,\ldots,T\}$
$\{\underline{X}_t,\ t=1,\ldots,T\}$
Compute a Set of Statistics
$\mathbf{S}= \{s_1, s_2, \ldots\}$
on micro/macro Time-Series
Repeat M ind. times
Generate Montecarlo Distribution for each Statistics in $\mathbf{S}= \{s_1, s_2, \ldots\}$
Studying how Montecarlo Distributions of Statistics in $\mathbf{S}= \{s_1, s_2, \ldots\}$ behave as initial conditions, micro and macro parameters change
Statistical Tests for difference between moments

Fig. 1 A procedure for studying the output of an AB model

The problems of developing a good representation are compounded when discrete-time models contain (as invariably AB models do): (a) non-linearities and randomness in individual behaviors (decision rules) and interaction networks; (b) micro and macro variables that are governed by complicated stochastic processes that can hardly be analyzed analytically (hence the need for computer simulation); (c) feedbacks between the micro and macro levels (due, for example, to macro-level variables affecting agents' adaptive expectations). Indeed, all these ingredients make it difficult to interpret the output of an AB model in terms of the set of its assumptions, as the causal links between the former and the latter become very weak.

To understand why, using Fig. 1, let us consider one possible procedure for studying the output of an AB model. Suppose the modeler knows (from a preliminary simulation study, or from some *ex ante* knowledge coming from the particular structure of the AB model under study) that the real-world system is ergodic, and that the *mDGP* displays a sufficiently stationary behavior for a time period after T^* for (almost all) points of the parameter space and initial conditions. Thus, (s)he will simulate the system for at least $T > T^*$ time steps.

Now suppose we are interested in a set of statistics $S = \{s_1, \ldots, s_j, \ldots\}$ that are to be computed on the simulated data generated by the *mDGP* $\{\underline{x}_{i,t}, t = 1, \ldots, T\}$ and $\{\underline{X}_t, t = 1, \ldots, T\}$.[5] For any given run ($m = 1, 2, \ldots, M$), the simulation will output a value for $s_j \in S$. Given the stochastic and possibly non-linear nature of the process,

[5] For example, one of the micro variables might be an individual firm's output and the corresponding macro variable may be GDP. In this case, we may be interested in aggregate statistics s_j such as the average rate of growth of the economy over T time-steps (e.g., quarters).

each run—and thus each value of s_j—will be different from the others, because each single microeconomic time series $\underline{x}_{i,t}$ will differ from run to run.

Therefore, after having produced M independent runs, one has generated a distribution for s_j containing M observations. This distribution can be summarized by computing, for example, its mean $E(s_j)$, its variance $V(s_j)$, and so on. Recall, however, that the moments will depend on the initial choices that were made for $\underline{\theta}_i$, Θ, and $\underline{x}_{i,0}$. By exploring a sufficiently large number of points in the space of initial conditions and parameter values, and by computing $E(s_j)$, $V(s_j)$, etc. at each point, one can gain a deep understanding of the behavior of the *mDGP* of the model system.[6] It is easy to see that the more the *mDGP* contains non-linear and stochastic elements, the looser is the mapping between initial conditions, micro- and macro-parameters, and the moments of the distributions for $E(s_j)$, $V(s_j)$, etc.

The comparison of the *mDGP* and the *rwDGP* in AB models opens up a whole range of new methodological problems. For example, how can one compare the single trace observed in the real world (or, by the same token, the stylized fact) with the distribution of traces generated by the model? How can one deal with the degrees of freedom allowed by the parameter space? To what extent can we truly consider simulated output traces to be stylized facts or, alternatively, counterfactuals? We will come back to these (and related) questions in more detail in the next sections. What is important to stress here is that different answers to the above questions have implied very diverse ways to perform empirical validation in AB models.

There is a second reason why empirical validation techniques are so different in AB models. Indeed, as mentioned in the introduction, there is no consensus among AB modelers on the techniques themselves that are to be employed to construct and analyze AB models. AB models tend in fact to differentiate along four key dimensions: the nature of the object(s) under study, the goal of the analysis, the modeling assumptions, and the method of sensitivity analysis that is used.

The *nature of the object(s) under study* regards the (empirically observed) stylized facts that the model is seeking to explain. Significant differences exist with respect to the nature of the object being studied. Where neoclassical modelers are interested in quantitative change, AB modelers are equally interested in qualitative change of economic systems themselves. For instance, there are AB models that investigate how R&D spending affects the qualitative nature of macroeconomic growth. Other AB models investigate its quantitative impact, e.g., seek to explain some statistically observed quantitative property of aggregate growth (e.g., its autocorrelation patterns). Another important distinction is between AB models that seek to investigate a single phenomenon, and those that jointly investigate multiple phenomena. For instance, a model may consider the properties of productivity and investment time-series, in addition to the properties of aggregate growth. Transient versus long-run impact is a further distinction. For example, there are AB models that examine the effect of R&D

[6] Consider the example of footnote 5 once again. One may plot $E(s_j)$, which is the Monte-Carlo mean of an economy's average growth rates, against key macro parameters such as the aggregate propensity to invest in R&D. This may allow one to understand whether the overall performance of the economy increases in the model with that propensity. Moreover, non-parametric statistical tests can be conducted to see if $E(s_j)$ differs significantly in two extreme cases, such as high versus low propensity to invest in R&D.

spending on growth along the diffusion path (the transient) of a newly introduced technology. Other AB models are only concerned with the magnitude of a technology's long-run impact (when the economic system has stabilized somewhat). Finally, an important distinction exists between AB models that investigate micro distributions and macro aggregates. The former are concerned with the dynamics of industry-level distributions, such as a cross-section of firm productivity or size distributions, for a particular sector, in a particular year. The latter are concerned with longer time-series data for nation states, or the world economy, over a number of years.

A *second* dimension in which AB models differ is in the *goal of the analysis*. AB models tend to deal with in-sample data. In-sample data is relevant when one is interested in describing or replicating observed phenomena. Out-of-sample exercises are less frequently carried out by AB economists. For instance, there are no examples of AB models, dealing with technological change and business cycles, that attempt to provide predictions of the out-of-sample behavior of GDPs over given periods of time (e.g., with the goal of answering control-related issues). Only recently have AB models have been employed to generate policy implications and to address issues related to market-design.[7]

A *third* dimension concerns *modeling assumptions*. Some models contain many degrees of freedom, others do not. For example, agents in AB models may be characterized by many variables and parameters. Their decision rules may, in turn, be highly-parameterized. Alternatively, agents and decision rules may be described in a very stylized way. Individual decision rules and interaction structures may be exogenously fixed or they may change over time. Change may be driven by exogenous, stochastic factors. Alternatively, change may be driven by agents endogenously selecting new decision rules and interaction structures according to some meta-criteria (as happens in endogenous network formation models, see Fagiolo et al. 2004b).

The *fourth* and final dimension is the *method of sensitivity analysis*. In order to thoroughly assess the properties of an AB model, the researcher needs to perform a detailed sensitivity analysis, along the lines sketched in Fig. 1. This sensitivity analysis should, at the very least, explore how the results depend on (i) micro-macro parameters, (ii) initial conditions, and (iii) across-run variability induced by stochastic elements (e.g., random initial conditions, and random individual decision rules).

These four key dimensions strongly inform the choice of empirical validation procedure that is used. The focus on qualitative or quantitative phenomena, on micro or macro phenomena, and on transients or long-run impacts, determine the type of data that is required for empirical validation, the statistical procedures to be followed, and the ability to generate empirically testable implications. Additionally, the extent to which sensitivity analysis is performed prior to empirical validation has important implications for the universality of the simulation results that are obtained.

[7] See Marks (2005; 2007), Koesrindartoto et al. (2005), and the special issue 'Agent-Based Models for Economic Policy Design,' edited by Herbert Dawid and Giorgio Fagiolo, forthcoming in the Journal of Economic Behavior and Organization.

4 Alternative Approaches to Empirical Validation in AB Models

In this section we review three of the most influential approaches to empirical validation developed in the AB literature and assess their strengths and weaknesses.

Of course, these three approaches to empirical validation are not the only ones present in the literature on simulation modeling. Our choice has been driven by the observation that they are indeed the main ones employed in the majority of AB economics models, in particular as far as models of general disequilibrium, industry and market dynamics are concerned.[8]

The approaches considered are: the indirect calibration approach, the Werker–Brenner approach, and the history-friendly approach. These three approaches are strongly shaping the debate on validation within AB economics. Our goal is to provide a general overview of each approach, and to consider how each tackles the methodological issues identified in Sects. 2 and 3. Each attempts to reduce the number of model parameters, and to reduce the space of possible 'worlds' that are explored, by tying the model down to an observed empirical reality. Each, however, does this in a very different way. The history-friendly approach constrains parameters, interactions, and decision rules of the model in line with a specific, empirically-observable history of a particular industry. It can be interpreted as a calibration exercise with respect to unique historical traces. The other two approaches do not impose any preliminary set of restrictions on parameters but, rather, indirectly employ empirical evidence to identify sub-regions in the potential parameter space. Within these sub-regions, a model is expected to replicate some relevant statistical regularities or stylized facts.

Prior to this discussion, however, we need to briefly consider qualitative AB models and their contribution to the validation debate. There exist a significant number of AB models that engage in purely qualitative theorizing, and which are not validated in any meaningful sense. In economics there is a long tradition, stretching back to the earliest classical economists, of using models as a means to engage in abstract *Gedankenexperimente*. Many AB models do this. In accordance with our thesis on the relationship between model content and empirical validation, a significant number of AB models seek to explain qualitative phenomena that are intrinsically closed to quantitative analysis. There is no rationale for testing such models against existing empirical data sets. Notable examples are evolutionary game-theoretic models (Vega-Redondo 1996), and Polya urn models (Arthur 1988, 1994). There is a weak relationship between the micro-macro variables/parameters of these models and their empirically observed counterparts. The focus of such models is the emergence of qualitative aggregate patterns, such as the emergence of coordination and cooperation. Forecasting exercises are possible but they typically generate unpredictability results. For example, one knows with certainty that users will lock into one of the

[8] It must be noted that quite a large literature exists on empirical validation of simulation models in other social sciences, computer science, engineering, etc.: for an introduction, see Leigh Tesfatsion's web site on empirical validation at http://www.econ.iastate.edu/tesfatsi/empvalid.htm. For some examples of alternative empirical validation techniques in simulation models, cf. Klejinen (2000), Sargent (1998), Barreteau (2003).

competing technologies in Arthur's (1994) Polya urn model but it is impossible to know *ex ante* which of the competing technologies will be selected.

Some AB economists, engaged in qualitative modeling, are critical of the suggestion that meaningful empirical validation is possible. They suggest there are inherent difficulties in trying to develop an empirically-based social science that is akin to the natural sciences. Socio-economic systems, it is argued, are inherently open-ended, interdependent and subject to structural change. How can one then hope to effectively isolate a specific 'sphere of reality', specify of all relations of phenomena within the sphere with the external environment, and build a model describing all important phenomena observed within the sphere (together with all essential influences of the external environment)? In the face of such difficulties, some AB modelers do not believe it is possible to represent the social context as vectors of quantitative variables with stable dimensions (Valente 2005).[9]

One possible reaction is to use the computer as an artificial laboratory in which basic, causal relationships can be tested in order to gain some knowledge of the underlying (much more intricate and convoluted) real-world causal structure. The danger of this strategy is that one ends up building auto-referential formalizations that have no link to reality (Edmonds and Moss 2005). Certainly there are those in other social science disciplines who have taken the step of accepting they are constructing and analyzing synthetic artificial worlds which may or may not have a link with the world we observe (Doran 1997). Those taking this position open themselves to the proposition that a model should be judged by the criteria that are used in mathematics: i.e., coherence, precision, soundness, and generality. This is hardly the case with AB models! The majority of AB modelers do not go down this particular path. Instead, they employ methodological approaches that seek to deal with the difficult issues and problems discussed in Sects. 2 and 3.

Nevertheless, advocates of qualitative simulation warn us about the problems that arise from the inherent structural non-stationarity and the interdependence of socio-economic systems. These points echo Kaldor's discussion of stylized facts (Kaldor 1961; Kwásnicki 1998). Kaldor proposed that when theory cannot assist, we should use empirical knowledge to restrict the dimension space of initial conditions and the micro/macro parameters of a model. Kaldor observed that, in practice, we tend to be hampered by a lack of large, high quality data sets. When this is the case, Kaldor argued, we should use stylized facts or statistical regularities to pin down values or value ranges for key parameters. By stressing the reproduction (explanation) or prediction of a set of stylized facts, one hopes to circumvent problems of data availability and reliability.

Calibration addresses the over-parameterization problem by reducing the space of possible 'worlds' explored by an AB model (Kwásnicki 1998). This is done through the use of empirical data, such that the model *mDGP* resembles as closely as possible the

[9] Between the position of Valente (2005), who regards empirical validation as unfeasible as priori, and empirically-calibrated models, there are of course modelers that take the intermediary position of referring both to non-validated qualitative aggregate patterns and quantitative parameters.

actual *rwDGP* that we observe.[10] This can be a sensible goal whenever the analysis aims to forecast the future or generate policy implications. However, on the down side, calibration has a strongly conservative tendency. It supports the continuation of orthodox theories and models for which empirical data is readily available. It disadvantages new theories and new models for which empirical research has not yet caught up, and mitigates against the study of qualitative phenomena that are either difficult to measure or are inherently immeasurable by their very nature. There exist numerous practical problems involved in actually carrying out calibration exercises. A notable problem is the availability of high-quality data in sufficient quantities over the many parameters of a typical AB model. Even if this is achieved, the modeler faces a range of problems such as knowing *ex ante* whether the real-world system being modeled is ergodic or non-ergodic, and the ability to correctly set the initial conditions of the model, the beginning and end points of the simulation runs, so as to match that of the real-world system captured in the empirical data sets. The upshot is that a complete calibration which leads to a clear value for each parameter is impossible.

4.1 The Indirect Calibration Approach

Drawing upon a combination of stylized facts and empirical data sets, many AB modelers have been developing a pragmatic four-step approach to empirical validation. As its name suggests, the indirect calibration approach first performs output validation, and then indirectly calibrates the model by focusing on the parameters that are consistent with output validation. In the first step, the modeler identifies a set of stylized facts that (s)he is interested in reproducing and/or explaining with Her model. Stylized facts typically concern the macro-level (e.g., the relationship between unemployment rates and GDP growth) but can also relate to cross-sectional regularities (e.g., the shape of the distributions of firm size). In the second step, along with the prescriptions of the empirical calibration procedure, the researcher builds the model in a way that keeps the microeconomic description as close as possible to empirical and experimental evidence about microeconomic behavior and interactions. This step entails gathering all possible evidence about the underlying principles that inform real-world behaviors (of, e.g., firms, consumers, and industries) so that the microeconomic level is modeled in a not-too-unrealistic fashion. In the third step, the empirical evidence on stylized facts is used to restrict the space of parameters, and the initial conditions if the model turns out to be non-ergodic.

The foregoing procedure is an exercise in 'indirect calibration.' Suppose, for example, that the Beveridge curve is one of the statistical regularities being investigated. Therefore, the model must be able to replicate a relationship in which unemployment rates decrease with vacancy rates in the labor market (cf. Fagiolo et al. 2004a). The researcher should restrict Her further analysis to all (and only) parameter combinations under which the model does not reject that hypothesis (at some confidence level). This step is the most sensible because it involves a fine sampling of the parameter space. It

[10] For a notable example of calibration on AB models, see Bianchi, C. et al. (Unpublished manuscript), who perform calibration on the CATS model developed in a series of papers by Gallegati et al. (2003, 2005).

is also computationally demanding and requires the use of Monte-Carlo techniques. Indeed, for any given point in the parameter space, one must generate a distribution for the statistics summarizing the stylized facts of interest (e.g., the slope of the relationship between unemployment and vacancy rate), and test the null hypothesis that the empirically observed valued can be generated by our model under that particular parameter combination (see Fig. 1).

In the fourth and final step, the researcher should deepen her understanding of the causal mechanisms that underlie the stylized facts being studied and/or explore the emergence of fresh stylized facts (i.e., statistical regularities that are different to the stylized facts of interest) which the model can validate *ex post*. This might be done by further investigating the subspace of parameters that resist to the third step, i.e., those consistent with the stylized facts of interest. For example, one might study how the absolute value of the Monte-Carlo average of the slope of the unemployment-vacancy rate relation varies with some macro-parameter (if any) that governs wage setting and/or union power in the model. This can shed light on the causal mechanism underlying the emergence of a Beveridge curve. Similarly, one can ask whether business cycle properties (e.g., average and volatility of growth rates) change with the slope of the Beveridge relation. If this is the case, a fresh implication generated by the model (under empirically plausible parameters) can be taken to the data—and further provide support for the AB model under scrutiny.

A stream of recent AB contributions to the fields of industry- and market-dynamics has been strongly rooted in the four-step empirical validation procedure just presented. For example, Fagiolo and Dosi (2003) study an evolutionary growth model that is able to reproduce several stylized facts about output dynamics, such as I(1) patterns of GNP growth, growth-rates autocorrelation structure, absence of size-effects, etc., while explaining the emergence of self-sustaining growth as the solution of the trade-off between exploitation of existing resources and exploration of new ones. Similarly, Fagiolo et al. (2004a) present a model of labor and output market dynamics that is not only able to jointly reproduce the Beveridge curve, the Okun curve and the wage curve, but also relates average growth rates of the system to the institutional set-up of the labor market. Finally, Dosi et al. (2005) analyze an evolutionary growth model that is capable of replicating and explaining the most important statistical properties of business cycles.

While appealing, the indirect calibration approach is open to criticism in at least two important respects. First, no attempt is made to calibrate micro and macro parameters using their empirical counterparts. There are two reasons for this. On the one hand, the models address in-sample exercises almost exclusively. On the other hand, due to the difficulties of matching theoretical and empirical observations, one is bound to be as agnostic as possible as to whether the details of a model (variables, parameters) can be really compared with empirically-observable ones. However, in order for this indirect calibration procedure to be effective, the empirical phenomena of interest should not be very general. Otherwise, they might not necessarily represent a difficult test for the model. If this is the case, the model might pass the validation procedure without providing any effective explanation of the phenomena of interest (e.g., no restrictions on the parameter space would be made). This parallels Brock's discussion of 'unconditional objects.' The fundamental issue of discriminating between the 'descriptions' and 'explanations' of reality pops up once again.

The second problem is far subtler, and has to do with the interpretation of the points belonging to the sub-region of the parameter space (and initial conditions) that resist the sort of 'exercise in plausibility' that one performs in the third step of the procedure. After a suitable sub-region of the parameter space (and initial conditions) has been singled out—according to the capability of the model to replicate the set of stylized facts of interests in that sub-region—how should one interpret all comparative exercises that aim at understanding what happens when one tunes the parameters within that sub-region? For instance, suppose one has found the range of values for the parameter 'strength of union power' that is consistent with the emergence of a Beveridge curve in the labor market, and is interested in asking the question: How does the average growth rate of the economy change when the strength of union power moves within those bounds? In all these cases, an interpretation problem arises: What does it really add—to our understanding of reality—providing an answer to this type of questions? How can one interpret alternative parameter values in an evolutionary world where history, indeterminacy, and non-linear feedbacks between the micro and macro levels may strongly affect the outcomes?[11]

4.2 The Werker–Brenner Approach to Empirical Calibration

Empirical calibration of AB models has been proposed by Werker and Brenner (2004) see also Brenner and Werker (2007), and applied in Brenner and Murmann (2003), and Brenner (2004). The Werker–Brenner approach is a three-step procedure for empirical calibration. The first two steps are consistent with all calibration exercises. The third step is novel. Step 1 uses existing empirical knowledge to calibrate initial conditions and the ranges of model parameters. Werker–Brenner propose that, where sensible data is not available, the model should be left as general as possible, i.e., wide ranges should be specified for parameters on which there is little or no reliable data. Step 2 involves empirical validation of the outputs for each of the model specifications derived from Step 1. Through empirical validation, the plausible set of dimensions within the initial dimension space is further reduced. The Werker–Brenner approach advocates the use of Bayesian inference procedures in order to conduct this output validation. Each model specification is assigned a likelihood of being accepted based on the percentage of 'theoretical realizations' that are compatible with each 'empirical realization.' In this way, empirically observed realizations are used to further restrict

[11] Notice that, at least in principle, initial conditions should always matter, irrespective of whether the investigated real-world DGP is ergodic or not. This is because in reality we always deal with limited time spans. Nevertheless, knowing whether our *mDGP* is ergodic or not can be important. There are four cases to be considered. First, if both the *rwDGP* and the *mDGP* are ergodic, initial conditions matter only until the model reaches some stationary state, while their impact tend to vanish in the long run. Second, if the investigated system is ergodic while the *mDGP* is not, we are actually employing a wrong model. Third, if neither the *rwDGP* nor *mDGP* are ergodic, we have to face the following insurmountable question: which time window of *mDGP* and *rwDGP* should we select to perform empirical validation? Finally, if the investigated system is non-ergodic but our model is ergodic, we are led astray. We thank an anonymous referee for pointing out this important issue.

the initial set of model specifications (parameter values) that are to be considered. However, all parameter values with a positive likelihood have to be included in the further analysis. Model specifications that conflict with current data are discounted. Step 3 involves a further round of calibration. This uses the surviving set of models and, where helpful, recourse to expert testimony from historians. This they call 'methodological abduction.' In effect, one is trying to identify an underlying structural model from the shared properties and characteristics of the surviving models. The authors argue that "these [shared] characteristics can be expected to hold also for the real systems (given the development of the model has not included any crucial and false premises)" (Werker and Brenner 2004, p. 13).

The Werker–Brenner approach is attractive in a number of respects. First, it addresses head-on many of the issues of model evaluation: it offers a means of reducing the degree of freedom in models, it advocates testing procedures for sensitivity analysis on large numbers of simulations, and it proposes the application of well-established Bayesian inference procedures for establishing rigorous empirical tests for both model parameters and outputs. It also avoids a number of potential pitfalls associated with developing models based on single case-study histories. Second, it appears to offer a powerful methodology for developing rigorous, empirically-grounded simulation models that explicitly take into account competing theories and assumptions. As with all approaches, there are strengths and weaknesses to empirical calibration. Let us consider some important methodological and operational issues associated with calibration. These, we hasten add, are not specific to Werker–Brenner, but are generic to all calibration approaches (cf. also beginning of Sect. 4).

First, assessing fitness amongst a class of models does not automatically help us identify a true (but unknown) underlying model.[12] If the initial set of models do not fit well, in the sense that they do not represent the *rwDGP*, then any likelihood-based method of selecting or averaging models can produce bad results. This is despite a claim, sometimes seen in the literature, that Bayesian methods work even on collections of false models. That claim is based on the result that Bayesian model comparison leads to the model that is closest to the true (unknown) model in Kullback–Leibler (KL) distance. KL distance can be an arbitrarily bad choice from a decision-theoretic perspective. Essentially, KL distance looks for where models make the most different predictions—even if these differences concern aspects of the data behavior that are unimportant to us. But if the set of models does not contain the true underlying model, and all models within the set are seriously flawed then we will not want to select a model based on KL distance. These points are discussed in Schorfheide (2000).

Second, there is a strong tendency for calibration to influence the types of models we develop. Notably, empirically calibrated models encourage the modeler to focus on variables and parameters that are readily calibrated and for which data already exists (Chattoe 2002). Yet, there are many potentially important variables and parameters for which data does not currently exist. Some may not be amenable to quantitative measurement. For instance, agents' mental models are an important component in many AB economics models. Yet the mental models used by real world agents tend

[12] Note that the Werker–Brenner approach is based on the assumption that the true model cannot be identified.

to be unobservable in practice. The calibration approach tends to induce the modeler either to abstract from the micro features of the economy, or to force calibration of those parameters using unreliable or inconsistent data. The approach also impacts on the types of model outputs that are considered. Again, there is a temptation to focus on outputs that are readily measured, and not to consider phenomena that cannot be measured or calibrated a priori. There is an inherent conservativeness here, a conservativeness which inhibits the search for new theories and new explanatory variables.

A third issue is the quality of the available empirical data. The most common reason for under-determination in economics is the bias and incompleteness of the available data sets. It is not always possible to exclude a particular model on the basis of existing empirical data because other types of data can potentially support the model, if they had been collected. Effective calibration requires a wealth of high quality data. Indeed, the Werker–Brenner calibration approach is particularly demanding because it requires the modeler to engage in two rounds of empirical validation. Unfortunately, in economics (and in the other social sciences, for that matter) empirical data is always scarce. There are three reasons for this lack of extensive data. First, there is the cost of organizing and collecting raw data, processing this raw data, and constructing organized data sets. The organization and construction of national and international data sets depends on the existence of specialist statistical offices. Even then, high costs mean that national and international statistical offices are not able to collect data on all matters. Second, there is an inherent bias in the collection process. People who collect data are informed by existing theory on which data to collect. Established theory not only informs choices about which variables to measure (and not to measure), but also how to measure the selected variables—i.e., the key indicators to be used. Hence, there is an inherent tendency to maintain the status quo and to lock out new, alternative concepts, theories and models. For new ideas to succeed in economics, new data sets may be required in order to test new theories and models. Perhaps the best-known example of this is the Keynesian revolution. Theoretical developments went hand-in-hand with the collection of aggregate national data for the first time, notably on household consumption expenditure and firm investment expenditure. Third, there is the nature of the phenomenon being observed. It may be that a particular phenomenon is rarely observed, or it is a unique event that is non-reproducible. The issue is widely discussed in statistics texts. The practical upshot is that, while in principle we could generate as many theoretical observations as we like, in practice we may only have a few of such empirical realizations (possibly only one!). If we believe that the empirical observations come from an underlying DGP that could have been 'played twice' (i.e., could have generated alternative observations, other than the one we have) the problem of comparing simulated with empirical data becomes very complicated. We will return to this issue in Sect. 4.3.

A fourth issue highlighted by calibration is the nature of the relationship between the model *mDGP* and the real-world *rwDGP*. First, there is the question of whether the *rwDGP* is ergodic or non-ergodic. If the underlying real-world *rwDGP* is thought to be non-ergodic (as well as the theoretical *mDGP* described in the AB model), then initial conditions matter (see footnote 11). This raises a whole host of problems for the modeler. The modeler needs to identify the 'true' set of initial conditions in the empirical data, generated by the *rwDGP*, in order to correctly set the initial parameters of the model. Even if perfect data exists (which is unlikely), this is a very difficult task. How far in the past does one need to go in order to identify the correct set initial values for the relevant micro and macro variables? There is a possibility of infinite regress. If this is the case, then one may need data stretching back a very long time, possibly before data started to be collected.

Fifth, even when the *mDGP* and *rwDGP* are thought to be (sufficiently) stationary processes, the problem of correctly setting t_0 remains. An important decision to make is about the particular sub-sample of simulated data (of length $\tau = t_n - t_0$) that is to be compared with the empirical data. The underlying *rwDGP* may generate a number of different regimes, e.g., the same macroeconomic structure may generate a diverse set of outcomes that include economic depression, full employment, inflation, hyper-inflation, and even stagflation. If this is the case, then one is faced with the problem of which sub-sample of simulated and observed time-series should be compared in order to carry out model validation. By incorrectly setting time t_0 in the model, one can generate a set of simulated output data that describes a different regime to that found in the empirical data. In addition to the issue of correctly setting t_0, one must identify the appropriate point at which to stop the simulation runs, i.e., to correctly set t_n. If t_n is set incorrectly then the simulated data may include multiple regimes that are covered by the empirical data. If the start or end points (or both) for the simulation runs are incorrectly set, there is the danger that one incorrectly rejects a 'true' model on the basis of its simulated outputs. We should also note that if, as is frequently the case, the modeler sets the simulation runs to end at a point where the model reaches a stationary or almost stationary behavior, one is implicitly assuming that the empirical evidence comes from a stationary DGP. This may, or may not, be the case.

Sixth, what if the observed micro and macro parameters are time dependent? One needs to be sure that the empirically estimated parameters that we assume are slow changing variables (and, hence, can reasonably be treated as fixed within the time-scale explored by the model) are not actually time dependent. If they are, then the researcher needs to go back and rethink the structural relationships between slow and fast variables, the timescale of the model,[13] or both.

Finally, issues of prediction and counterfactuals are core to calibration. To what extent do the predictions of the models take into account data that lies outside the current regime? The Lucas critique (Lucas 1976; Lucas and Sargent 1979) is relevant here. Real economic agents not only use statistical processes based on past experience (adaptive expectations) but use current data to project into the future. In this way, agents are able to respond to exogenous economic shocks. This was the basis for the rational expectations critique of Keynesian behavioral models. Exogenous economic shocks alter behaviors but leave the underlining structure unchanged. As a consequence Keynesian theories seriously mispredict the consequences of a shock, whereas a model of the micro fundamentals—individual production functions and utility functions—would not.

[13] An important issue related to time-scales in AB models, which we shall just mention here, concerns the choice made about the timing in the model. Whether we assume that the time-interval $[t, t+1]$ describes a day, or a quarter, or a year (and whether one supposes that the 'updating scheme' is asynchronous or parallel), has non-trivial consequences for calibration and empirical validation.

4.3 The History-Friendly Approach

The history-friendly approach offers an alternative solution to the problem of over-parametrization. Like the calibration approaches discussed above, it seeks to bring modeling more closely 'in line with the empirical evidence' and thereby reduce the dimensionality of a model. The key difference is that this approach uses the specific historical case studies of an industry to model parameters, agent interactions, and agent decision rules. In effect, it is a calibration approach which uses particular historical traces in order to calibrate a model.

In part, the history-friendly approach represents an attempt to deal with criticisms leveled at early neo-Schumpeterian AB models of technological change. Two of the key protagonists of history-friendly modeling, Richard Nelson and Sydney Winter, were founding fathers of neo-Schumpeterian AB modeling. While the early models were more micro-founded and empirically-driven than contemporary neoclassical models, empirical validation was weak. There was a lack of thorough sensitivity and validation checks and empirical validation, when carried out, tended to consist of little more than a cursory comparison of outputs generated by just a handful of simulation runs with some very general stylized facts. Further, the early models contained many dimensions and so it was rather easy to generate a few outputs that matched some very general observations (the over-parametrization problem).[14]

In terms of our taxonomy, the history-friendly approach is strongly quantitative and mainly focuses on microeconomic transients (industrial paths of development). In this approach a 'good' model is one that can generate multiple stylized facts observed in an industry. The approach has been developed in a series of papers. Key amongst these is Malerba et al. (1999), and Malerba and Orsenigo (2001). In the first of these papers, the authors outlined the approach and then applied it to a discussion of the transition in the computer industry from mainframes to desktop PCs. In the second of these papers, the approach was applied to the pharmaceutical industry and the role of biotech firms therein. Here we shall keep the description of the approach succinct.[15] Through the construction of industry-based AB models, detailed empirical data on an industry informs the AB researcher in model building, analysis and validation. Models are to be built upon a range of available data, from detailed empirical studies to anecdotal evidence to histories written about the industry under study. This range of data is used to assist model building and validation. It should guide the specification of agents (their behavior, decision rules, and interactions), and the environment in which they operate. The data should also assist the identification of initial conditions and parameters on key variables likely to generate the observed history. Finally, the data is to be used to empirically validate the model by comparing its output (the 'simulated trace history') with the 'actual' history of the industry. It is the latter that truly distinguishes the history-friendly approach from other approaches. Previous researchers have used historical case studies to guide the specification of agents and environment, and to identify possible key parameters. The authors of the history-friendly approach

[14] See Windrum (1999) for a detailed discussion of early neo-Schumpeterian models.

[15] Interested readers are directed to Windrum (2007) for a detailed critique of history-friendly modeling.

suggest that, through a process of backward induction one can arrive at a satisfactory approximation of structural assumptions, parameter settings, and initial conditions. Having identified the approximated set of 'history-replicating parameters,' one can carry on and conduct sensitivity analysis to establish whether (in the authors' words) 'history divergent' results are possible.

There are many points here that deserve closer inspection. Let us begin with issues that concern the structure of the model and the object of analysis. First, the modeling activity that has been conducted is, in practice, informed by the history of a few, key companies rather than the history of an entire industry. For instance, Malerba et al. (1999) is calibrated to capture one particular computer company—IBM—rather than the entire industry. This severely restricts the universality of the model. As a consequence, the micro-economic description of the supply-side of the industry is highly stylized. The demand-side of the computer industry model is also highly stylized. Indeed, many of the behavioral assumptions made about the supply and demand sides do not appear to be driven by industry-specific empirical observations. Windrum (2007) suggests that this reflects practical difficulties in collecting sufficient amounts of high quality data at the industry level.

This leads us to an important question: to what extent can one hope to acquire all the relevant data needed to build an empirically sound industry-level model? If this is not possible, then a further question follows: what are we to do if the empirical evidence is incomplete, offers no guidance on a particular point, or else seems to contain alternative, competing viewpoints?

Finally, limited attention is given to sensitivity analysis in the history friendly models, as parameters and rules are supposed to be deduced from the industry under study. The lack of sensitivity analysis is particularly noticeable with regards to cross-run variability.

Aside from the issues relating to implementation, the history-friendly approach raises a set of fundamental methodological issues. First, the approach to empirical validation that is advocated involves comparing the output traces of a simulated model with detailed empirical studies of the actual trace history of an economic system. We are immediately confronted, once again, with problems associated with comparing individual output traces generated by the model *mDGP* with individual traces generated by the real-world *rwDGP*. This does not move us much further on from ascertaining whether a model is 'capable' of generating an output trace that resembles an empirically observed trace. It is not a very strong test. An individual simulated trace may, or may not, be typical of the model. A second issue is the ability to backwardly induce the 'correct' set of structural assumptions, parameter settings, or initial conditions from a set of traces—even if we have a model that generates an appropriate distribution of output traces. Simply stated, there are, in principle, a great many combinations of alternative parameter settings that can produce an identical output trace. We cannot deduce which combination of parameter settings is correct, let alone the appropriate set of structural assumptions. A third issue is implication that we are can construct counterfactual histories (although the authors do not themselves engage in this in their papers). For example, we need to be able to construct a world in which IBM did not enter the PC market. This poses a very serious question. Could the PC market have developed in much the same way had IBM not invented the PC? Can we meaningfully construct a counterfactual history? As Cowan and Foray (2002) discuss, it is exceedingly difficult in practice to construct counterfactual histories because economic systems are stochastic, non-ergodic, and structurally evolve over time.

Finally, a fourth key methodological issue concerns the meaning of history. To what extent can we actually rely on history to be the final arbiter of theoretical and modeling debates? To pose the question another way, can simulations, in principle, be guided by history? In practice, it is unlikely that we will be able to appeal to history, either to bear witness, or to act as a final arbiter in a dispute. This is because history itself is neither simple nor uncontested, and any attempt to develop a historically based approach to modeling faces deep level methodological problems.[16] The development of high quality accounts, open to critical scrutiny, is essential to the history-friendly approach (and indeed any other historically based methodology). It is, after all, on the basis of these accounts that guidance is taken on particular modeling choices, on parameter testing, and output evaluation. In recognizing the limitations of any historical account, we simultaneously identify the limitations of decisions based on that account. But this is a strength, not a weakness, of open academic discourse. How, then, are we to proceed? Let us suggest the following possibility. While a single 'typical' history may not exist, we may be able to draw some generalizations on the basis of a large collection of historical case studies. To use an analogy used by Jerry Silverberg, rather than seeking to develop a model that describes the fall of one particular leaf from a tree (the history friendly approach), we should seek to develop general models, such as the bromide diffusion model in physics, that can be used to explain the fall of many leaves from many trees (and other phenomena). To get to this point, what is needed is the construction of high quality data sets. A wealth of empirical studies within the neo-Schumpeterian tradition, written over the last 20 years, can be drawn upon. It is also likely that new databases will be needed to be constructed in order to develop greater understanding of micro, meso, and macro phenomena. We shall return to this issue later. For the moment, it is important to observe that modelers need to ensure they do not prematurely restrict their models, given the lack of high quality data that currently exists. If the AB models that are developed are not flexible enough to consider alterative scenarios, then we will be left with a set of models that are less, not more, compatible with one another.

4.4 A Taxonomy of Empirical-Validation Approaches in AB Models

Applying the taxonomy developed in Sect. 3, we can identify important differences between the three approaches with respect to the types of empirical data that is used and how these are applied in the process of empirical validation. First, there is the empirical domain addressed by each approach. The indirect calibration and Werker–Brenner approaches can, in principle, be applied to micro and macro AB models (e.g., to describe the dynamics of firms, industries, and countries). By contrast, the

[16] A well-known example of the contestability of history is evidenced by the ongoing debate about whether inferior quality variants can win standards battles (Leibowitz and Margolis 1990; Arthur 1988). As Carr (1961) observed in his classic work, history can be contestable at more fundamental and unavoidable levels.

history-friendly approach only addresses micro dynamics. Second, there are differences between the types of empirical observations (data) used for empirical validation. In addition to empirical data sets, the Werker–Brenner approach advocates the use of historical knowledge. The history-friendly approach allows one to employ casual and anecdotic knowledge as well. Third, there are differences in the way data is actually used. All three approaches use data to assist model building, as well as model validation. Unlike the other two approaches, indirect calibration does not directly employ data to calibrate initial conditions and parameters. Fourth, there are differences in the order in which validation and calibration is performed. Both the Werker–Brenner and the history-friendly approaches first perform calibration and then validation. By contrast, the indirect calibration approach first performs validation, and then indirectly calibrates the model by focusing on the parameters that are consistent with output validation.

Table 1 illustrates the ways in which the indirect calibration, Werker–Brenner, and history-friendly approaches differ to each other.

5 Conclusions and Future Research Issues

This paper has critically examined a set of core issues concerning empirical validation of AB simulation models. The discussion has been wide ranging and extensive. Section 2 defined what an AB model is, and the methodological basis of empirical validation. The modeler develops a theoretical DGP (*mDGP*), that captures the salient features of the real-world DGP (*rwDGP*) while also being simpler than the *rwDGP*. The extent to which the *mDGP* is a good representation of the *rwDGP* is evaluated by comparing the simulated outputs of the *mDGP* with the real-world observations of the *rwDGP*.

Our main starting point was that there is still little consensus on how (and if) one should perform empirical validation in AB economics models. We have then suggested that the reasons for such a lack of agreement may depend on three related factors. First, the very process of empirical validation may be hampered, as happens for any backward-induction procedure, by many, still unresolved problems. We have identified four of them: (1) concretization versus isolation—the identification and isolation of the set of salient mechanisms, forces and causal relations present in the *rwDGP*; (2) the balance between instrumentalist and realist approaches; (3) the choice between methodological pluralism and strong apriorism; (4) the identification or under-determination problem.

These four core issues are generic to all forms of empirical validation. There may be however two other factors, more specific to the AB philosophy, that might explain the existing heterogeneity in approaching empirical validation of AB models in economics. First, the very nature of AB models and their assumptions implies that it is often complicated to obtain sharp relationships between model inputs and outputs. Second, there is still a lot of variety in the techniques employed to build and analyze AB models. To characterise this variety, Sect. 3 presented a novel taxonomy which contains four significant dimensions along which the various approaches differ. These are (1) the nature of the object under study (i.e., the stylized fact(s) under analysis), (2)

Table 1 Differences between the types of data collected and their application

	Empirical domain	The types of data used	The application of data	Order of application
Indirect calibration approach	• Micro (industries, markets) • Macro (countries, world economy)	• Empirical data	• Assisting in model building • Validating simulated output	• First validate, then indirectly calibrate
Werker–Brenner Approach	• Micro (industries, markets) • Macro (countries, world economy)	• Empirical data • Historical knowledge	• Assisting in model building • Calibrating initial conditions and parameters • Validating simulated output	• First calibrate, then validate
History-friendly approach	• Micro (industries, markets)	• Empirical data • Casual, historical and anecdotic knowledge	• Assisting in model building • Calibrating initial conditions and parameters • Validating simulated output	• First calibrate, then validate

the goal of the analysis (in-sample explanation vs. out-of-sample prediction), (3) the nature of the main modeling assumptions (e.g., parameters, decision rules, interaction, and timing), and (4) the method of sensitivity analysis. Having identified the nature and causes of heterogeneity amongst AB modelers, Sect. 4 narrowed the focus by discussing three important approaches to validation within AB economics: indirect calibration, the Werker–Brenner calibration approach, and the history-friendly approach.

In analyzing the strengths and weaknesses of each approach, we were also able to identify a set of unresolved problems that require future research. Let us draw the paper to a close by outlining future prospects with regards to these unresolved problems. These can fruitfully be classified under five headings:

1. Alternative strategies for constructing empirically-based models;
2. Problems due to over-parametrization;
3. Counterfactuals and their use in policy analysis;
4. Defining sufficiently strong empirical tests; and
5. Problems due to the availability, quality and bias of available data sets.

5.1 Alternative Strategies for Constructing Empirically-Based Models

This concerns the crucial question of whether (and how) one should build an AB model that is based on empirical observations. We have discussed how methodological realism embodies the principle that a model should (in addition to other criteria) be judged on the realism of its assumptions. An assumption has a higher degree of realism when it is supported by robust empirical evidence (Kagel and Roth 1995; Plott and Smith 1998). Thus, the model *mDGP* should capture key observed features of the real-world *rwDGP*. Yet there is intense debate about the best way to actually construct empirically-based models, and to select between alternative models. What happens, for instance, if there are alternative assumptions and existing empirical data does not assist in choosing between them? This is the under-determination problem in a new context.

A number of different strategies exist for selecting assumptions in the early stages of model building.[17] One strategy is to start with the simplest possible model, and then proceed to complicate the model step-by-step. This is known as the KISS strategy: ‘Keep it simple, stupid!.’ In effect, it is a reformulation of Ockham’s razor. A very different strategy is the KIDS strategy: ‘Keep it descriptive, stupid!.’ Here one begins with the most accurate model one can devise, and then simplifies it as much as possible. A third strategy, common amongst neoclassical economists, is TAPAS: ‘Take A Previous model and Add Something.’ Here one takes an existing model and successively explores the assumption space through incremental additions and/or the relaxation of initial assumptions.

[17] See Mäki (1994), Edmonds and Moss (2005), Pyka and Fagiolo (2005), Frenken (2005).

5.2 Problems due to Over-Parametrization

Whichever strategy is employed, the AB modeler often faces an over-parametrization problem. AB models with realistic assumptions and agent descriptions invariably contain many degrees of freedom. First, the model may contain a large number of micro–macro parameters. Second, the modeler may explore different interaction setups and agent decision rules. While the latter can reasonably be considered a modeling choice (justified by the particular issue under study), the former is a dimension that is often non-reducible. This is true, even if one adopts a KISS approach to model building and model selection. There are two aspects to the over-parametrization problem. First, the dimensions of the model may be so numerous that it can generate any result. If this is the case, then the explanative potential of the model is little better than a random walk. Second, the causal relations between assumptions and results become increasingly difficult to study the more degrees of freedom there are in the model. The over-parametrization of a model (in the spaces of micro–macro parameters, variables, and/or decision rules) can seriously impair any validation or calibration exercise because there are a number of different combinations of parameter settings that can produce the same output. Since the parameter space describes all possible regimes (behavioral, technological, institutional, etc.) in which an economic system can find itself, discriminating among the different combinations means choosing among possible realities. Which one should then be compared with the empirical evidence?

We have discussed a number of escape strategies for these problems. First, one can use empirical evidence to restrict the degrees of freedom, by directly calibrating initial conditions and/or parameters (i.e., the set of possible 'worlds' modeled by the *mDGP*). Second, one can indirectly calibrate the model by focusing on the subspace of parameters and initial conditions under which the model is able to replicate a set of stylized facts. Unfortunately, this procedure still tends to leave the modeler with multiple possible 'worlds.' The modeler needs to address the issue of how the remaining worlds should be interpreted. What does it mean when one is comparing the model outputs that are generated under alternative sets of initial conditions/parameters that resist direct/indirect calibration? In fact, each combination represents a different economy or world, as it defines a different institutional, technological, market/industry and behavioral setup.

The issue is particularly relevant for the indirect calibration approach because it is frequently the case that many combinations of parameters and initial conditions are consistent with the set of stylized facts of interest. In the Werker–Brenner approach the modeler can also find her/himself in a situation where many parameters and/or initial conditions cannot be directly estimated. Contrasting the output of any two alternative setups—as is often done with the direct and indirect calibration approaches—means one must perform comparative dynamics exercises. Such exercises are informative from a theoretical point of view because they allow for a better understanding of the properties of the model. However the interpretation of the results is unclear as far as policy prediction is concerned.

5.3 Counterfactuals and their use in Policy Analysis

The interpretation of alternative points in the space of parameters/initial conditions brings us to another problem. How does one interpret the counterfactual outputs generated by a model? It is tempting to suggest that outputs which do not accord with empirical observations are counterfactuals, and that the study of these counterfactuals are useful for policy analysis. Cowan and Foray (2002) discuss the issue at length. They suggest that, in practice, it is exceedingly difficult to construct counterfactual histories because economic systems are stochastic, non-ergodic, and structurally evolve over time. As AB models typically include all these elements in their structure, Cowan and Foray argue that using AB models to address counterfactual-type questions may well be misleading. The critique may also apply to the interpretation of empirically plausible parameters and initial conditions. In light of the Cowan–Foray critique, one must consider if any comparative dynamics exercise is informative.

More generally, comparing the outputs generated by AB models with real-world observations involves a set of very intricate issues. For example, Windrum (2007) observes that the uniqueness of historical events sets up a whole series of problems. In order to move beyond the study of individual traces, we need to know if the distribution of output traces generated by the model *mDGP* approximates the actual historical traces generated by the *rwDGP* under investigation. A way to circumvent the uniqueness problem is to employ a strong invariance assumption on the *rwDGP*, thereby pooling data that should otherwise be considered a set of unique observations. For example, one typically supposes that cross-country aggregate output growth rates come from the same DGP. Similarly, it is supposed that the process that drives firm growth does not change across industries or time (up to some mean or variance scaling). This allows one to build cross-section and time-series panel data.

Unfortunately we cannot know if the suppositions are valid. Consider the following example. Suppose the *rwDGP* in a particular industry does not change over time (i.e., it is ergodic). Even if this is the case, we do not typically observe the entire distribution of all observations but rather a very limited set of observations—possibly only *one*, unique roll of the dice. The actual history of the industry we observe is only one of a set of possible worlds. So how do we know that the actual historical trace is in any sense 'typical' (statistically speaking) of the potential distribution? If we do not know this, then we have nothing against which to compare the distributions generated by our model. We cannot determine what is typical, and what is atypical.

5.4 Defining Sufficiently Strong Empirical Tests

Defining a strong empirical test for an AB model is a thorny problem. A common criticism of early models of technological change (in the Nelson–Winter tradition) was that they were not given sufficiently strong tests. In effect, they were evaluated on the basis of whether they could generate outputs which resemble very general, macro observations. The models were not subject to rigorous testing procedures, either on model variables or model outputs. Indeed, it was very common to find that authors had not engaged in any form of sensitivity analysis but rather provided illustrative outputs from just a handful of simulation runs.

Difficulties in defining strong tests for model outputs is highlighted by Brock (1999) discussion of 'unconditional objects' in economics (another aspect of the underdetermination problem). Empirical regularities need to be handled with care because we only have information on the properties of stationary distributions. The data that we observe does not provide information on the dynamics of the stochastic processes that actually generated them. Therefore, replication does not necessary imply explanation. For example, many evolutionary growth models can generate similar outputs on differential growth-rates between countries, technology leadership and catch-up, even though they differ significantly with respect to the behavior and learning procedures of agents, and in their causal mechanisms (Windrum 2004). Similarly, the Nelson and Winter (1982) model replicates highly aggregated data on time paths for output (GDP), capital and labor inputs, and wages (labor share in output), but these outputs can also be replicated by conventional neoclassical growth models. In the same vein, there might be many different stochastic processes (and therefore industry dynamic models) that are able to generate, as a stationary state, a power-law distribution for the cross-section firm size distribution.

Although one may be unable to narrow down a single model, we may be able to learn about the general forces at work, and to restrict the number of models that can generate a set of statistical regularities (Brock 1999). Therefore, as long as the set of stylized facts to be jointly replicated is sufficiently large, any 'indirect' validation (see Sect. 4.1) could be sufficiently informative, because it can effectively help in restricting the set of all stochastic processes that could have generated the data displaying those stylized facts. Another way out the conditional objects critique would be to not only validate the macro-economic output of the model, but also its micro-economics structure, e.g., agents' behavioral rules.[18] This requires one to only include in the model individual decision rules (e.g., learning) that have been validated by experimental/empirical evidence. Of course, this would require highly detailed and reliable data about microeconomic variables, possibly derived from extensive laboratory experiments.

5.5 Problems due to the Availability, Quality and Bias of Available Data Sets

This points us to a final core problem; the availability, quality and bias of available data sets. Empirically-based modeling depends on high quality data sets. Unfortunately, the data sets that exist are invariably pre-selected. Not all potential records are retained; some are fortuitously bequeathed by the past but others are not captured. The data sets that do exist are invariably biased. Data sets are constructed according to criteria that reflect certain choices and, as a consequence, have inbuilt biases. As econometricians know only too well, it may simply be the case that data that would have assisted in a particular discussion has simply not been collected. Such problems

[18] This point was made by John Duffy in his plenary talk at the 2005 International Workshop on "Agent-Based Models for Economic Policy Design" (ACEPOL05) in Bielefeld (Germany). Also see Gilbert (2004).

exist with data from the recent past, just as they do for data from the more distant past. Further, linking to a point raised in Sect. 4.3, econometrics is influenced by prevailing theoretical orthodoxy. As a consequence, it is very difficult to test data on new, alternative theories because suitable data are not available. The most famous example of this is the Keynesian revolution. Theoretical developments following the publication of Keynes' *General Theory* could not be tested, or put into policy practice until government agencies started to collect aggregate national data, notably on household consumption expenditure and firm investment expenditure. The data that had been collected up to this point had been informed by pre-Keynesian economic theory.

To summarize, the AB economics community has been extremely successful in developing models that address issues that are amenable using traditional neoclassical models (Dosi et al. 1994). Moreover, these AB models are able to explain how some crucial macroeconomic phenomena can be generated by the evolving networks of interactions among boundedly-rational agents in economies where the fundamentals may endogenously evolve over time. Examples range from growth and development patterns, to industry and market dynamics, to technological innovation, to the evolution of consumption and demand. What is more, they do so by taking on board methodological pluralism and avoiding the apriorist view that characterizes neoclassical economics. Having said this, there are a set of core issues that need to be addressed by the AB economics community if it is to proceed successfully. Notably, there is an excess of heterogeneity with respect to the range of competing models and a lack of consensus on core methodological questions. Drawing upon the findings of this paper, we suggest two fruitful directions. First, a commonly accepted, minimal protocol for the analysis of AB models should be developed and agreed upon (here we concur with Leombruni (2002) and Leombruni et al. (2006)). This would allow AB models to become more comparable and reach more methodologically sound conclusions. Second, far more work needs to be done to address the four core issues of empirical validation discussed in the paper. We believe that the recent trend, which seems to indicate a growing interest in methodological questions within the AB community, is an optimistic move in this direction.

References

Arthur, W. B. (1988). Competing technologies: An overview. In G. Dosi, C. Freeman, R. Nelson, G. Silverberg, & L. Soete (Eds.), *Technical change and economic theory* (pp. 590–607). London: Pinter.

Arthur, W. B. (1994). *Increasing returns and path-dependency in economics*. Ann Arbor: University of Michigan Press.

Barreteau, O. (2003). Our companion modeling approach. *Journal of Artificial Societies and Social Simulation, 6*, 1.

Brenner, T. (2004). Agent learning representation—advice in modelling economic learning. Papers on Economics and Evolution #0416, Jena: Max Planck Institute.

Brenner, T., & Murmann, J. P. (2003). The use of simulations in developing robust knowledge about causal processes: Methodological considerations and an application to industrial evolution. Papers on Economics and Evolution #0303, Jena: Max Planck Institute.

Brenner, T., & Werker, C. (2007). A taxonomy of inference in simulation models. *Computational Economics*, http://dx.doi.org/10.1007/s10614-007-9102-6.

Brock, W. (1999). Scaling in economics: A reader's guide. *Industrial and Corporate Change, 8*, 409–446.

Carr, E. H. (1961). *What is history?* London: Macmillan.

Chattoe, E. (2002). Building empirically plausible multi-agent systems: A case study of innovation diffusion. In K. Dautenhahn (Ed.), *Socially intelligent agents: Creating relationships with computers and robots*. Dordrecht: Kluwer.

Cowan, R., & Foray, D. (2002). Evolutionary economics and the counterfactual threat: On the nature and role of counterfactual history as an empirical tool in economics. *Journal of Evolutionary Economics, 12*(5), 539–562.

Dawid, H. (2006). Agent-based models of innovation and technological change. In L. Tesfatsion & K. Judd (Eds.), *Handbook of computational economics II: Agent-based computational economics*. North-Holland: Elsevier.

Doran, J. (1997). From computer simulation to artificial societies. *SCS Transactions on Computer Simulation, 14*(2), 69–78.

Dosi, G., Freeman, C., & Fabiani, S. (1994). The process of economic development: Introducing some stylized facts and theories on technologies, firms and institutions. *Industrial and Corporate Change, 3*, 1–46.

Dosi, G., Marengo, L., & Fagiolo, G. (2005). Learning in evolutionary environment, In K. Dopfer (Ed.), *Evolutionary principles of economics*. Cambridge: Cambridge University Press.

Dosi, G., & Nelson, R. R. (1994). An introduction to evolutionary theories in economics. *Journal of Evolutionary Economics, 4*, 153–172.

Edmonds, B., & Moss, S. (2005). From KISS to KIDS—an 'anti-simplistic' modelling approach. In P. Davidsson, B. Logan, & K. Takadama (Eds.), *Multi agent based simulation 2004* (Vol. 3415, pp. 130–144). *Lecture notes in artificial intelligence*, Springer.

Fagiolo, G. (1998). Spatial interactions in dynamic decentralized economies: A review. In P. Cohendet, P. Llerena, H. Stahn, & G. Umbhauer (Eds.), *The economics of networks. Interaction and Behaviours*. Berlin - Heidelberg: Springer Verlag.

Fagiolo, G., & Dosi, G. (2003). Exploitation, exploration and innovation in a model of endogenous growth with locally interacting agents. *Structural Change and Economic Dynamics, 14*, 237–273.

Fagiolo, G., Dosi, G., & Gabriele, R. (2004a). Matching, bargaining, and wage Setting in an evolutionary model of labor market and output dynamics. *Advances in Complex Systems, 14*, 237–273.

Fagiolo, G., Marengo, L., & Valente, M. (2004b). Endogenous networks in random population games. *Mathematical Population Studies, 11*, 121–147.

Frenken, K. (2005). *History, state and prospects of evolutionary models of technical change: A review with special emphasis on complexity theory*. The Netherlands: Utrecht University, mimeo.

Friedman, M. (1953). *The methodology of positive economics, in essays in positive economics*. Chicago: University of Chicago Press.

Gallegati, M., Delli Gatti, D., Di Guilmi, C., Gaffeo, E., Giulioni, G., & Palestrini, A. (2005). A new approach to business fluctuations: Heterogeneous interacting agents, scaling laws and financial fragility. *Journal of Economic Behavior Organization, 56*, 489–512.

Gallegati, M., Giulioni, G., Palestrini, A., & Delli Gatti, D. (2003). Financial fragility, patterns of firms' entry and exit and aggregate dynamics. *Journal of Economic Behavior and Organization, 51*, 79–97.

Gilbert, N., & Troitzsch, K. (1999). *Simulation for the social scientist*. Milton Keynes: Open University Press.

Gilbert, N. (2004). Open problems in using agent-based models in industrial and labor dynamics. In R. Leombruni & M. Richiardi (Eds.), *Industry and labor dynamics: The agent-based computational approach* (pp. 401–405). Singapore: World Scientific.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica, 12*, 1–115.

Kagel, J. H., & Roth, A. E. (Eds.), (1995). *The handbook of experimental economics*, (Vol. 1). Princeton, N.J: Princeton University Press.

Kaldor, N. (1961). Capital accumulation and economic growth. In F. A. Lutz & D. C. Hague (Eds.), *The theory of capital* (pp. 177–222). London: Macmillan.

Klejinen, J. (2000). *Validation of models: Statistical techniques and data availability. Proceedings of 1999 Winter Simulation Conference*. San Diego, CA.

Koesrindartoto, D., Sun, J., & Tesfatsion, L. (2005). An agent-based computational laboratory for testing the economic reliability of wholesale power market designs. *IEEE Power Engineering Society Conference Proceedings*. (Vol. 3, pp. 2818–2823), San Francisco, California USA.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: Chicago University Press.

Kwásnicki, W. (1998). Simulation methodology in evolutionary economics. In F. Schweitzer & G. Silverberg (Eds.), *Evolution und selbstorganisation in der konomie*. Berlin: Duncker and Humblot.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge: Cambridge University Press.

Lane, D. (1993a). Artificial worlds and economics, part I. *Journal of Evolutionary Economics, 3*, 89–107.

Lane, D. (1993b). Artificial worlds and economics, part II. *Journal of Evolutionary Economics, 3*, 177–197.

Leamer, E. E. (1978). *Specification searches, ad hoc inference with nonexperimental data*. New York: John Wiley.

Leombruni, R. (2002). *The methodological status of agent-based simulations*. Working Paper No. 19, LABORatorio R. Revelli, Centre for Employment Studies, Turin, Italy.

Leombruni, R., Richiardi, M., Saam, N., & Sonnessa, M. (2006). A common protocol for agent-based social simulation, *Journal of Artificial Societies and Social Simulation, 9*(1) <//jasss.soc.surrey.ac.uk/9/1/15.html>.

Liebowitz, S. J., & Margolis, S. E. (1990). The fable of the keys. *Journal of Law and Economics, 22*, 1–26.

Lucas, R. (1976). Econometric policy evaluation: A critique. In K. Brunner & A. H. Meltzer (Eds.), *The phillips curve and labor markets*. *Carnegie-Rochester Conference Series on Public Policy*. (Vol. 1. pp. 161–168). North Holland: Spring, Amsterdam.

Lucas, R., & Sargent, T. (1979). After Keynesian macroeconomics. Reprinted in R. Lucas & T. Sargent (Eds.), *Rational expectations and econometric practice*. 1981, (pp. 295–320). London: Allen & Unwin.

Mäki, U. (1992). On the method of isolation in economics. *Poznan Studies in the Philosophy of the Sciences and the Humanities, 26*, 19–54.

Mäki, U. (1994). Reorienting the assumptions issue. In R. Backhouse (Ed.), *New Directions in economic methodology*. London and New York: Routledge.

Mäki, U. (1998). Realism. In J. B. Davis, D. Wade Hands, & U. Mäki (Eds.), *The handbook of economic methodology*. (pp. 404–409), Cheltenham, UK: Edward Elgar.

Malerba, F., Nelson, R. R., Orsenigo, L., & Winter, S. G. (1999). History friendly models of industry evolution: The computer industry. *Industrial and Corporate Change, 8*, 3–41.

Malerba, F., & Orsenigo, L. (2001). Innovation and market structure in the dynamics of the pharmaceutical industry and biotechnology: Towards a history friendly model. *Conference in Honour of Richard Nelson and Sydney Winter, Aalborg*. 12th–15th June 2001.

Marks, B. (2005). Agent-based market design. Australian Graduate School of Management, *mimeo*.

Marks, B. (2007). Validating simulation models: A general framework and four applied examples. *Computational Economics*, http://dx.doi.org/10.1007/s10614-007-9101-7.

Nelson, R. R. (1995). Recent evolutionary theorizing about economic change. *Journal of Economic Literature, 33*, 48–90.

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Cambridge: Harvard University Press.

Plott, C. R., & Smith, V. L. (Eds.). (1998). *Handbook of experimental economics results*. North Holland: Elsevier press.

Pyka, A., & Fagiolo, G. (2005). Agent-based modelling: A methodology for neo-schumpeterian economics. In H. Hanusch & A. Pyka (Eds.), *The elgar companion to neo-schumpeterian economics*. Cheltenham: Edward Elgar.

Richiardi, M. (2003). *The promises and perils of agent-based computational economics*. Working Paper No. 29, LABORatorio R. Revelli, Centre for Employment Studies, Turin, Italy.

Sargent, R. P. (1998). Verification and validation of simulation models. *Proceedings of 1998 Winter Simulation Conference*. San Diego, CA.

Sawyer, K. R., Beed, C., & Sankey, H. (1997). Underdetermination in economics. The Duhem-Quine Thesis, *Economics and Philosophy, 13*, 1–23.

Schorfheide, F. (2000). Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics, 15*, 645–670.

Silverberg, G., Dosi, G., & Orsenigo, L. (1988). Innovation, diversity and diffusion: A self-organisation model, *Economic Journal, 98*, 1032–1054.

Silverberg, G., & Verspagen, B. (1995). *Evolutionary theorizing on economic growth*. Working Paper, WP-95–78. Laxenburg, Austria: IIASA.

Tesfatsion, L. (1997). How economists can get a life. In W. Arthur, S, Durlauf, & D. Lane (Eds.), *The economy as an evolving complex system II*. MA, Addison-Wesley: Santa Fe Institute, Santa Fe and Reading.

Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up. Working Paper 1, Iowa State University, Department of Economics.

Valente, M. (2005). *Qualitative simulation modelling*. Faculty of Economics, University of L'Aquila, L'Aquila, Italy, mimeo.

Vega-Redondo, F. (1996). *Evolution, games, and economic behavior*. Oxford: Oxford University Press.

Werker, C., & Brenner, T. (2004). Empirical calibration of simulation models. Papers on Economics and Evolution # 0410, Jena: Max Planck Institute for Research into Economic Systems.

Windrum, P. (1999). Simulation models of technological innovation: A review. *American Behavioral Scientist, 42*(10), 1531–1550.

Windrum, P., & Birchenhall, C. (1998). Is life cycle theory a special case? Dominant designs and the emergence of market niches through co-evolutionary learning. *Structural Change and Economic Dynamics, 9*, 109–134.

Windrum, P. (2004). Neo-Schumpeterian simulation models, merit research memoranda 2004-004, MERIT, University of Maastricht. In H. Hanusch & A. Pyka (Eds.), *The elgar companion to neo- schumpeterian economics*. Cheltenham: Edward Elgar, forthcoming.

Windrum, P. (2007). Neo-Schumpeterian simulation models. In H. Hanusch & A. Pyka (Eds.), *The elgar companion to neo-schumpeterian economics*. Cheltenham: Edward Elgar.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice, *Knowledge Engineering Review, 10*, 115–152.