



Towards a Validation Methodology for Macroeconomic Agent-Based Models

Sebastiaan Tieleman¹ 

Accepted: 30 August 2021 / Published online: 12 September 2021
© The Author(s) 2021

Abstract

Agent-based models provide a promising new tool in macroeconomic research. Questions have been raised, however, regarding the validity of such models. A methodology of macroeconomic agent-based model (MABM) validation, that provides a deeper understanding of validation practices, is required. This paper takes steps towards such a methodology by connecting three elements. First, is a foundation of model validation in general. Second is a classification of models dependent on how the model is validated. An important distinction in this classification is the difference between mechanism and target validation. Third, is a framework that revolves around the relationship between the structure of models of complex systems with emergent properties and validation in practice. Important in this framework is to consider MABMs as modelling multiple non-trivial levels. Connecting these three elements provides us with a methodology of the validation of MABMs and allows us to come to the following conclusions regarding MABM validation. First, in MABMs, mechanisms at a lower level are distinct from, but provide input to higher levels of mechanisms. Since mechanisms at different levels are validated in different ways we can come to a specific characterization of MABMs within the model classification framework. Second, because the mechanisms of MABMs are validated in a direct way at the level of the agent, MABMs can be seen as a move towards a more realist approach to modelling compared to DSGE.

Keywords Model validation · Agent-based models · Complexity economics · Economics methodology · Macroeconomics

✉ Sebastiaan Tieleman
S.B.Tieleman@uu.nl

¹ Utrecht School of Economics, Utrecht University, 3584 CE Utrecht, Netherlands

1 Introduction

Macroeconomic agent-based models (MABMs) are a promising new tool in the analysis of macroeconomic phenomena (Farmer and Foley 2009). The models do not rely on *ex ante* equilibrium assumptions which make them particularly suitable for the analysis of economic crises. MABMs have presented us with several methodological innovations compared to dynamic stochastic general equilibrium (DSGE) models. These same innovations, however, have also been the cause of critique on the use of MABMs. Most of this critique has been focused on how MABMs are validated (Fagiolo et al. 2007). The criticism is partially due to the relative novelty of the MABMs. A deeper understanding of the relationship between validation theory and the MABMs is required in order to judge the correctness of current validation practices.

The goal of this paper is to present a methodology for the validation of MABMs. Such a methodology must be able to answer the following question: How do validation practices in MABM work to enhance model validity? Furthermore, I will show that such a methodology allows us to shed light on some more fundamental issues regarding our characterization of MABMs. The structure of this paper is as follows. First, I will provide an account of the foundations of model validation in general. I will present a definition of model validation and relate this to the concept of model domain, where model domain can be seen as the scope of the model. Second, I will introduce a classification scheme based on how models are validated based on the account by Barlas (1996) and Boumans (2009). An important distinction in this classification is the difference between mechanism and target validation. The model target is the phenomenon that the model is constructed to reproduce. Model mechanisms are the relationships between model entities generating the model target. Third, I will present the framework of the structure of complex systems with emergent properties by Baas and Emmeche (1997). Connecting these three elements, we can take steps towards a methodology for the validation of MABMs. The basis of this methodology will be to analyse how the structure of complex systems relates to how MABMs are validated in practice, and from there, situate MABMs within the classification by Boumans (2009). The analysis will imply that an insightful way to look at the validation of MABMs is to consider them as modelling multiple non-trivial levels that are subject to distinct forms of validation.

The methodology of MABM validation that I will present reveals several fundamental insights into MABMs. First, it allows us to pinpoint what the mechanisms of MABMs are constituted by. The mechanisms at a lower level are distinct from, but are input to, the higher level mechanisms. Since mechanisms at different levels are validated in different ways I come to a specific characterization of MABMs within the classification of Barlas (1996) and Boumans (2009), that is, in some ways, distinct from other types of models in macroeconomics. Second, I will show that because the mechanisms of MABMs are validated in a direct way at the level of the agent, MABMs can be seen as a more realist view on modelling compared to the DSGE approach.

With this paper, I will contribute to several strands of literature. First there is the literature that seeks to explicate some of the general issues that modellers run into the validation of MABMs, as well as discuss the up- and downsides of the different validation methods used in practice. The first publications in this series were Fagiolo et al. (2005) Fagiolo et al. (2007) and Windrum et al. (2007). Later, important updates followed to discuss new developments (Fagiolo and Roventini 2017; Gatti et al. 2018). In this series of papers the most commonly used validation approaches are put forward. In Fagiolo et al. (2005) the main approaches are qualitative simulation modelling, replication of stylized facts, empirical calibration and the history friendly approach. In qualitative simulation modelling, the relationship between the behaviour of the model and empirical data is only required in a qualitative dimension. That is, as long as the model behaviour is roughly in line with some qualitative empirically observed features, the model is considered valid (if A increases than B also increases etc.). Such models are most applicable for exploratory and experimental purposes. Replication of stylized facts is the approach in which the model is considered valid if it is able to reproduce a set of relevant (given the model purpose) stylized facts. Importantly, all of the model parameters are calibrated indirectly, meaning that they are quantified such that the model is able to reproduce the set of stylized facts. The empirical calibration approach is similar to the replication of stylized facts approach, in the sense that stylized facts are used to calibrate the model parameters. In addition, however, some of the parameters are calibrated directly. This entails that the empirical data used for the calibration concerns the individual relationship in which the parameter occurs, instead of comparing the output of the model as a whole. In the case of agent-based models, most of the parametrization occurs at the micro level, implying that, in direct calibration, empirical data at the micro level are used. This type of validation is considered to be a more strict type of validation. Finally, the history friendly approach, in which the validation criterion is to reproduce a precise data history, at least in qualitative terms. Often, this comes in the form of reproducing a specific time series. In this approach, there is an additional importance of matching the initial conditions of the model to the one's observed in time series to be reproduced.

Most recently Fagiolo et al. (2019) has been published, which provides us with the latest developments regarding validation techniques of MABMs. This includes the use of machine learning techniques to select regions of the model parameter space that exhibit interesting behaviour. Such computational techniques have become increasingly important as agent-based models have become more elaborate, and the increasingly important criterion of sensitivity analysis. Sensitivity analysis is an assessment of how robust certain model behaviour is to changes in the parameter space, or to changes in assumptions. An additional interesting development that is discussed in Fagiolo et al. (2019) is the validation approach described in Guerini and Moneta (2017). In this approach, a VAR-estimation is performed on the aggregated (macro) variables of the simulated data (generated by the agent-based model). The coefficients of this simulated VAR-estimation are then compared to the coefficients of the same VAR-estimation performed on empirical data. If the coefficients match in terms of their sign and, to some extent, their size, the model passes the validation test. This is an interesting approach because it deals with what

is known as the conditional object critique originally put forward by Brock (1999). The core of this critique is that stylized facts may be too general such that multiple distinct models may be able to reproduce these stylized facts. The approach in Guerini and Moneta (2017) provides a more selective validation measure within this context.

The publications discussed above have been critical in providing overviews of validation approaches in MABMs in order to build a more standardized approach and are important in making clear in which directions new research should develop. This strand of literature, however, has not sought to provide in depth methodological foundations for the methods they present. It does list an overview of some deeper methodological issues in model validation, such as realism versus instrumentalism and underdetermination [see for example Windrum et al. (2007)], but does not go the step further and look at how these issues specifically apply to MABMs. This is, in turn, precisely what this paper aims to do.

An alternative angle to consider the validation of MABMs is to construct a benchmark model, the performance of which, serves as a minimum criterion for model validation. In a.o. Caiani et al. (2016) and Lengnick (2013), for example, the aim has been to present a benchmark or baseline model. The difficulty with such a common benchmark may be that actual validation criteria may differ given various different questions models are built to answer. Benchmark models, therefore, seek to incorporate only features that are seen as essential for most models. If we look at Lengnick (2013), for example, we see that the incorporated features include households, firms and banks that behave and interact according to simple rules. Moreno et al. (2019) does not consider particular models but rather looks at setting up a benchmark for various agent-based model software platforms. It looks at which features are typically deemed essential for an agent-based model, such as the ability to incorporate interaction rules and to simulate large numbers of agents. By implementing these features in various agent-based modelling platforms, the performance is assessed by looking at the computational performance (such as execution time) of the platforms. An interesting avenue for future work would be to extend the approach of Moreno et al. (2019) from agent-based modelling platforms to agent-based models.

The literature discussed above shows that there are a variety of approaches to model validation of macroeconomics agent-based models. Some approaches may be more suitable for particular types of models than others. For the purpose of this paper I will focus on the stylized fact (or indirect calibration) approach, because this is the approach that is most frequently used in practice (Fagiolo et al. 2019), and because more recent approaches such as Guerini and Moneta (2017) are ultimately extensions of this approach.

In important building block in the methodology I will present is the literature that relates the type of model to the type of validation. This literature originates in the system dynamics modelling literature with Barlas (1996) as one of its most relevant publications. Boumans (2009) has specified this way of relating a model's purpose and validation further by applying it to models in economics. This literature is rather unique in that it explores validation in the context of model purpose. I will use this

classification to characterize what type of model MABMs are and what this implies for their domain of validation.

As we will see, the notion of complex systems and complexity economics plays an important role in the analysis of this study. My understandings of the economy as a complex system have been strongly influenced by works such as Arthur (2013). Furthermore, the concepts of reductionism and emergence in the context of macroeconomic modelling will be relevant. Hoover (2015) and to some extent Gatti et al. (2011) are important contributions in this regard. This study will contribute by connecting the concepts of complexity and emergence to model validation.

Finally, this contribution stands within a strand of literature looking at the fundamental methodology of agent-based modelling. Most well-known are Epstein (1999) and Epstein (2006), in which the idea of agent-based models as tools to generate explanation is brought forward. This also links to validation since a necessary condition within this concept is that agent-based models are valid only if they are able to generate an explanation starting from interacting agents. Furthermore, contributions such as Elsenbroich (2012) and Grüne-Yanoff (2009) have helped me to gain a deeper understanding of agent-based methodology in relation to explanation. An analysis of validation in the light of the fundamental methodology of agent-based models is not present in the current literature, however, which is where I hope to contribute.

2 Model Validation

First, it is helpful to discuss model validation in more general terms, in order to start the analysis on equal footing. I will give a definition of model validation and explain how model validation is given by the purpose of the model. Second, I will introduce the concept of model domain to show under which conditions validation tests enhance model validity.

2.1 Defining Model Validation

Before we are able to understand what model validation entails I should first discuss, briefly, a bit about models in general. There are many ways to consider models in science. The viewpoint presented here is roughly based on Morrison and Morgan (1999), because I consider it a useful way to understand MABMs. Models can be viewed as instruments that are constructed for a specific purpose. This purpose can be to measure, to predict or to provide understanding of some sort. To the extent that models are instruments to investigate real world phenomena they must also latch on the real world in some way. Models accomplish this by being a representation to some extent. Models are useful tools for investigating phenomena because they represent some aspects of the real world in some way. Models, however, are also necessary simplifications and idealizations of reality. It is the idealization and simplification that allows the model to provide us with enhanced understanding compared to looking at the world directly. Models can give us insights into the world because they connect to it, but are at the same time also

independent from it. A central issue in modelling is, therefore, how we should construct models in such a way that they include necessary representational elements while still being useful instruments. Whether a model is correct in this respect can only be assessed relative to the purpose of the model.

In economics, a way in which models are often used is for a so-called computational experiment, as was put forward by Kydland and Prescott (1996), where models are seen as computational tools to answer “well-posed questions”. In fact, posing a question is the first step in the modelling process. If the question is of purely quantitative nature, the model can be constructed as a measurement or prediction instrument. If the question asks for some type of explanation, the mechanisms by which the model generates certain output are of interest. For this type of models, the tension one is faced with in model construction is that the modeller has to determine which mechanisms can be left out and which can be abstracted in such a way that the model is still able to accurately answer the question without being so complex that the answer the model provides to the question does not enhance our understanding.

The above description of what a model is provides us with the following definition of model validation: If a model is a tool to answer a question, then, model validity is the assessment of the model’s capability to provide a correct answer to this question. Later, I will go into more depth regarding which sort of validation is appropriate given particular kinds of model purpose. Validation is often a broad assessment, in which various types of criteria play a role, including theoretical, mathematical and empirical criteria. The methodology put forward in this paper concerns mostly the empirical criteria of the validation process. Concretely, empirical validation means the assessment of a model’s answer by its correspondence to relevant empirical data. From here on, when I refer to validation I mean empirical validation.

2.2 Model Domain

Let us now look a bit closer at when empirical data may be considered relevant for a specific model validation exercise. Model domain can be seen as the scope of the model. It is determined by what the model is intended to do and limited by what it is not to do. Looking back at our definition of what a model is in the context of a computational experiment, the model domain is thus given by the question the model is intended to answer. Imagine empirically observable reality as a space. This space contains all empirical data that could potentially be used in any model validation exercise. The question to ask is what subset of this space enhances a model correctness, if reproduced by the model. Given our definition of validation this can only be assessed relative to the question the model is built to answer. The question determines which subset of this space is relevant, and thus enhances a model’s correctness. This subset is the model domain. It contains the parts of the question’s answer that can be verified or falsified using data. Let us consider an example in which a model is built to answer a question regarding the mechanisms generating the business cycle. The domain of this model is constituted by the model target: the business cycle, and the mechanisms that generate this business cycle.

Empirical reflection of the business cycle and the model mechanisms generating the business cycle, can be used in model validation. Data unrelated to the model target or model's mechanisms will not enhance the model's validity. It is thus important that the information we are using to validate our model is relevant to the model's purpose. As long as information is relevant for this purpose, however, there is an incentive to use as much information as possible. This is related to the notion of underdetermination (Stanford 2009). Underdetermination is a fundamental problem of science. Contrastive underdetermination entails that there are always multiple theories or models that are able to explain a certain set of data. For validation this implies that for any validated model we can always come up with a different model that is also able to pass the same empirical validation tests. The more relevant information we use in testing the model, however, the more we are able to reduce underdetermination. Model validation can thus be seen as reducing underdetermination within the model domain.

3 Black-, White and Grey-Box Models

We have seen how model validity can only be assessed given the purpose of the model. In this context, Barlas (1996) presents us with a classification of model types given different forms of empirical validation. Later, this account was extended in Boumans (2009) to specifically apply to the type of validation that is often observed for economic models.

First, it is important to clarify a bit further the concepts of model target and model mechanisms. Any model is constituted by both mechanisms and a target. The target of the model is the phenomenon the model is intended to reproduce. In economic models this is usually some economic phenomenon observed empirically. Examples are business cycles, economic growth or competition dynamics. The model mechanisms are, broadly speaking, the relationships between model entities by which the target is generated. These mechanisms could be in the form of equations, but also arrows in diagrams or entities interacting in a computer simulation. Models differ with respect to how and whether the target and mechanisms are validated. Following Barlas (1996) we can identify three types of validation tests. The first are behaviour pattern tests. These tests entail comparing the model target with its empirically observed counterpart. From here on, the use of this type of tests will be labelled simply as target validation. Next, are direct structure tests. These types of tests consist of comparing empirically the individual relationships between model entities that make up a model's mechanisms. An example of such validation is seeking unbiased estimation of the parameters of these relationships. I will label these type of tests as direct mechanism validation. Finally, there is a third group of validation tests known as structure-oriented behaviour tests. This type of tests validates the model mechanisms, but only indirectly through the output the model's mechanisms generate together. An important example of this type of testing is the Turing Test, in which we present the model with questions that have well-known answers. If the model is able give the right answers in response to these questions, it is evidence that the models mechanisms are, in some sense,

correct. An example of a Turing Test in economics is how DSGE models are validated using impulse response functions. We shock the model in several ways and look how the model behaves. We can then compare this to how the economy behaves empirically if it is presented with such shocks. I will refer to this type of model validation as indirect mechanism validation.

Barlas (1996) makes a classification of models types based on which of the above three types of validation are used. Barlas (1996) distinguishes between two types of models: white-box and black-box models. Boumans (2009) extends this classification to include grey-box models. Grey-box models, as we will see, are of particular relevance for models in economics. The three types I will consider are, therefore, white-, black- and grey-box models. First, in white-box models all three types of validation tests are appropriate. The model target is validated, and the mechanisms are validated directly and can be validated indirectly in addition. In such models the mechanisms of the model are judged as a causal-descriptive account. Looking back at our notions of model domain, white-box models give answers to “why-questions” (Boumans 2009). Second, there are black-box models. In this type of models the mechanisms are not of interest and, therefore, not subject to validation. The only type of validation appropriate is target validation. In black-box models we “just” want a model to give the correct output and we do not care whether the mechanisms by which it arrives at this output are in any sense related to the mechanisms operating the real world. Third, we have grey-box models. In this type of models the mechanisms are of interest and are, in turn, subject to validation. These mechanisms are, however, only validated indirectly, through, for example, Turing Tests. This type of model validation does not require each model entity relationship to correspond to its empirical counterpart individually as long as the model entities together generate a range of empirically valid outputs. Importantly, however, it also does not exclude the possibility that these relationships are also individually empirically valid. This is also why, in white-box models, both direct and indirect mechanism validation can be used. Generally, we could say that a model that has passed direct mechanism validation tests should also pass indirect mechanism validation tests. The converse, however, does not necessarily hold. Why this is the case relates to views regarding instrumentalism and realism and is something I will come back to extensively later in this paper. In Table 1, we can see a schematic overview of this classification. Note how model classification is given by which types of validation tests are appropriate.

One may wonder for which type of models only indirect mechanisms validation would apply. As Boumans (2009) points out, typically, grey-box models have a so-called modular structure. Following Simon (1962), models that have a modular structure are built from several, in some sense, autonomous parts or submodels. The reason for modelling in this way is that it might be a daunting task to model directly all of the mechanisms that are at work in a larger system. Instead, we can seek to partition the system into smaller subsystems, or modules. If we put these modules together we have a model of the larger system. These modules may interact in complex ways that make it difficult to validate the relationships between model entities individually, which leaves us with indirect mechanism validation.

Table 1 Model validation classification based on Boumans (2009)

	Direct mechanism validation	Indirect mechanism validation	Target validation
White-box model	✓	✓	✓
Grey-box model		✓	✓
Black-box model			✓

4 Complex Systems with Emergent Properties

Let us now discuss the third element necessary to come to a validation methodology of MABMs. MABMs, generally speaking, are tools to model the economy as a complex system. But what does this mean? There are many ways to define what a complex system is. A definition that is useful for our analysis comes from Ladyman et al. (2013): A system in which elements react to the patterns they together create through interactions. So according to this definition, a complex system is one comprised of a multitude of interacting elements. Through these interactions they create patterns the elements, in turn, react to. This implies that the created patterns are not always consistent with individual behaviours, which, in turn, implies that the natural state of the system is not, at least ex ante, equilibrium (Arthur 2013). This discrepancy between patterns at different levels is known as emergence. Emergence is generally a key feature of complex systems. There are many ways to define what emergence is exactly (O'Connor and Wong 2015). In order to analyse properly the role of emergent properties in models of complex systems, I start from the framework introduced by Baas and Emmeche (1997). In this framework a system is defined in terms of structures S^n , interactions Int^n and an observation mechanisms Obs^n . n represents the level of interactions we are considering. Structures at a certain level are constituted by the lower level structures and their interactions. Formally, levels are connected in the following way:

$$S^{n+1} = M(S^n, Int^n, Obs^n) \quad (1)$$

where M can be considered as some process of mathematical induction. Each of these levels of structures can have properties. Emergence can be defined through the following definition regarding these properties. A property P is emergent if $P \in Obs^{n+1}$ while $P \notin Obs^n$. This means that a property of a certain structure is emergent if it is not observed as a property of the lower-level components that constitute this structure in isolation. Isolation in this sense means that we do not take into account the implications of the interactions of structures on each other. This also relates to the observation mechanisms Obs . Looking at a structure at a level n in isolation means we are looking at the structures through Obs^n .

Let us now see how this structure is reflected by models of complex systems such as MABMs. Such a model always starts from a base level $n = 1$. This level is the input to the model and consists of the elementary structures of the model. All model assumptions are at this level. We can generate the next level in such a model by

taking into account the implications of interactions between these structures S^1 . Considering the interactions between structures of this model implies that we use some methods of mathematical induction, such as iteration, to generate the structures at the next level. These structures will, in turn, interact to generate the next level structures and so on. If a structure at a level higher than $n = 1$ has properties that can only be observed by considering the implications of interactions of lower level structures, we can say a structure has emergent properties. For micro-founded modelling macroeconomics, in which the micro level can be seen as $n = 1$, this implies that, contrary to the “strong reductionist” (Gatti et al. 2011) approach of the representative agent, macro properties do not reduce to properties of individual households or firms in isolation. Rather, we require the interactions between agents from which macro properties emerge that are qualitatively different than the properties of the agents.

Most MABMs can best be described as three level systems: the micro level ($n = 1$), the meso level ($n = 2$) and the macro level ($n = 3$). Interactions at the micro level give rise to a second order structure which consists of networks of agents. These networks, in turn, can interact resulting in the macro level and third level structure. Levels are defined by whether there exist interactions between structures at the same level. It is, however, important to consider that in reality the boundaries between the meso and macro level are much more blurry, than the framework of Baas and Emmeche (1997) would consider them to be. There are interactions and feedbacks between different levels, which can make a level in between the agent and the aggregate model output difficult to set apart.

5 A Validation Methodology of Macroeconomic Agent-based Models

We now have now discussed all elements necessary to come to validation methodology of MABMs. First, I will show how current validation practices relate to the structure of MABMs as a complex system as discussed in Sect. 4. Second, I will characterize MABMs according to how they are validated in practice, in terms of the classification scheme discussed in Sect. 3.

5.1 Validation at Multiple Levels

Let us now look at which validation tests, observed in practice, apply at the micro, meso and macro level. As stated before, the distinction between the meso and macro level is, in practice, often not so easy to make. I still think it is useful, however, to consider the meso level to better distinguish between different types of stylized facts. Importantly, the core of the analysis would remain unchanged if we were to consider a MABM as a two level system. In what follows, I will relate the concept of emergent properties to validation in practice. I will do so by looking at how the properties of the micro, meso and macro level are validated. As examples of validation in practice, I will refer to Lengnick (2013) since this paper is, to some extent, representative for the general validation practices in the MAMB literature.

5.1.1 The Micro level

Agents are the structures at the $n = 1$ level. They are the lowest level structures of a MABM. This implies that they are the model input. They have properties that directly result from the assumptions made by the modeller. These properties can be tested empirically in order to conduct validation at the micro level. In Fig. 1, we can see a schematic representation of validation at the micro level. In this representation, firms could be seen as squares while agents could be seen as circles. Note that the squares and circles are not connected in any way. This is because we are looking at the system through the Obs^1 lens; we are not considering the implications of agent behaviour on each other. P^1 are the properties observed through Obs^1 .

What does validation at this level entail? The assumptions at the micro level are considered to be of particular importance to MABM practitioners. MABM papers often start with the critique that DSGE models are not based on realistic assumptions. In turn, MABM modellers pride themselves on modelling from more realistic assumptions. But what does it mean in the case of MABMs to have more realistic assumptions? MABM practitioners state that one of the crucial differences between DSGE models and MABMs is in the realism of the assumptions that determine agent behaviour. The following quote is one of the essential differences between neoclassical economics (in which DSGE are one of the main tools) and agent-based computational economics (ACE):

ACE can be seen as a substitute to standard neoclassical approaches to economics that tries to build more reasonable models based on reality to better address its behaviour, a new approach that rejects the idea that models can be built using false assumptions and trying instead to explore models based on assumptions more in line with what we know about how real-world agents behave and interact. (Gatti et al. 2018)

MABMs are thus an attempt to move from the homo oeconomicus towards a more empirically validated agent. In most MABMs papers, therefore, specific attention

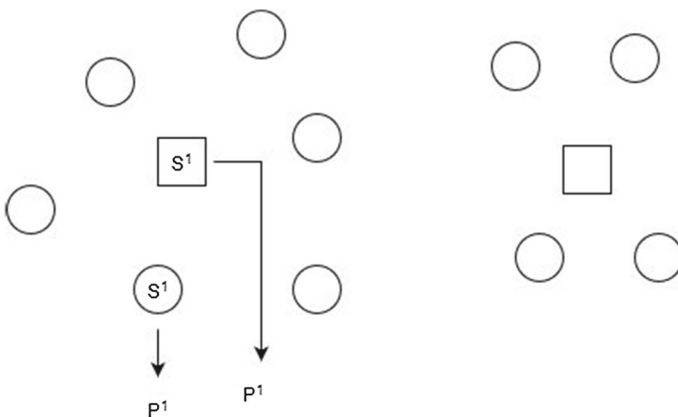


Fig. 1 Validation at the micro level

will be paid to validation at the micro level, in which the assumptions regarding the behaviour of agents are compared to outcomes of economics experiments or other insights from psychology or strategy literature. Most often these rules will imply a certain form of limited knowledge regarding the economic system, compared to the rational expectations approach of DSGE. This is in line with the notion of procedural rationality (Simon 1976) and heuristics (Gigerenzer and Todd 1999). In the schematic overview, properties P^1 at the micro level are such behavioural rules.

Looking at how agent validation takes place in Lengnick (2013), we can observe a similar way of agent validation. For example, to describe consumption behaviour, the following equation is used:

$$c_h^r = \left(\frac{m_h^\alpha}{P_h^r} \right) \quad (2)$$

where c^r is the individual household consumption, m^h are the monetary holdings of the household and P_h^r is the average price of the producers the households buys from. $0 < \alpha < 1$ such that the relative share of income that is consumed decreases when the household's monetary holdings increase. Importantly, the functional form of this equation is defended not on the basis of homo oeconomicus theory but rather by citing an empirical study, namely Souleles (1999). For other domains of household behaviour and firm behaviour similar sources are cited which are either empirical studies or theories that are directly supported by micro data. It is useful to note that in neoclassical economics such empirical validation at the agent is generally not present.

5.1.2 The Meso Level

Let us now consider how properties at levels higher than the micro level relate to validation as we observe it in practice. In addition to validation at the micro level $n = 1$, we observe the validation of cross-sectional properties as well as macro variables over time, or relationships between macro variables. The use of cross-sectional properties can best be characterized as validation at the meso level or $n = 2$. We can label these properties as P^2 . S^2 arises from the first-order interactions between, for example, firm agents and their consumers. A property of these structures is the distribution of firm size, as these structures will vary in terms of how many links they have with consumers. These distributions can then be compared to empirical data to perform empirical validation. In Fig. 2, we can see a schematic overview of this. The second-level structures are given by the units represented as the big the circles drawn around the network of squares and circles. These could represent a network of two groups of consumers and a firm. In Lengnick (2013), we observe such use of cross-sectional properties in validation. For example, model data on the distribution of firm size is used. The model data follows a right-skewed distribution, more specifically, a power law. Practically, it means that in the model data there is a large group of small firms and a small group of very large firms. In Lengnick (2013), these model data characteristics are compared to the empirical data on firm size and it is concluded that the model data

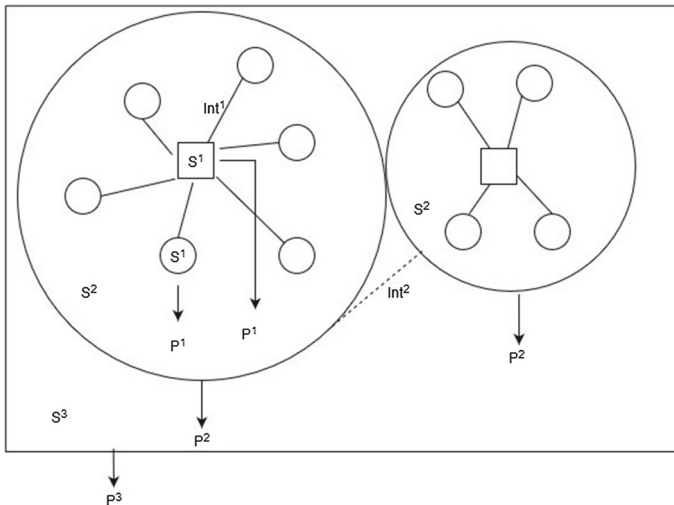


Fig. 2 Validation at different levels

and empirical data are distributed in a similar way. In addition, Lengnick (2013) uses model data regarding the distribution of the price changes for validation. Validation of these distributions can be seen as validation of the meso level. The reason for this is that, especially in the case of firm size, these distributions are the product of interaction between agents rather than the product of interaction between networks of agents. Firm size is determined by its consumer network and the distribution of these consumer networks can thus be seen as a property of S^2 observed through Obs^2 .

5.1.3 The Macro Level

Finally, we can look at the properties at the macro level S^3 , with properties P^3 . These usually consist of patterns of aggregate variables over time or relationships between macro variables. In Fig. 2 we can see a schematic overview of validation at the micro, meso and macro level. The square around the groups of networks represents the model as a whole. Properties at the macro level are different from the properties at the meso level, because for the macro level we take the interactions Int^2 between S^2 into account which yield S^3 . If we look at how validation in practice is conducted at this level in Lengnick (2013), we observe the use of empirical regularities in the form of relationships between macro variables. Examples are the relation between price level and employment, better known as a Philips curve and the relationship between vacancies and unemployment, a Beveridge curve. For these relationships the model data are compared to empirical data in order to conduct the validation. In addition to looking at relationships between macro variables, we can also look at the characteristics of macro variables over time. An example of this is the unemployment level over time. In Lengnick (2013), model unemployment data

alternates between low unemployment and deep downturns. This is in line with what we observe empirically regarding the business cycle.

Again, whether there exists a distinction between the meso and macro level depends on whether the properties of the structures at the meso level can be obtained without considering the interactions between these meso structures, and whether the properties at the macro level require interaction between the meso structures in order for the model to generate these. A well known example of this would be a skewed distribution of firm size that arises out of the interactions between firms and consumers, and a business cycle that arises through the default of one large firm which affects smaller firms because of a decrease in wage income of the defaulted firm's employees (Gabaix 2011).

5.2 Mechanism and Target Validation in MABMs

Thus far, we have seen how MABMs are validated at several emergent levels. In order to situate MABMs within the white-, black- and grey-box framework we have seen in Sect. 3, however, it is important to relate such validation practices to the different forms of validation discussed before. How does the validation at multiple levels apply to mechanism and target validation?

First, it is important to look at what actually constitutes the mechanisms of an agent-based model. Broadly speaking, the mechanisms of a model are processes between model entities that contribute in some way to generating the model target. Given that MABMs are structured as multiple levels, mechanisms operate at multiple levels as well. In order to understand this, let us look again at the framework by Baas and Emmeche (1997) and clarify a bit further the role of the observation mechanisms *Obs*. *Obs* states at which level we are observing *S* and *Int*. The difference between a lower and higher level of *Obs* is given by whether or not we take interactions into account. But what does this mean exactly? Let us consider a two-level example starting from the micro level $n = 1$:

$$S^2 = M(S^1, Int^1, Obs^1) \quad (3)$$

Now, if we put on the *Obs*¹ glasses, we observe the behavioural rules of agents as properties of *S*¹. Embedded in these rules are the mechanisms that generate output at the level of the agent, without considering interactions. Importantly, however, these rules can, and will in practice, also include rules that are a function of variables of other agents. It could be, for example, that a firm sets its price as a mark-down on the average price of other firms. These types of behavioural rules can still be observed through *Obs*¹ and embed the mechanisms at the first level. Such a rule does in the definition of this framework not take into account the interactions, because interaction requires taking into account the implications of behavioural rules of agents on each other. In the price setting example, taking into account the interactions would require taking into account the price setting behavioural rules of other firms as well. At the level of the agent in isolation observed through *Obs*¹, influences of other agents are exogenous. If we take into account interactions, however, we observe, through *Obs*², a system of connected agents such that

exogenous variables at the micro level are now endogenous in relation to the second order structures. This system observed through Obs^2 will thus be constituted of mechanisms that take into account the effects of agent's behaviour on each other. We must note at this point, that the difference between observed levels is thus "merely" one of steps in a mathematical induction. This is what is also implied by the fact that S^2 is generated through an inductive process M . The reason for this is that the interactions Int do not constitute any new input to the model, but are, rather, an inductive consequence of the behavioural rules that are part of S^1 . Finally, it is important to see how the notion of emergence fits into this distinction between levels. We have defined emergence as $P \in Obs^{n+1}$ while $P \notin Obs^n$, which implies that the properties of S^2 cannot be observed without taking into account the implication for agent's behaviour on each other. In the case of mechanisms, this means that by taking into account interaction, mechanisms will emerge that are constituted by, but different from, the behaviour of agents observed through Obs^1 . If a property was non-emergent, it would mean that it could be generated from the mechanisms observed through Obs^1 . Given the number of heterogeneous agents in MABMs, the inductive steps that allow us to observe a new level can, in the case of emergent phenomena, only be taken taken through simulation. It is, however, essential to understand that the different levels of mechanisms in MABMs are all inductible from the same input. This also means that all of the mechanisms observed at any level are not independent from the mechanisms observed through Obs^1 . Dawid and Gatti (2018) also come to this conclusion by showing that any macro property can be derived through mathematical induction.

This leads us to an interesting conclusion regarding how mechanism validation takes place in MABMs. First, the mechanisms observed at Obs^1 are validated directly in the sense that agent behaviour is modeled based on empirical evidence related to agents, in the way I have shown in the previous section. This agent behaviour observed through Obs^1 , also constitutes the input for the mechanisms observed through Obs^2 , the system of interacting agents. These mechanisms, contrary to the mechanisms observed through Obs^1 , are usually not observed directly by the modeller. The reason for this is that MABMs can, in practice, only be analysed through numerical simulation. This is due to the large number of heterogeneous agents engaging in non-linear interactions. Mechanisms that are not observed directly can only be validated indirectly through, for example, Turing Tests at both the cross-sectional and time series level. In the previous section, I have shown examples of such tests. Importantly, both the tests of properties at the meso level as well as the macro level can be seen as Turing Tests of the mechanisms as long as they are not the model target. The model target in the example of the previous section were the observed business cycle dynamics. Looking back at the classification by Boumans (2009), we observe that MABMs are thus validated as grey-box models. Importantly, MABMs have a modular structure. The modules in an MABM are the agents. Specific to MABMs is that the modules are validated directly which means we can view these modules as white-box sub-models. A MABM can, within this classification, be seen as a grey-box model built from white-boxes.

Now, how does this combination of white-box validation within grey-box validation enhance model validity of the model's mechanisms? If the relationship between multiple levels is merely one of inductive steps, does this mean validation at higher levels is superfluous? The answer is no for the following reasons. It is useful to think about this in terms of reducing underdetermination, and consider the notion of model domain of Sect. 2. When modelling agent behaviour, the behaviour of the agent is constraint by empirical data through direct comparison of agent behaviour with empirical evidence related to agents. This does, however, not mean that choices regarding agent behaviour do not have to be made. First, modelling is always a matter of simplification and isolation. The modeller has to decide which elements of agent behaviour are significant in relation to the model target and which elements can be left out. Second, underdetermination can only be reduced to some degree, since there will always be degrees of freedom and several competing models of agent behaviour even if we apply strict empirical testing. The above facts are important to state, since they imply that indirect testing of mechanisms at Obs^2 will still enhance the validity of the model's mechanisms, because they further reduce the number of possible specifications of agent behaviour that are in line with empirical reality. The modeller makes choices regarding agent behaviour when modelling the agent. The innovation that MABMs have brought is that these choices are constraint by empirical data regarding the behaviour of the individual agent. The choices the modeller makes, however, also determine the mechanisms at higher levels of observation. By simulating the model the implications of the choices made at the level observed through Obs^1 , for the level observed through Obs^2 , can be uncovered and can be indirectly validated. This further constraints the choices the modeller can make regarding the model input. The role of emergence in this process is thus merely one of veiling the implications of choices at one level for the next level. Mechanisms at higher levels can only be uncovered through some inductive process, which in the case of MABMs, is numerical simulation. The simulation methodology implies that the generating these implication mechanisms can only be validated indirectly through output characteristics. If there was no emergence relationship between two level it would mean that there are no veiled implications for the choices of one level for the next. Validation at these levels, in such a case would therefore require the same data, which would mean that validating at multiple levels would not help to reduce underdetermination.

6 Realism and Instrumentalism

We have seen that MABMs are structured as grey-box models built from white-boxes. This has helped us to gain an understanding how certain validation tests enhance validity of a model's mechanisms and target. We can now go one step further and look at what such a validation structure could imply for how we should characterize MABMs. More specifically, how we should characterize MABMs regarding instrumentalist or realist view of modelling. It is useful to look into this because it will help us gain better insight into how MABMs can actually enhance our understanding of the world. Dawid and Gatti (2018) also regard discussions on

instrumentalism and realism as vital in understanding validation of MABMs. In the argument I will make, the difference between an instrumentalist vision and realist vision is tied to how we characterize a model's mechanisms. In pure instrumentalist visions on modelling, the correctness of the mechanisms is given solely in terms of the model output. As long as the mechanisms together generate some required model output the mechanisms are correct. At the other extreme, in a purely realist view, the correctness of mechanisms is given by their degree of homomorphism to the causal structure in the real world. What I will do in this section, is explain how these notions of an instrumentalist and realist vision towards modelling relate to the classification black-, white- and grey-box models. From here, I will compare mainstream DSGE models to MABMs in order to compare these two approaches to modelling in terms of their instrumentalist or realist vision.

So let us first look at how black-, white- and grey-box models relate to instrumentalism and realism. Black-box models are only validated in terms of how well they match the target of the model. A model that is to describe the business cycle would thus only be validated in terms of how well the model generates such a cycle. In such a model, the mechanisms are not of interest and are thus not validated in any way. Depending on the question the model is built to answer, such models are usually associated with an instrumentalist view on modelling since their correctness is only tied to output. White-box models are validated in relation to the target. The mechanisms in the form of relationships between entities are validated individually, usually through estimation of these individual relationships. Unbiased estimation of such individual relationships implies causality, which is why white-box models are usually associated with a more realist view on modelling. Grey-box models are validated in terms of their target, but are also subject to indirect tests of mechanisms, indirect meaning here that the mechanisms are judged through certain output of the model, that the mechanisms generate in concert. Grey-box models relate to instrumentalism in an interesting way.

Grey-box validation implies that we are in fact interested in the mechanisms, but do not validate them in a way that requires them to represent the causal structure of the real world. Rather, we require them to operate in a way that mimics this causal structure in terms of a wider range of outputs. This is important since an understanding of a grey-box model can provide an understanding of the workings of the economy without knowing the real world causal mechanisms at play. In this way, grey-box models can for example inform policy interventions. It is useful to note that an instrumentalist view does not exclude the real world causal structure to be captured by the mechanisms in the model, they just are not the criterion. DSGE models are grey-box models since their workings are mainly validated using impulse response functions. If the model is able to reproduce the correct response to a variety of economic shocks the model is deemed correct. Grey-box models generally speaking have a "modular" structure, meaning that they are comprised of multiple sub-models the interaction of which is part of the model's mechanisms. In DSGE models these modules could be seen as consumer and firm behaviour. These modules in the case of DSGE, are only validated in so far they are useful for generating correct output at the level of the model as a whole. These modules, in the

case of DSGE models can, therefore, be considered black-boxes. DSGE models can be characterized as grey-boxes built from black-boxes.

As I have described thoroughly in the previous section, MABMs are grey-box models, but contrary to DSGE models, are built from directly validated modules, making these modules white-boxes. How should we consider such a model in relation to instrumentalism versus realism? It is helpful to look at this issue in terms of reducing underdetermination within a model's domain, as I have discussed in Sect. 2. In an instrumentalist view, one would seek to reduce the possible models by looking at which model performs better in terms of output generated by the model as a whole. In MABMs, through validation at the micro level, however, we are eliminating only potential models whose mechanisms, that take us from the micro level to the macro level, would only be correct from an instrumental perspective (in terms of model output). The reason for this is that any model that starts from agent behaviour and that does not isolate elements of agent behaviour that can be fitted to the empirical information we have about agents, can necessarily only be correct in an instrumentalist sense. Or, to put it differently, the mechanisms in the real world macroeconomy must in some sense be derived from the behaviour of real world economic agents. As discussed in the previous section, however, the converse is not necessarily true. Because of the fundamental problem of underdetermination there will always exist multiple agent-specifications that fit the information we have about agents. Therefore, there will always remain a possibility that an MABM validated at both the agent and the macro level, contains mechanisms that are only correct from an instrumentalist perspective. Again, put differently, it is possible that within the degrees of freedom we have in modelling agent behaviour, we make choices that lead to a model whose mechanisms are not correct from a realist perspective, but still manages to be correct instrumentally. The exercise of validation at the micro level, however, reduces the space of instrumentalist mechanisms. This increases the validity of the model from a realist perspective, compared to, for example, DSGE models. Therefore MABMs present a shift towards a more realist perspective on how to model economic phenomena.

7 Conclusion

In summary, in order to understand validation practices conducted in MABMs, I have introduced three elements. First, I have defined model validation and introduced the concept of model domain. The analysis has shown that a model's purpose creates a limited model domain. A model can only be correctly validated within this domain. Second, I presented a classification of three types of models based on how these models are validated. We can distinguish black-box models in which only target validation applies, white-box models in which both the target and the mechanisms are subject to validation in a direct sense, and grey-box models in which the mechanisms are subject to validation in an indirect sense. This classification serves to understand how we should consider the purpose of validation tests used in practice. Third, I have looked at how models of complex systems with emergent properties are structured. Complex systems can be seen as systems with

different levels of interacting structures. Properties observed at one level that are not observed as properties at a lower structure, can be defined as emergent properties. MABMs can also be understood as systems with multiple levels of structures; the micro level, the meso level and the macro level.

The combined insights of these three elements allows us to formulate a methodology of the empirical validation of MABMs, that provides a deeper understanding regarding how certain forms of validation observed in practice relate to mechanism and target validation. First, I have shown that in practice, each of the levels (micro, meso and macro) of MABMs are validated using different tests. The micro level is the input to the model. Specific to MABMs is that this input is subject to validation. This means that the agent behaviour should be similar to the agent behaviour we observe empirically in relevant domains. The meso level is validated by comparing model output to the distributions of aggregate variables. Finally, properties at the macro level are validated by comparing model output to behaviour of aggregate variables over time or correlations between aggregate variables. Next, we have looked at how we should understand MABM validation in the context of mechanisms and target validation. Again the starting point is the layered structure of MABMs. It implies that the mechanisms at one level provide input to the mechanisms at the next level. The mechanisms embedded in the agent behaviour are validated by comparing their behaviour with empirical behaviour of agents directly. These mechanisms are then input for the mechanisms at higher levels. The mechanisms at higher levels cannot be observed at the micro level since they are emergent properties. They arise out of agent interaction through an inductive process. Due to the complexity of MABMs this process usually comes in the form of numerical simulation. This implies that the mechanisms at higher levels can only be validated indirectly. This combination of direct and indirect assessment of mechanisms means that MABMs can best be described as grey-box models built from white-boxes.

Having situated MABMs within the white-, grey- and black-box classification we can look at how MABMs fit into discussions regarding realism versus instrumentalism. This is because white-, grey- and black-box types can be viewed as having different implications regarding realist versus instrumentalist views on modelling. Indirect testing, i.e. validation through joint output, leaves open possibilities for mechanisms in models to only be correct from an instrumentalist sense, while direct validation reduces these possibilities. Compared to DSGE models in macroeconomics whose mechanisms are only validated indirectly, MABMs can be seen as a move towards a more realist view on modelling. The reason for this is that direct validation at the micro level reduces the number of possible model specifications that are correct only from an instrumentalist view.

Concluding, this paper serves as a step towards a methodology of empirical validation of MABMs, that provides a more fundamental understanding of current MABM validation practices. We have seen that a proper assessment of validation practices in MABMs requires an understanding of both model validation and complex systems with emergent properties. Validation of MABMs cannot be compared one-to-one with DSGE, or other macroeconomic model types. MABMs have a unique structure that requires specific forms of validation new to

macroeconomics. Furthermore, since the Lucas Critique there has been a desire in macroeconomics to micro-found macroeconomic models. By a desire to micro-found macroeconomic models while maintaining mathematical elegance, model validation in macroeconomics has shifted towards a more instrumentalist view regarding the mechanisms by which agents create macroeconomic patterns. Through understanding how MABMs are validated we have seen how MABMs can be regarded as a move towards a view of realism compared to DSGE models. The MABM methodology can thus be seen as an attempt to be an answer to the Lucas Critique while maintaining a more realist view towards the model's mechanisms.

Funding The study was funded by Utrecht University - department Utrecht School of Economics.

Availability of Data and Material No data was used in the preparation of this paper.

Code Availability No code was used in the preparation of this paper.

Declarations

Conflicts of interest The author has no relevant financial or non-financial interest to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arthur, W. B. (2013). *Complexity economics*. Oxford: Oxford University Press.
- Baas, N. A., & Emmeche, C. (1997). On emergence and explanation. *Intellectica*, 25(2), 67–83.
- Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review: The Journal of the System Dynamics Society*, 12(3), 183–210.
- Boumans, M. (2009). Understanding in economics: Gray-box models. In H. W. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 210–229). Pittsburgh: University of Pittsburgh Press.
- Brock, W. A. (1999). Scaling in economics: A reader's guide. *Industrial and Corporate change*, 8(3), 409–446.
- Caiani, A., Godin, A., Caverzasi, E., Gallegati, M., Kinsella, S., & Stiglitz, J. E. (2016). Agent based-stock flow consistent macroeconomics: Towards a benchmark model. *Journal of Economic Dynamics and Control*, 69, 375–408.
- Dawid, H., & Gatti, D. D. (2018). Agent-based macroeconomics. In C. Hommes & B. LeBaron (Eds.), *Handbook of computational economics* (Vol. 4, pp. 63–156). Amsterdam: Elsevier.
- Elsenbroich, C. (2012). Explanation in agent-based modelling: Functions, causality or mechanisms? *The Journal of Artificial Societies and Social Simulation*. <https://doi.org/10.18564/jasss.1958>.
- Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, 4(5), 41–60.
- Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton: Princeton University Press.

- Fagiolo G., Roventini A. (2017) Macroeconomic policy in dsge and agent-based models redux: New developments and challenges ahead. *Journal of Artificial Societies and Social Simulation*, 20(1),1, <http://jasss.soc.surrey.ac.uk/20/1/1.html>
- Fagiolo G., Moneta A., Windrum P. (2005) Empirical validation of agent-based models. In ACEPOL 2005-international workshop on agent-based models for economic policy design.
- Fagiolo, G., Moneta, A., & Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, 30(3), 195–226.
- Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., & Roventini, A. (2019). Validation of agent-based models in economics and finance. In C. Beisbart & N. Saam (Eds.), *Computer simulation validation* (pp. 763–787). Cham: Springer.
- Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256), 685.
- Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, 79(3), 733–772.
- Gatti, D. D., Desiderio, S., Gaffeo, E., Cirillo, P., & Gallegati, M. (2011). *Macroeconomics from the bottom-up*. New York: Springer Science & Business Media.
- Gatti, D. D., Fagiolo, G., Gallegati, M., Richiardi, M., & Russo, A. (2018). *Agent-based models in economics: A toolkit*. Cambridge: Cambridge University Press.
- Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.
- Grüne-Yanoff, T. (2009). The explanatory potential of artificial societies. *Synthese*, 169(3), 539–555.
- Guerini, M., & Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control*, 82, 125–141.
- Hoover, K. D. (2015). Reductionism in economics: Intentionality and eschatological justification in the microfoundations of macroeconomics. *Philosophy of Science*, 82(4), 689–711.
- Kydland, F. E., & Prescott, E. C. (1996). The computational experiment: An econometric tool. *Journal of economic perspectives*, 10(1), 69–85.
- Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1), 33–67.
- Lengnick, M. (2013). Agent-based macroeconomics: A baseline model. *Journal of Economic Behavior and Organization*, 86, 102–120.
- Moreno, A., Rodríguez, J. J., Beltrán, D., Sikora, A., Jorba, J., & César, E. (2019). Designing a benchmark for the performance evaluation of agent-based simulation applications on HPC. *The Journal of Supercomputing*, 75(3), 1524–1550.
- Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. *Ideas in Context*, 52, 10–37.
- O'Connor T., Wong H. (2015) Emergent properties (stanford encyclopedia of philosophy). <https://plato.stanford.edu/entries/properties-emergent/>, retrieved 11 Apr 2021.
- Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106, 467–482.
- Simon, H. A. (1976). From substantive to procedural rationality. In J. G. De Gooijer & R. J. Hyndman (Eds.), *25 years of economic theory* (pp. 65–86). Cham: Springer.
- Souleles, N. S. (1999). The response of household consumption to income tax refunds. *American Economic Review*, 89(4), 947–958.
- Stanford K. (2009) Underdetermination of scientific theory (stanford encyclopedia of philosophy). <https://plato.stanford.edu/entries/scientific-underdetermination/>, retrieved 11 Apr 2021.
- Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *The Journal of Artificial Societies and Social Simulation*, 10(2), 8.