4

Assessment and Validation of Multi-agent Models

Frédéric Amblard, Pierre Bommel and Juliette Rouchier

4.1. INTRODUCTION

A frequent criticism of multi-agent models concerns their "validation". During the presentation of a model at a conference, it is quite frequent that *the* question about validation occurs, inducing a visible embarrassment in the speaker. It is then often answered that the presented work is on-going and that the phase of validation will be shortly undertaken, or that the model is too greatly abstracted at the moment and should then be refined before the question of validation is tackled. This last answer, even if it is not always sincere, is an infinite escape, if the person who adopts it hopes to reach one day a perfect model of observed phenomena. The best answer would consist in asking back: "What do you mean by validation?" "According to you, what are the criteria which would allow us to say that a multi-agent model is validated?" It is then frequently answered that a model is validated once its output values are "close enough" to observed data, by applying a classical method of validation for descriptive models such as in econometrics. Of course the comparison of simulation results with sets of empirical data constitutes an important exercise that takes place in the modelling process. But is it enough to conclude that the

model is valid? Disregarding accidental correlations, a large number of logical, theoretical or practical problems arise when we want to compare the "predictions" of a model with empirical data. A model may display consistent results for instance with empirical data even though its content is clearly far from the dynamics that it aims to represent.

Furthermore, if we retain this criterion of validation, the quantity of data necessary to assess the validity of a multi-agent model is a problem in itself [Chapter 5] and especially if we consider the application of these models in the humanities and social sciences where experiments are difficult to carry out, the collected data are called into question and the collection itself extremely costly, even when possible. But these questions about validation apply as much to multi-agent models as for standard mathematical models. For instance, Schaefer's model [SCH 57] which describes the evolution of fishing hauls according to the state and dynamics of fish stocks, shows its limits [GIL 89] and can be considered as non-valid from the criteria identified now. Nevertheless, it is used by numerous managers to estimate the optimum level of exploitation [NAT 99]. And its extension to the Gordon–Schaefer model is used to decide on conservation principles and on economic fishing policies (quotas of production, taxes on production or on investment). Questions concerning the validation of models should not be dissociated from those relating to their uses. Does the conclusion about the validity of a model permit making decisions on the basis of this sole model that can have important consequences? Would even a hypothetical perfect model bring the *Truth* we could follow blindly? Would its precise explanation of the world allow us to predict the future, like Laplace's demon? If the majority of scientists acknowledge that the range of models is restricted and that there is no definitive and guaranteed validation, the general public would not take the same precautions when facing results produced by a "validated model" (scientifically of course!).

In this chapter, we shall thus discuss various criteria to tackle several aspects, going from the conception of the model to its use, to assess it rather than validate it. A more general discussion about the assessment of knowledge through an epistemological perspective is presented in the appendix 1 to this book. It also deals with the question of simulation as "experimentation".

4.2. WHAT DOES MODELLING MEAN, WHAT ARE MODELS?

4.2.1. The Modeller's Project

The first definition we could use for the notion of model is that of Minsky [MIN 65]: "*To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A* ". This very simple definition refers to key concepts of modelling in general, concepts which are little known or, at least, not much taken into account by modelling in the social sciences. From a domain including a set of entities and of empirical phenomena, called the "object domain" or the "target system", Minsky invites us to define a frame and a question regarding this system. Modelling corresponds then to a process of abstraction relative to the question identified. The idea is that to take into account some phenomena from the target system *A* and to answer the question asked by *B*, it is "sufficient" to study an abstraction of *A*: the model *A**. This point of view introduces the notion of a model's *boundaries*: answering this question means selecting some entities and some relations from the target system, and then to eliminate those that should be considered as external to the *A** model. This demarcation also concerns related processes: which ones must be taken into account or put aside, so that the model, at least in its first version is relevant enough to answer the given question? The design of a model for simulation means to focus on dynamic processes and to conceive formal hypotheses concerning its changes of state.

To define an operational notion of model for the social sciences, we may say in a synthetic way that the model we are interested in, is an abstraction of the object domain formalized by means of a non-ambiguous language. The abstraction is carried out according to a purpose, to a question or to a particular aspect of the system studied. In this framework, multi-agent modelling can be defined as a collective phenomenon to be studied or understood, and on which we put forward hypotheses at individual and collective levels. As a matter of fact, from hypotheses and from simplifications that we abstract from the target system, we try to better understand the emergence of collective phenomena. Contrary to classical models, these hypotheses are presented at several levels. They are abstracted in:

1. A model of the entities that we consider as belonging to the system, and on which we associate a behavioural pattern at the individual level;

2. A model of the organization (a social network for example) of these entities and its evolution;
3. A model of the environment and its evolution;
4. A model of the interactions between these individuals but also between these individuals and their environment or between these individuals and the organization;
5. Hypotheses on the initial conditions of the simulation (the state of the system at an instant t_0 from which the simulation should start).

All these hypotheses when translated into models and initialized (a value being appointed to every variable), allow to define precisely a *simulation experiment* This experiment can be carried out, that is to say calculated or run by a computer. Furthermore, this experience has to be observed, i.e. some probes have to be designed (attributes or aggregated variables for which we would like to monitor their evolution during the simulation). The modeller becomes also progressively an experimenter, taking measurements in experimental conditions on a virtual object. To loop on the modelling process, the simulation outputs are confronted to the initial hypotheses or perhaps compared with observed data from the target system. At the end this comparison often makes it possible to find new ways to accomplish new experiments, to refine or simplify the model, or even to restructure it completely. This process may question former knowledge and representations.

It is now necessary to clarify the specificity of multi-agent models compared with classical ones. Indeed, the common representation of scientific models is typically the model of descriptive statistics, that is to say a model constructed first of all to represent a phenomenon as simply as possible without trying necessarily to offer an explanation. Validating these types of models generally implies measuring the deviation between the model outputs and the empirical data collected on a phenomenon by direct observations or by experiments. From our point of view, the multi-agent model belongs to another category that regroups models constructed to explain and understand a phenomenon. From hypotheses put forward on mechanisms at the individual level [MAN 05], the modeller tries to assess the explicative range of these hypotheses and to identify possible individual behaviours that may generate the collective phenomenon. In this framework, the research of similarities[1] with data, even useful, cannot be a sole and final criterion of validation.

4.2.2. The States of Models

With these various purposes, various states of the model are added during the elaboration process, even models that are qualitatively rather different. So, during the elaboration stages of a simulation model, one can identify several stages of design, which show that several "models" are developed in fact during the full process.

A first stage results in a theoretical model, which describes the target system through an agent approach. Often, the UML formalism (Unified Modelling Language) is used to produce a schematic description of the system. UML allows to present the same model from various points of view called *views*. One of the most famous view is called the class diagram which shows the entities of the system (agents and objects), their characteristics and their relations. The temporal sequence of the system is represented by other views such as sequence diagrams, states-transition diagrams or diagrams of activities, which allow for instance to represent the behaviour of the agents. This model, designed on paper, should ideally contain all the information concerning the details of the simulation, to make it possible to implement the model in any computer language without requiring additional assumption. In fact, specification languages such as UML are partly ambiguous (always much less than a discursive text), but however, they give a good description of the implementation.

The second model to be considered is the implemented model, also called the "code" that is expressed in a programming language, often of high level (SmallTalk, Java, C++). This program must be a faithful translation of the UML model. It is used to run simulations. Besides the code of the model itself, it is necessary to instrument it in order to observe it, for example, to observe certain data of the simulation and to let the user see the evolution of this virtual system. Indeed, a simulation creates a world in which the agents interact according to the implemented model. So, at the initialization of a simulation, agents and objects are created, as computing entities, which contain data and capacities of calculation. Then, a sequence of processes intervenes which follows the temporal order defined by the modeller. Without outside intervention, the computer calculates these sequences, which cause changes of state of the agents and of the objects. But if the computer uses data for its calculation and if it generates some, it only provides those that the user asks for, particularly as indicators (often aggregated), which are considered as important. It is thus necessary to have a model of observation of the system (an epiphyte system according to [GIR 94]).

This task is made easier by using simulation platforms. The Cormas platform allows for instance to easily define indicators and to provide graphs which help the user to follow the evolution of its simulation. In SDML, all the information (interactions between the agents and their internal states) is kept in memory. Despite a slowing down of the execution, it allows a very complete understanding of the sequence of simulation.

Finally, the result of this work involves detachment. The user of the model conceives a *model of the model* , that is to say, he organizes the knowledge kept during the stage of experimentation; he gets a better understanding of its functioning, far from a particular execution of an experiment [DEF 03]. From that point, he will be able to transmit his knowledge textually to the scientific community, selecting the essential elements of the model and its functioning. But the knowledge expressed by the model must be clear enough to run the risk of criticism and to be debated. Without even discussing empirical tests, the description of the model must be sufficient to make it possible to reproduce the computing model for readers [EDM 03]. This perfect specification of the model remains, however, an ideal, because numerous experiments of re-implementing models eventually failed. The main reason for these failures is due to the incompleteness of the description, particularly concerning the operational aspects but also the functional elements [HAL 03]. Therefore, a good specification of a model has become nowadays a key criterion for evaluation, not of the model itself, but of the modelling work achieved.

4.2.3. The Various Usages of Multi-agent Modelling

4.2.3.1. Modelling is a Learning Process

A good model is characterized by the fact that once implemented and studied, the modeller does not need it any more. Presented a bit provocatively, this is the thesis of Grimm [GRI 99], who insists on the purpose of learning from modelling or more exactly from the process of building the model. Notwithstanding this view, it seems reasonable to underline, as Grimm did, the essential usage of models. By designing a model of its studied system, the modeller illustrates its cognitive capacities by choosing the essential elements to be taken into account to explain the given phenomenon. By designing his model, he also assumes the qualitative hypotheses about the basic mechanisms that induce the phenomenon observed on the real system. By simulating his model, he can test and analyze these hypotheses and discuss them.

He extracts from the artefact *in silico* the properties of his model and the consequences of his hypotheses. Thus, the set of all the tasks constituting the modelling process facilitates a better representation of the functioning of the target system, which, in the opinion of Grimm, is the main advantage of modelling. A criterion of evaluation of this learning process seems difficult to set up and very subjective: How to assess the fact that the modeller learns from his model? Nevertheless, it is a criterion that is necessary to take into account and that looks somewhat like the Legay criterion of produced knowledge [LEG 73].

4.2.3.2. A Model to Predict, to Understand Or to Act?

In practice, it is possible to distinguish between models conceived to predict or to describe, and those conceived to understand. This difference of purpose in the activity of modelling does not exclude theoretical considerations: nothing prevents a model conceived for understanding to be used after all in purposes of prediction. A difference can however be noticed in the method of designing the model. As part of research leading to description, the reproduction of observed data is of major importance. In the framework of classical modelling, the modelling purpose will be to find a function representing for instance the growth of the purchasing power of a population. But no explanatory effort is provided to understand the reasons of this shape. The use of the model conceived for prediction will serve for extrapolating results in a reasonable time scale. The question of validation for this type of model consists merely in a simple confrontation with data since the first purpose is definitely to reproduce empirical series. Unfortunately, this model can hardly be used to explore alternative situations on the basis of the same social situation, since the factors influencing this social situation are not even clear.

Models conceived to understand a given phenomenon or the evolution of a system are designed through hypotheses among a set of possible hypotheses which concern, in the particular case of multi-agent models, is behind individual behaviour, the interactions between these individuals or between the individuals and their environment. This type of purpose matches with a majority of multi-agents models. The work of formalization and implementation, then of exploration of simulations produces an explanation from basic hypotheses and from their involvements calculated by the simulations. In this framework, if the confrontation of the model with data can bring a lot of information and knowledge for the evaluation of the model, one can see more clearly

still that it cannot constitute the unique criterion of decision-making concerning the validity of the model. In this sense, it is necessary to assess not only the results, which can be compared with qualitative data, but it is also necessary to check the individual behaviours of the agents. These behaviours have to be qualitatively adequate in terms of knowledge about the social system, and the influencing parameters must be able to be interpreted as part of the modelled system.

A third category for the usage of models for which multi-agent models are particularly meaningful, are the models designed for intervention or more precisely to help in cooperation over management choices. For this purpose, the model is seen as an artefact, a visual support, shared by all to help to elaborate, for instance, rules of management of a renewable resource. This type of usage is widely discussed in this book (see *Companion Modelling*, [Chapter 9]).

4.2.4. The Various Modelling Practices

The practice of multi-agent modelling, has seen the emergence of several "paradigmatic" rules of "good practice" for the design and elaboration of models. These rules, often summarized in a leitmotiv, strongly influence the design of models. They re-use some classical criteria for model evaluation. Thus, it seems important to introduce them shortly here.

The first, called the KISS approach (*Keep It Simple, Stupid !*) is a direct application of Ockham's razor, also called the principle of parsimony. It recommends the design of models which are analysable and simple enough to be dissected by a human being who observed the simulations attentively [AXE 97]. The positioning of this approach can be seen as follows: it is useless to conceive models without being able to study their properties seriously. With such complex models, one cannot achieve internal validation defined as the existence of the good properties of the model in its own formal framework. External validation, implying data comparisons, is also difficult for such complex models, which require too many data to assess their adequacies with a real phenomenon. They would be over-determined in comparison with available data. Thus, the number of parameters of such models makes it possible to fit them to any series of collected data.

The systematic application of the parsimony principle leads to deviations as the one that consists in introducing into the multi-agent model some elements of descriptive (or even analogical) modelling at the level of the individual behaviours. An example could be the massive

use of probabilistic laws for the modelling of agents' choices; another example could be the use of some heuristics like genetic algorithms to simulate an optimizing behaviour. These problems led some scientists to promote the KIDS approach (*Keep It Descriptive, Stupid!*). Its idea is to keep as much as possible to an explicative approach, and to try to let the model be fully isomorphic to the phenomena it is supposed to mimic.

The Companion Modelling approach (ComMod) does not set systematic limits for KISS or KIDS, even if the majority of those who apply it tend to use KISS. In this approach, it is not rare that the environment and social relations are modelled with a high degree of precision. However, as the model is designed generally in several stages and with non-professional modellers, some strong simplifications of the different models occur during the process, passing gradually from a KIDS model to a KISS model for which the evolution shows the advance of process.

4.3. INTERNAL AND EXTERNAL VALIDATIONS OF MULTI-AGENT SIMULATIONS

Usually, two stages in validation are distinguished: "internal" and "external" validation. First of all, the internal validation includes a check of the conformity between specifications and the implemented program. This stage aims at answering the following question: is the implemented model definitely the one that I wanted to implement? (Particularly in case the modeller is not the one who implemented the code [MEU 01]). Then, internal validation aims at looking for and identifying model properties. In the case of multi-agent simulations, one cannot achieve logical proofs. So, the question raised is: "Does my model possess the expected properties?" Among the good properties, it is necessary to assess for instance robustness [Chapter 11] or to run sensitivity analysis to verify that the answers are well differentiated on the parameters space. Indeed, this phase of internal validation aims at verifying that the intrinsic logic of the model is achieved.

The second stage of validation,*external* validation, is related to the adequacy between the model and the real phenomenon for which it has been conceived. For this last stage, the comparison with empirical data (or the fact that the model is able to display stylized statements identified from the target system) constitutes key criteria. Thus, the studies carried out via simulations concern, first of all, the systematic

properties of the model (structural and dynamic properties) and the patterns that can emerge because of the stated hypotheses (internal validation). Then the pertinence of the model is assessed with respect to expected situations that we like to represent or to anticipate (external validation). These two stages can loop iteratively between them. Of course, they need a clear description of the model, so that the different methods of valuation can be applied.

4.3.1. Pre-requisites for Validation

Before dealing in detail with model validation, it seems essential that the model *can* be validated. The model should, therefore, have been built using some criteria that minimally allow the reproducibility of conducted experiments. A modelling phase driven in a context where simulation experiments are not reproducible cannot be seriously considered. For scientists who do not participate in the initial experiment, it would be the same as blind trust in the results, that is contrary to the basic principles of a scientific approach. In order to ensure the reproducibility of experiments or minimally the results of the experiments, we then have to be sure that the model is described formally with enough details to enable replication. In particular, it is important to mention explicitly the points on which there could be any ambiguity and then the risk of an implementation with very different properties (as, for instance, in the case of updating mechanisms, synchronous or asynchronous, in discrete-time simulations). As we will discuss, a method if identifying those points could be to get the model re-programmed by the model by another developer.

We can distinguish experiments that are replicable statistically from experiments that are replicable unitarily. Experiments that are classically achieved in experimental sciences are statistically reproducible in the sense that a singular experiment cannot be replicated exactly in all its details, but statistically it can be shown that if we take the same experimental conditions the results are similar. In order to ensure the unitary reproducibility of the experiments, that is possible for simulations (the algorithm being executed is finally deterministic), we have to be cautious: choose a multi-platform language (like Java or SmallTalk) for which the execution does not depend on the computer on which it is executed; control the execution sequence in the simulation of the different processes; and finally control the seed of pseudo-random number generator (PRNG) (it is therefore important to specify the PRNG that is used). This set of criteria makes sure that

anybody can exactly reproduce the same experiment and then test or simply observe the singular elements in the model that are linked to a particular experiment: "why at the 15th iteration, the agent no. 134 decides to buy 120 goods of the firm 28?" Agent-based simulation corresponds sometimes to a kind of *in silico* historical process [AXE 97], being able to reproduce at will exactly the same process and, for instance, to change the kind of measures done can greatly help to understand a phenomenon.

4.3.2. Internal Validation

As explained during the introduction to this section, the first step identified for internal validation is divided into two points: delimitation of the model's properties concerning its dynamics and verification of the programming task. Verification is routine in computer sciences, when a formal system is transformed into a program, we check that the program effectively executes the theoretical model as it has been described—decision algorithms, articulation of communications, internal evolutions of objects and agents. More than the correction of programming mistakes for which verification tools (method B or Z) can be used (but are not much used practically), there clearly exist errors that can be determined only by running the model, for instance during the identification and the detailed analysis of unexpected behaviours in the model. In this case, by studying logical coherence we can know if some errors have been introduced in the code. This point requires a good knowledge of the model. It follows a second step of observation.

The second step, essential in internal validation, as well as in the knowledge and communication of the dynamical structure of the model, consists in identifying the model's properties in the context of its own logic. Some of these properties, as the possibility to attain all states or identification of interlocking phenomena in parallel processes, can be studied by using formal verification methods as Petri nets but are not much used in multi-agent simulation [BAK 03], often because of the important cost concerning the deployment of these techniques on big programs as multi-agent models. The properties linked to the models robustness, generally evaluated by using *sensitivity analyses* [Chapters 3 and 11], are in practice much more frequently used. This is surely due to one of the first aims of sensitivity analyses that consist in identifying the parameters that most influence model dynamics; which allows a choice, and to focus on important parameters of the multi-agent model where we have to evaluate an impor-

tant number of parameters. Sensitivity analyses help to evaluate the *filter quality of the model*, a criterion we find in [LEG 73]. According to his arguments, an interesting property of a model concerns its abil - ity to classify or discriminate the elements it receives in input (the parameters) in different outputs. A model, even built upon reasonable behavioural hypotheses, which would behave as a random number generator whatever is entered in input, would not be much use in the end. This capacity to discriminate a given input of the model should be related to the selected output variable, this latter being frequently an indicator built by the modeller. The observation of an important noise on this output variable can question the model itself but also the chosen output indicator. Sensitivity analysis, if it could be applied to test the robustness of the results of a model can be used also to test the robustness of the model's structure. Modifying the hypotheses taken, for instance, by modifying the organizational structures, the model - ler can obtain indicators related to the stability of his model and his hypotheses. These indicators help to evaluate the importance of the choice of a hypothesis and the influence of its replacement by another on a particular aspect of the model.

Another important property we have to study during this step of internal validation concerns the classes of behaviours generated by the model. Multi-agent simulations produce what is commonly named *emergent behaviours* [Chapters 14, 16, 17], that is to say behav - iours that can not be expressed by using only hypotheses based on individual behaviour. If we take as an example the opinion dynamics model proposed by [DEF 02], the three qualitative classes in presence of extremists are central convergence (the fact that all individuals con- verge on the centre of the opinion distribution), the convergence on both extremes and the convergence on one extreme of the distribution only. The identification of these classes is of course the result of a large number of experimentations achieved on the model. We will see in the following section how the identification of the behavioural classes with stylized facts is a kind of external validation commonly used for multi-agent simulations.

Once the identification of the behavioural classes is achieved, a final point to discuss concerns the identification of the conditions that favour their emergence; i.e., identification of the parameter values for which one or another form tends to emerge. This step, added to a deeper study of generative mechanisms of emerging forms ("what in the program makes that a chosen input value gives another particular

output?") is surely one of the most difficult phases to build concerning multi-agent models. But this is a crucial step in order to control the model, as it serves to explain the reason why each phenomenon has been observed in the simulation. Several attempts to give a more operational approach for this phase are currently under elaboration either as tools and methods to assist this search [AMB 03a, AMB 03b] or as tools enabling an automatic detection of production rules of identified emergent behaviours [YAH 05].

Among all these data on system robustness, an important group should be furnished along with the system description in order to enable a counter-verification to be made by another programmer. This is on the basis of the model description, and on a consequent set of knowledge of its dynamical properties and the pertinent emergences that are produced, through which a replication can be effected [AXE 97]. This supplementary validation exercise is becoming more widely used in order to establish the model's validity (and could even be a simple exercise in the context of a university teaching [BIG 05]). To replicate a model, we rewrite the program starting from the original theoretical model, generally using another software or programming language, in order to check if structural results described above are not biased (due to the implementation), but reliable results that are effectively the direct consequences of theoretical hypotheses. This exercise has been practised frequently during M2M workshops ("*Model to model*", as we will see later on) and serves to demonstrate that in fact numerous articles do not provide enough information for the model to be replicated (and then to check their inner dynamics) [ROU 03] and that sometimes even the published results are impossible to retrieve [EDM 03].

4.3.3. External Validation

External validation, the second phase for validation, raises the question of the adequacy between the model and the target system, taking into account the aims or the planned use of the model. The aim of this phase consists in aligning the model and its results with observations or experiences achieved on the target system. The rapid conclusion that would assimilate this step to a "simple" quantitative comparison with empirical data, gives an idea of the methods usually adopted in that framework. As underlined in our introduction, this is frequently what is meant when the validation of models is mentioned (see discussions about inductive inferences and falsificationism in Appendix 1).

A method that is employed in agent-based simulation consists in comparing *classes of behaviours* (identified during the phase of internal validation) to emergent behaviours from the target system: the *stylized* facts. This comparison, even if it could be done with stylized facts identified *a posteriori* (serving, for instance, to discover in empirical phenomena, some stylized behaviours that wouldn't have been identified), is strengthened when stylized facts are determined before the modelling process as desired behaviours we attempt to rebuild with the model. We could then use them as a validation criterion among others. The capacity to generate stylized facts, often at the macroscopic or global scale, emerging from individual hypotheses, is one of the characteristics, and one of the main motivations, for using an agent-based approach. A question that could be raised here, that we will discuss later on, is whether or not it would be the only way to make external validation on such models, taking into account the difficulties linked to quantitative comparison with empirical data. Then, this qualitative mode of comparison between the outputs, the productions of the model and the target system, even if not totally satisfying for the modeller, helps us at least to realize that the model produces some effects that are qualitatively close to the ones we are trying to obtain. This criterion of empirical adequacy, even if far from being definitive, states that the model *structure* possesses some properties, given these hypotheses that urge the modeller to keep these latter, and suggest other hypotheses which would not reproduce qualitative behaviour so efficiently should be ruled out. [2] Nevertheless, the weakness of this method can be criticized: it is by human observation that we bring closer two stylized facts, the first one produced by the model, the other one extracted from empirical phenomena. This raises the arguments about the *fallibility of the observer,* which when added to the qualitative comparison with data, is often criticized. Despite this, observation is more and more frequently assisted by the use of indicators on both the target system and the model that would be comparable. It is clear that this second step in validation requires, in order to be fully achieved, a good expertise in the domain on the part of the person who controls the model. The relevance of the indicators chosen to significantly characterize the situation has to demonstrate the functional similarity between the model and the target system. For several researchers, the creation of these indicators and what serves to indicate if they are satisfying and sufficient, is, as such, a complex task but a fundamental one that should be justified in the modelling

process [ROU 00, DEF 03]. There are fewer interpretations of models that are translated into stories (depending a lot on the talent of the storyteller) that might be plausible and recall the modelled target system. And there are more comparisons of forms between the outputs of the virtual system and the data of empirical observation.

However, the quantitative comparison with data should not be given up. On the one hand because the trend mentioned above tends to move toward quantitative approaches and on the other hand because the modeller who might have gathered data on his target system would like to compare them with the output data of his model. Despite this, we have to underline that this task is difficult for two main reasons. The first reason deals with the lack of data or the frequent difficulty if not impossibility of repeating field experiments a great number of times. The second reason, linked to the preceding one, concerns the over-determination of data by the model parameters. [3] This comparison, we can identify several criteria relating to, especially the organization level where the comparison takes place, whether temporal or not. Dealing with the organizational level, we may want to compare gathered data to model indicators defined at an individual level. It may involve, for instance, the comparison of individual trajectories in the model with individual data on the target system under consideration. Or we may wish to compare gathered data with indicators defined at a macroscopic level. We would compare then a global indicator as the population size with a global indicator of the considered system. It could be done also at intermediary levels [JEA 03], on the scale of a cohort or a small group of individuals for instance. The fact that gathered data could be either temporal series or data gathered at a given time step helps to determine the kind of comparison that could be achieved. It is, for instance, useless to compare individual trajectories with the model if we do not have longitudinal data gathered on particular individuals. However, one might stand on the use of partial comparison with data, they are both difficult and expensive to obtain in the social sciences, and the link between these data and the model variables is not always as automatic as it seems. In order to evaluate the quantitative gap that exists between the model and the data, let us consider a model where 1,000 agents have an initial opinion, coded as a continuous variable and an associated uncertainty. These individuals are situated in a social network and change their opinion according to their state and the state of their neighbours in the network. Ideally, we would like to have data concerning the time evolution of

the opinion of these 1,000 individuals, as well as data to determine the initial state of the social network. It would be necessary also to link gathered opinions to a real value between –1 and 1. This example shows that we have to give up the aim of getting sufficient data that would correspond to the data that are possible to produce with a series of simulation experiments. And we have to stand on partial comparisons, either on a macroscopic scale or on individual trajectories that still bring some elements for the validation, even if far from being definitive.

The other problem raised by the comparison with data (mainly macroscopic) concerns over-determination by the model parameters. This problem holds classically for descriptive models that have more parameters (or freedom degrees) than available data. This question is rarely asked for agent-based simulation but merits further consideration. Let us consider that we have twenty empirical points to qualify the evolution of the target system on a global scale at twenty different moments in time. If our multi-agent model is composed one hundred agents assuming that the initial state of each one of these agents as well as the initial organization influence the simulation results, we would then face a huge number of degrees of freedom over which we can play and we could even say that whatever the amount of available data, the model can under certain conditions produce results that are very close to the empirical data. Even if we are rarely in this caricature case of having a model able to fit whatever data series—we should then reconsider its capacity to serve as a *discriminating filter* on the inputs—agent-based models often have sufficient flexibility to fit a curve locally that has approximately the same shape. This problem should obviously be treated case by case, but for a model that is flexible enough, it would be reasonable to conduct only a qualitative comparison as this is the only possible comparison.

From a formal point of view, external validation consists in creating indices to affirm that a model is adequate to observed phenomena. It helps to draw several useful conclusions: explanations about functioning by making the parameters precise that mainly drive model behaviour; predictions concerning the influence of several institutions to solve a problem or organize a situation; demonstration concerning the functioning of the actors' rationality; previsions about the evolution of the situation regarding the current context. Depending on the target and the original data, the elements used to evaluate the pertinence of the model can be purely quantitative—we make the values of one of

the simulation indicators correspond to obtained values from the real situation; these data could be qualitative, and we make an evaluation about the pertinence of the system behaviours identifying the qualitative classes of phenomena. The observed elements can be only global, or could concern also the logic of the behaviours of the individual agents and the communication between them.

4.3.4. Comparison of Models

The comparison of models is also one of the methods used to evaluate a model. Grimm [GRI 99] recommends this activity in a more restrictive way than the works developed around the workshop series Model-to-Model we will discuss. For Grimm, *individual-based simulations*[4] should be compared with the existing global descriptive models that carry elements of explanation about the phenomena. Grimm's position should be resituated in his domain, ecology, for which there exists a tradition of global descriptive models, that have been elaborated and compared successfully with numerous empirical series. These descriptive models, dealing with particular phenomena that are also treated in individual-based simulations studied by Grimm, can be used therefore as substitutes for reality. These classical models which tend to explain global phenomena are generally built using formalisms like differential equations or compartment models (such as Leslie matrices). They constitute for Grimm a robust reference framework to build individual-based simulations. Grimm's position, though very attractive, could be adapted only with difficulty to the social sciences where the modelling tradition, if existing in some domains (mathematical economy for instance), has not produced descriptive models that are convincing enough to be unanimously adopted, contrary to the situation in ecology. The works achieved so far in the*Model to Model* (M2M) workshop series [HAL 03], even more recently, propose a framework that is more adapted to the situation in the social sciences by listing a certain number of approaches to evaluate models using comparison, this is the case in particular for models from the literature. Among the proposed methods we can mention for instance:

- The replication, as mentioned above, in another programming language of published models helps to understand all their subtleties and to reproduce published results [AXE 97, BIG 05]. This point helps to check the reproducibility of both models and results and often show that the published information is incomplete [ROU 03]

(it is even true for the classical model by Epstein and Axtell [EPS 96, BIG 05]) or involves false results [EDM 03].

- The coupling of models where different scales (of time and space) are interconnected—the results of one model being used by another. The inter-connection enables, more than the extension of a model, to treat more realistic data where theoretical distributions have been used for the internal validation.
- The comparison of different models dealing with the same kind of results, trying to identify if they produce effectively similar, or even identical, results. This method is sometimes called *model alignment* [AXT 96] and serves among a set of models either to compare the effects of different hypotheses, or, if the results are similar, to select the simplest or the easiest model to be interpreted, according to the parsimony principle.
- The comparison of different models in relation to their adequacy to a data series. This technique, known as *docking*, helps evaluate the qualitative reproduction of empirical data but does not constitute a definitive criterion for the validation of models, as we underlined above.
- The use of a simpler model as an abstract built *a posteriori* or an abstraction of the results of another model. This exercise helps, at the same time, to build the model of the model while facilitating the understanding of the first one or facilitating its use, the second model being typically a descriptive model of the results produced by the first one.
- The use of models by changing structures and hypotheses in order to test the structural robustness of the model.

4.4. CONCLUSION: HOW TO VALIDATE A COMPLEX SYSTEM OF SIMULATION?

Now at the end of the chapter, the reader may be disappointed to not have at his disposal a definitive criterion of validity that would enable him to answer clearly the question "is my model valid?", especially when preparing to face the question for a conference. He would be even more frustrated if he had in mind a representation of validation as it is practised for statistical and descriptive models where validity seems to be evaluated only according to the comparison with the data that the model should describe. It would forget too quickly that validation remains a subjective human judgment: a decision justified by

criteria provided by the researcher. From this point of view, it seems that the main work in modelling is not to provide ready-made answers, but rather to provide as clearly and as detailed a fashion as possible, the criteria that would serve, if not to definitely validate the model, at least to evaluate it. The impossible formal validation of an agent-based model imposes a reflection concerning the final use of the model in order to be valid, the interpretation of the results being done in relation to a specific context.

Among these criteria, listed throughout this chapter, all being useful, but not absolutely necessary in our opinion, we could also add some proposed by Jean-Marie Legay [LEG 73], about other kinds of models applied to other kinds of systems. Jean-Marie Legay articulated his thesis about validation around three crucial points: (1) Models are always imperfect, (2) The value we can give to models is always linked to their aim, a model being first of all a tool. (3) Validation is a decision, a judgment, taking into account some validity criteria, of which none, if taken in isolation, has a decisive value, and all of them serve to describe a profile of the studied model and to take a decision concerning its validity. From these elements, Legay gives a precise description of different criteria, some of them being useful to decisions concerning the model validation:

- *Usefulness of the model*: the set of results and successes produced by the model. In agent-based simulation, we could identify the ComMod approach as giving an answer for placing the reflection in a context that makes the model useful. It is the notion of the model considered as an object for mediation.
- *Simplicity*: it corresponds to the parsimony criterion we mentioned. We should notice that Legay specified that simplicity is not a quality as such and that modelling has to take into account the complexity of the real system. The question of the equilibrium between simplicity and complexity is, as we saw, at the centre of the dispute between pro KISS and pro KIDS.
- *Non-contradiction* The model should respect observed relations. In this aim, numerous statistical techniques could be used, as for instance the χ^2 test to evaluate the correspondence between the model productions, observables of the simulation experiments and the empirical data gathered in the target system. Non-contradiction signifies simply that the model is not to be rejected, but it is not sufficient to defend its adequacy. In the framework of agent-based

simulations, we saw that questions were asked at the same time concerning macroscopic results and individual trajectories [JAN 03]: we can easily obtain good results for bad reasons.

- *Fecundity*: A criterion that takes into account the unanticipated consequences produced by the model, when usefulness concerned only the anticipated consequences.
- *Convergence*: The validity of a model increases with its use, i.e. the number of independent experiments which confirm it. For instance, when some equilibrium of the model of Esptein and Axtell [EPS 96] is reproduced, it can be said that the model demonstrates more solidity [BIG 05].
- *Stability*: the fact that the model is not sensitive to secondary factors, which do not directly concern the important hypotheses of the model, but that it is sensitive to primary factors. This selective sensitivity makes the model a good instrument of measure and exploration. We saw that study of model dynamics deals mainly with research into its primary factors, and the determination of the parameter space inside of which they can vary while the system structure is maintained.
- *Non-identity*: A model is efficient because it differs from its target, and particularly because it is simpler and easier to use to obtain knowledge.

Even if they are pertinent, these criteria require, in order to become operational, to be associated with confirmed evaluation methods. Considering that simulation in its widest sense is relatively recent, it is not so surprising that the validation of agent-based simulation, which is one of the youngest subfields, does not yet propose efficient tools that are shared by the community. Nevertheless, a structuring around research teams with multiple competencies, comprising at the same time computer scientists, statisticians and researchers from the application field, begins to produce or adapt tools to study these complex systems that are agent-based simulations. Nevertheless, one of the characteristics of the agent-based community from its origin is its consciousness that as far as we can go in the validation of a model, it will however only be known in a limited way, just like the target system it represents. In this context, the questions concerning the validation of models are not dissociated from the ones related to their use and the positive evaluation of the model does not transform it in to a tool that is separate from the social context of its production.

4.5. NOTES

1. The reader will find more detailed explanations about various studies on similarities in Chapter 8.
2. Cf. the structural adequacy criterion of Poincaré discussed in appendix 1 for instance.
3. It does not concern the thesis about the "over-determination of facts by theories" (facts being "charged with theories") or the thesis concerning the under-determination of theories by experience (several theories being compatible with the same facts)—see the appendix 1.
4. A denomination of agent-based simulation used currently in ecology and sometimes in the social sciences.

4.6. REFERENCES

[AMB 03a] AMBLARD F., HILL D.R.C., BERNARD S., TRUFFOT J., DEFFUANT G., "MDA compliant design of SimExplorer, a software to handle simulation experimental frameworks", *Proceedings of the 2003 SCS Summer Simulation Conference*, Montreal, pp. 279–284, July 2003.

[AMB 03b] AMBLARD F., *Comprendre le fonctionnement de simulations sociales individus-centrées: Application à des modèles de dynamiques d'opinions*, doctoral thesis in computing at Université Blaise Pascal, Clermont-Ferrand, 2003.

[AXE 97] AXELROD R., *The Complexity of Cooperation: Agent-based Models of Conflict and Cooperation*, Princeton University Press, 1997.

[AXT 96] AXTELL R., AXELROD R., EPSTEIN J., COHEN M.D., "Aligning simulation models: a case study and results", *Computational and Mathematical Organization Theory*, vol. 1, pp. 123–141, 1996.

[BAK 03] BAKAM I., *Des systèmes multi-agents aux réseaux de pétri pour la gestion des ressources naturelles: Le cas de la faune dans l'est cameroun*, doctoral thesis, Université de Yaoundé 1, Yaoundé, Cameroon, 2003.

[BIG 05] BIGBEE A., CIOFFI-REVILLA C., LUKE S., "Replication of sugarscape using Mason", *Fourth International Workshop on Agent-based Approaches in Economic and Social Complex Systems (AESCS'05)*, Tokyo, May 2005.

[DEF 02] DEFFUANT G., AMBLARD F., WEISBUCH G., FAURE T., "How can extremism prevail? A study based on the relative agreement model", *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 4, 2002.

[DEF 03] DEFFUANT G., AMBLARD F., DUBOZ R., RAMAT E., "Une démarche expérimentale pour la simulation individus-centrée", J.-P. Müller dir, *Rochebrune 2003: épistémologie de la simulation*, Paris ENST, pp. 45–64, 2003.

[EDM 03] EDMONDS B., HALES D., "Replication, Replication and Replication: Some Hard Lessons from Model Alignment", *Journal of Artificial Societies and Social Simulation*, vol. 6, no. 4, 2003.

[EPS 96] EPSTEIN J., AXTELL R., *Growing Artificial Societies, Social Science From the Bottom Up*, Cambridge, Mass., MIT Press, 1996.

[GIL 89] GILLY B., "Les modèles bio-économiques en halieutique: démarches et limites", in Verdeaux, Fr. (ed.), La pêche: enjeux de développement et objet de recherche, *Cahiers Sciences Humaines* , vol. 25, no. 1–2, pp. 23–33, 1989.

[GRI 99] GRIMM V., "Ten years of individual-based modelling in ecology: what we have learned and what could we learn in the future?", *Ecological Modelling* , vol. 115, pp. 129–148, 1999.

[HAL 03] HALES D., EDMONDS B., ROUCHIER J., "Model to model analysis", *Journal of Artificial Societies and Social Simulation* , vol. 6, no. 4, 2003.

[JAN 03] JANSSEN M., AHN T.K., "Adaptation vs. anticipation in public-good games", Rouchier J., Edmonds B., Hales, D. (eds), *Model to Model Workshop Electronic Proceedings* , 2003.

[JEA 03] JEANSON R., BLANCO S., FOURNIER R., DENEUBOURG J.L., FOURCASSIE V., THERAULAZ G., "A model of animal movements in a bounded space", *Journal of Theoretical Biology* , vol. 225, pp. 443–451, 2003.

[LEG 73] LEGAY J.M., *La méthode des modèles, état actuel de la méthode expérimentale* , Informatique et Biosphère, 1973.

[MAN 05] MANZO G., "Variables, mécanismes et simulations. Une combinaison des trois méthodes est-elle possible? Une analyse critique de la littérature", *Revue Française de Sociologie,* vol. 46, no. 1, 2005.

[MEU 01] MEURISSE T., VANBERGUE D., "Problématique de conception de simulations multi-agents", *Actes des 9eme Journées Francophones pour l'Intelligence Artificielle Distribuées et les Systèmes Multi-Agents* , 2001.

[MIN 65] MINSKY M., "Matter, mind and models", *Proceedings of IFIP Congress* , pp. 45–49, 1965.

[NAT 99] NATIONAL RESEARCH COUNCIL , Sharing the fish: toward a national policy on individual fishing quotas. National Research Council, (ed)., National Academy Press USA, July 1999.

[ROU 00] ROUCHIER J., *La Confiance à travers l'échange. Accès aux pâturages au Nord-Cameroun et échanges non-marchands: des simulations dans des Systèmes Multi-Agents*, Université d'Orléans, doctoral thesis, Université d'Orléans, 2000.

[ROU 03] ROUCHIER J., "Re-implementation of a multi-agent model aimed at sustaining experimental economic research: the case of simulations with emerging speculation", *Journal of Artificial Societies and Social Simulation* , vol. 6, no. 4, 2003.

[SCH 57] SCHAEFER M.B., "A study of the dynamics of the fishery for yellowfin tuna in the eastern tropical Pacific Ocean", *Inter-American Tropical Tuna Communication Bulletin* , vol. 2, no. 6, pp. 245–285, 1957.

[YAH 05] YAHJA A., CARLEY K.M., "WIZER: An Automated Intelligent Tool for Model Improvement of Multi-Agent Social-Network Systems", *Proceedings FLAIRS,* Miami, 2005.

[ZEI 00] ZEIGLER , B. P., PRAEHOFER , H., KIM , T. G., *Theory of Modelling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems* , New York, Academic Press, 2000.