Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

The quality of social simulation: an example from research policy

Petra Ahrweiler
EA European Academy of Technology and Innovation Assessment GmbH, Wilhelmstrasse 56, D-53474 Bad Neuenahr-Ahrweiler, Germany
Email: petra.ahrweiler@ea-aw.de

Nigel Gilbert
University of Surrey, Guildford GU2 7XH, UK
Email: n.gilbert@surrey.ac.uk

Abstract— **This contribution deals with the assessment of the quality of a simulation. After discussing this issue on a general level, we apply and test the assessment mechanisms using an example from policy modelling.**

"The construction of a scientific social simulation implies the following process: "We wish to acquire something from a target entity *T*. We cannot get what we want from *T* directly. So we proceed indirectly. Instead of *T* we construct another entity *M*, the "model", which is sufficiently similar to *T* that we are confident that M will deliver (or reveal) the acquired something which we want to get from *T*. [...] At a moment in time the model has structure. With the passage of time the structure changes and that is behaviour. [...] Clearly we wish to know the behaviour of the model. How? We may set the model running (possibly in special sets of circumstances of our choice) and watch what it does. It is this that we refer to as "simulation" of the target" (quoted with slight modifications from [1]).

We also habitually refer to "simulations" in everyday life, mostly in the sense that a simulation is "an illusory appearance that manages a reality effect", cf. [2], or as Baudrillard put it, "to simulate is to feign to have what one hasn´t" while "substituting signs for the real" [3]. In a previous publication [4], we used the example of the Caffè Nero in Guildford, 50 km southwest of London, as a simulation of a Venetian café – which will serve as the 'real' to illustrate this view. The purpose of the café is to "serve the best coffee north of Milan". It tries to give the impression that you are in a real Italian café – although, most of the time, the weather outside can make the illusion difficult to maintain. The construction of everyday simulations like Caffè Nero has some resemblance to the construction of scientific social simulations (see Table 1). In both cases, we build models from a target by reducing the characteristics of the latter sufficiently for the purpose at hand; in each case, we want something from the model we cannot achieve easily from the target. In the case of Caffè Nero, we cannot simply go to Venice, drink our coffee, be happy and return. It is too expensive and time-consuming. We have to use the simulation. In the case of a science simulation, we cannot get data from the real system to learn about its behaviour. We have to use the simulation.

TABLE I.
COMPARING SIMULATIONS

	Caffè Nero Simulation	**Science Simulation**
Target	Venetian Café	"Real System"
Goal	Getting "the feeling" (customers) and profit (owners) from it	Getting understanding and/or predictions from it
Model	By reducing the many features of a Venetian Café to a few parameters	By reducing the many features of the target to a few parameters
Question	Is it a good simulation, i.e. do we get from it what we want?	Is it a good simulation i.e. do we get from it what we want?

I. METHODS TO EVALUATE SIMULATIONS

The question, whether one or the other is a good simulation, can therefore be re-formulated as: do we get from the simulation what we constructed it for? Heeding these similarities, we shall now try to apply evaluation methods typically used for everyday simulations to scientific simulation and vice versa. Before doing so, we shall briefly discuss the "ordinary" method of evaluating simulations called the "standard view" and its adversary, a constructivist approach asserting, "anything goes".

A. The standard view

The standard view refers to the well-known questions and methods of *verification*, namely whether the code does what it is supposed to do and whether there are any bugs, and *validation*, namely whether the outputs (for given inputs/parameters) resemble observations of the target, although (because the processes being modelled are stochastic and because of unmeasured factors) identical

This work was partly supported by the FP7 CSA EGovPoliNet.

outputs are not to be expected, as discussed in detail in [5]. This standard view relies on a realist perspective because it refers to the observability of reality in order to compare the 'real' with artificial data produced by the simulation.

Applying the standard view to the Caffè Nero example, we can find quantitative and sometimes qualitative measures for evaluating the simulation. Using quantitative measures of similarity between it and a "real" Venetian café, we can ask, for example,

- whether the coffee tastes the same (by measuring, for example, a quality score at blind tasting),
- whether the Caffè is a cool place (e.g. measuring the relative temperatures inside and outside),
- whether the noise level is the same (using a dB meter for measuring purposes), whether the lighting level is the same (using a light meter) and whether there are the same number of tables and chairs per square metre for the customers (counting them) and so on. In applying qualitative measures of similarity we can again ask
- whether the coffee tastes the same (while documenting what comes to mind when customers drink the coffee),
- whether the Caffè is a 'cool' place (this time meaning whether it is a fashionable place to hang out),
- whether it is a vivid, buzzing place, full of life (observing the liveliness of groups of customers),
- whether there is the same pattern of social relationships (difficult to operationalise: perhaps by observing whether the waiters spend their time talking to the customers or to the other staff), and
- whether there is a ritual for serving coffee and whether it is felt to be the same as in a Venetian café.

The assumption lying behind these measures is that there is a 'real' café and a 'simulation' café and that in both of these, we can make observations. Similarly, we generally assume that the theories and models that lie at the base of science simulations are well grounded and can be validated by observation of empirical facts. However, the philosophy of science forces us to be more modest.

1) The problem of under-determination

Some philosophers of science argue that theories are under-determined by observational data or experience, that is, the same empirical data may be in accord with many alternative theories. An adherent of the standard view would respond that one important role of simulations (and of any form of model building) is to derive from theories as many testable implications as possible, so that eventually validity can be assessed in a cumulative process[1]. Simulation is indeed a powerful tool for testing theories in that way if we are followers of the standard view.

[1] We owe the suggestion that simulation could be a tool to make theories more determined by data to one of the referees of [4]).

However, the problem that theories are under-determined by empirical data cannot be solved by cumulative data gathering: it is more general and therefore more serious. The under-determination problem is not about a missing quantity of data but about the relation between data and theory. As [6] presents it: if it is possible to construct two or more incompatible theories by relying on the same set of experimental data, the choice between these theories cannot depend on "empirical facts". Quine showed that there is no procedure to establish a relation of uniqueness between theory and data in a logically exclusive way. This leaves us with an annoying freedom: "sometimes, the same datum is interpreted by such different assumptions and theoretical orientations using different terminologies that one wonders whether the theorists are really thinking of the same datum" ([7], own translation).

The proposal mentioned above to solve the under-determination problem by simulation does not touch the underlying reference problem at all. It just extends the theory, adding to it its "implications", hoping them to be more easily testable than the theory´s core theorems. The general reference between theoretical statement – be it implication or core theorem - and observed data has not changed by applying this extension: the point here is that we cannot establish a relation of uniqueness between the observed data and the theoretical statement. This applies to any segment of theorising at the centre or at the periphery of the theory on any level – a matter that cannot be improved by a cumulative strategy.

2) The theory-ladenness of observations

Observations are supposed to validate theories, but in fact theories guide our observations, decide on our set of observables and prepare our interpretation of the data. Take, for example, the different concepts of the two authors concerning Venetian cafés: For one, a Venetian café is a quiet place to read newspapers and relax with a good cup of coffee, for the other a Venetian café is a lively place to meet and talk to people with a good cup of coffee. The first attribute of these different conceptions of a Venetian café is supported by one and the same observable, namely the noise level, although one author expects a low level, the other a high one. The second attribute is completely different: the first conception is supported by a high number of newspaper readers, the second by a high number of people talking. Accordingly, a "good" simulation would mean a different thing for each of the authors. A good simulation for one would be a poor simulation for the other and vice versa. Here, you can easily see the influence of theory on the observables. This example could just lead to an extensive discussion about the "nature" of a Venetian café between the two authors, but the theory-ladenness of observations again leads to more serious difficulties. Our access to data is compromised by involving theory, with the consequence that observations are not the "bed rock elements" [8], our theories can safely rely on. At the very base of theory is again theory. The attempt to validate our theories by "pure"

theory-neutral observational concepts is mistaken from the beginning.

Balzer et al. summarise the long debate about the standard view on this issue as follows: "First, all criteria of observability proposed up to now are vulnerable to serious objections. Second, these criteria would not contribute to our task because in all advanced theories there will be no observational concepts at all – at least if we take `observational´ in the more philosophical sense of not involving any theory. Third, it can be shown that none of the concepts of an advanced theory can be defined in terms of observational concepts" [8]. Not only can you not verify a theory by empirical observation, but you cannot even be certain about falsifying a theory. A theory is not validated by "observations" but by other theories (observational theories). Because of this reference to other theories, in fact a nested structure, the theory-ladenness of each observation has negative consequences for the completeness and self-sufficiency of scientific theories, cf. [9]. These problems apply equally to simulations, which are just theories in process.

We can give examples of these difficulties in the area of social simulation. To compare Axelrod´s *The evolution of cooperation* [10] and all the subsequent work on iterated prisoners´ dilemmas with the 'real world', we would need to observe 'real' IPDs, but this cannot be done in a theory-neutral way. The same problems arise with the growing body of work on opinion dynamics (e.g. [11], [12], [13]). The latter starts with some simple assumptions about how agents' opinions affect the opinions of other agents and shows under which circumstances the result is a consensus, polarisation or fragmentation. However, how could these results be validated against observations without involving again a considerable amount of theory?

Important features of the target might not be observable at all. We cannot, for example, observe learning. We can just use some indicators to measure the consequences of learning and assume that learning has taken place. In science simulations, the lack of observability of significant features is one of the prime motivations for carrying out a simulation in the first place.

There are also more technical problems. Validity tests should be "exercised over a full range of inputs and the outputs are observed for correctness" [14]. However, the possibility of such testing is rejected: "real life systems have too many inputs, resulting in a combinatorial explosion of test cases". Therefore, simulations have "too many inputs/outputs to be able to test strictly" (ibid.).

While this point does not refute the standard view in principle but only emphasises difficulties in execution, the former arguments reveal problems arising from the logic of validity assessment. We can try to marginalise, neglect or even deny these problems, but this will disclose our position as mere "believers" of the standard view.

B. The constructivist view

Validating a simulation against empirical data is not about comparing "the real world" and the simulation output; it is comparing *what you observe as the real world* with what you observe as the output. Both are constructions of an observer and his/her views concerning relevant agents and their attributes. Constructing reality and constructing simulation are just two ways of an observer seeing the world. The issue of object formation is not normally considered by computer scientists relying on the standard view: data is "organized by a human programmer who appropriately fits them into the chosen representational structure. Usually, researchers use their prior knowledge of the nature of the problem to hand-code a representation of the data into a near-optimal form. Only after all this hand-coding is completed is the representation allowed to be manipulated by the machine. The problem of representation-formation […] is ignored" [15].

However, what happens if we question the possibility of validating a simulation by comparing it with empirical data from the "real world"? We need to refer to the modellers/observers in order to get at their different constructions. The constructivists reject the possibility of evaluation because there is no common "reality" we might refer to. This observer-oriented opponent of the realist view is a nightmare to most scientists: "Where anything goes, freedom of thought begins. And this freedom of thought consists of all people blabbering around and everybody is right as long as he does not refer to truth. Because truth is divisible like the coat of Saint Martin; everybody gets a piece of it and everybody has a nice feeling" [16].

Clearly, we can put some central thoughts from this view much more carefully: "In dealing with experience, in trying to explain and control it, we accept as legitimate and appropriate to experiment with different conceptual settings, to combine the flow of experience to different 'objects'" [17].

However, this still leads to highly questionable consequences: there seems to be no way to distinguish between different constructions/simulations in terms of "truth", "objectivity", "validity" etc. Science is going coffeehouse: everything is just construction, rhetoric and arbitrary talk. Can we so easily dismiss the possibility of evaluation?

C. The user community view

We take refuge at the place we started from: what happens if we go back to the Venetian café simulation and ask for an evaluation of its performance? It is probably the case that most customers in the Guildford Caffè Nero have never been in an Italian café. Nevertheless, they manage to "evaluate" its performance – against their concept of an Italian café that is not inspired by any "real" data. However, there is something "real" in this evaluation, namely the customers, their constructions and a "something" out there, which everybody refers to, relying on some sort of shared meaning

and having a "real" discussion about it. The philosopher Searle shows in his work on the *Construction of Social Reality* [18] how conventions are "real": they are not deficient for the support of a relativistic approach because they are constructed.

Consensus about the "reality observed by us" is generated by an interaction process that must itself be considered real. At the base of the constructivist view is a strong reference to reality, that is, conventions and expectations that are socially created and enforced. When evaluating the Caffè Nero simulation, we can refer to the expert community (customers, owners) who use the simulation to get from it what they would expect to get from the target. A good simulation for them would satisfy the customers who want to have the "Venetian feeling" and would satisfy the owners who want to get the "Venetian profit".

For science equally, the foundation of every validity discussion is the ordinary everyday interaction that creates an area of shared meanings and expectations. This area takes the place left open by the under-determination of theories and the theoreticity problem of the standard view.[2] Our view comes close to that of empirical epistemology which points out that the criteria for quality assessment "do not come from some *a priori* standard but rest on the description of the way research is actually conducted" [19].

If the target for a social science simulation is itself a construction, then the simulation is a *second order* construction. In order to evaluate the simulation we can rely on the ordinary (but sophisticated) institutions of (social) science and its practice. The actual evaluation of science comes from answers to questions such as: Do others accept the results as being coherent with existing knowledge? Do other scientists use it to support their work? Do other scientists use it to inspire their own investigations?

An example of such validity discourse in the area of social simulation is the history of the tipping model first proposed by Schelling and now rather well known in the social simulation community. The Schelling model purports to demonstrate the reasons for the persistence of urban residential segregation in the United States and elsewhere. It consists of a grid of square cells, on which are placed agents, each either black or white. The agents have a 'tolerance' for the number of agents of the other colour in the surrounding eight cells that they are content to have around them. If there are 'too many' agents of the other colour, the unhappy agents move to other cells until they find a context in which there are a tolerable number of other-coloured agents. Starting with a random distribution, even with high levels of tolerance the agents will still congregate into clusters of agents of the same colour. The point Schelling and others have taken from this model is that residential segregation will form and persist even when agents are rather tolerant.

The obvious place to undertake a realist validation of this model is a United States city. One could collect data about residential mobility and, perhaps, on 'tolerance'. However, the exercise is harder than it looks. Even US city blocks are not all regular and square, so the real city does not look anything like the usual model grid. Residents move into the city from outside, migrate to other cities, are born and die, so the tidy picture of mobility in the model is far from the messy reality. Asking residents how many people of the other colour they would be tolerant of is also an exercise fraught with difficulty: the question is hypothetical and abstract, and answers are likely to be biased by social desirability considerations. Notwithstanding these practical methodological difficulties, some attempts have been made to verify the model. The results have not provided much support. For instance, Benenson [21] analysed residential distribution for nine Israeli cities using census data and demonstrated that whatever the variable tested - family income, number of children, education level – there was a great deal of ethnic and economic heterogeneity within neighbourhoods, contrary to the model's predictions.

This apparent lack of empirical support has not, however, dimmed the fame of the model. The difficulty of obtaining reliable data provides a ready answer to doubts about whether the model is 'really' a good representation of urban segregation dynamics. Another response has been to elaborate the model at the theoretical level. For instance, Bruch [22] demonstrates that clustering only emerges in Schelling's model for discontinuous functional forms for residents' opinions, while data from surveys suggests that people's actual decision functions for race are continuous. She shows that using income instead of race as the sorting factor also does not lead to clustering, but if it is assumed that both race and income are significant, segregation appears. Thus the model continues to be influential, although it has little or no empirical support, because it remains a fruitful source for theorising and for developing new models. In short, it satisfies the criterion that it is 'valid' because it generates further scientific work.

Summarising the first part of this article, we have argued that a simulation is good when we get from it what we originally would have liked to get from the target. It is good if it works. As Glasersfeld [23] puts it: "Anything goes if it works". The evaluation of the simulation is guided by the expectations, anticipations and experience of the community that uses it – for practical purposes (Caffè Nero), or for intellectual understanding and for building new knowledge (science simulation).

[2] Thomas Nickles claims new work opportunities for sociology at this point: "the job of philosophy is simply to lay out the necessary logico-methodological connections against which the under-determination of scientific claims may be seen; in other words, to reveal the necessity of sociological analysis. Philosophy reveals the depths of the under-determination problem, which has always been the central problem of methodology, but is powerless to do anything about it. Under-determination now becomes the province of sociologists, who see the limits of under-determination as the bounds of sociology. Sociology will furnish the contingent connections, the relations, which a priori philosophy cannot" [20].

Miguel, Amblard, Barceló & Madella (eds.) Advances in Computational Social Science and Social Simulation
Barcelona: Autònoma University of Barcelona, 2014, DDD repository <http://ddd.uab.cat/record/125597>

II. An example of assessing quality

In this part, we will apply and test the assessment mechanisms outlined using as an example our work with the Simulating Knowledge dynamics in Innovation Networks (SKIN) model in its application to research policy modelling.

There are now a number of policy modelling studies using SKIN [24]. We will here refer to just one recent example, on the impact assessment and ex-ante evaluation of European funding policies in the Information and Communication Technologies (ICT) research domain [25].

A. A policy modelling application of SKIN

The basic SKIN model has been described and discussed in detail elsewhere (e.g. [26], [27], [28]). On its most general level, SKIN is an agent-based model where agents are knowledge-intensive organisations, which try to generate new knowledge by research, be it basic or applied, or creating new products and processes by innovation processes. Agents are located in a changing and complex social environment, which evaluates their performance; e.g. the market if the agents target innovation or the scientific community if the agents target publications through their research activities. Agents have various options to act: each agent has an individual knowledge base called its "kene", cf. [29], which it takes as the source and basis for its research and innovation activities. The agent kene is not static: the agent can learn, either alone by doing incremental or radical research, or from others, by exchanging and improving knowledge in partnerships and networks. The latter feature is important, because research and innovation happens in networks, both in science and in knowledge-intensive industries. This is why SKIN agents have a variety of strategies and mechanisms for collaborative arrangements, i.e. for choosing partners, forming partnerships, starting knowledge collaborations, creating collaborative outputs, and distributing rewards. Summarising, usually a SKIN application has agents interacting on the knowledge level and on the social level while both levels are interconnected. It is all about knowledge and networks.

This general architecture is quite flexible, which is why the SKIN model has been called a "platform", cf. [30], and has been used for a variety of applications ranging from the small such as simulating the Vienna biotech cluster [31] to intermediate such as simulating the Norwegian defence industry [32], to large-scale applications such as the EU-funded ICT research landscape in Europe [25]. We will use the latter study as an example after explaining why the SKIN model is appropriate for realistic policy modelling in particular.

The birth of the SKIN model was inspired by the idea of bringing a theory on innovation networks, stemming mainly from innovation economics and economic sociology, onto the computer – a computer theory, which can be instantiated, calibrated, tested and validated by empirical data. In 1998, the first EU project developing the model "Simulating Self-Organizing Innovation Networks" (SEIN) consisted of a three-step procedure: theory formation, empirical research collecting data both on the quantitative and on the case study level, and agent-based modelling implementing the theory and using the data to inform the model [33].

This is why the SKIN model applications use empirical data and claim to be "realistic simulations" insofar as the aim is to derive conclusions by "inductive theorising". The quality of the SKIN simulation derives from an interaction between the theory underlying the simulation and the empirical data used for calibration and validation. In what way does the SKIN model handle empirical data? We will now turn to our policy modelling example to explain the data-to-model workflow, which is introduced in greater detail in [34]

B. Policy modelling for ex-ante evaluation of EU funding programmes

The INFSO-SKIN application, developed for the Directorate General Information Society and Media of the European Commission (DG INFSO), was intended to help to understand and manage the relationship between research funding and the goals of EU policy. The agents of the INFSO-SKIN application are research institutions such as universities, large diversified firms or small and medium-sized enterprises (SMEs). The model simulated real-world activity in which the Calls of the Commission specify the composition of consortia, the minimum number of partners, and the length of the project; the deadline for submission; a range of capabilities, a sufficient number of which must appear in an eligible proposal; and the number of projects that will be funded. The rules of interaction and decision implemented in the model corresponded to Framework Programme (FP) rules; to increase the usefulness for policy designers, the names of the rules corresponded closely to Framework Programme terminology. For the Calls 1-6 that had occurred in FP7, the model used empirical information on the number of participants and the number of funded projects, together with data on project size (as measured by participant numbers), duration and average funding. Analysis of this information produced data on the functioning of, and relationships within, actual collaborative networks within the context of the Framework Programme. Using this data in the model provided a good match with the empirical data from EU-funded ICT networks in FP7: the model accurately reflected what actually happened and could be used as a test bed for potential policy choices, cf. [25].

Altering elements of the model that equate to policy interventions such as the amount of funding, the size of consortia, or encouraging specific sections of the research community, enabled the use of INFSO-SKIN as a tool for modelling and evaluating the results of specific interactions between policies, funding strategies and agents. Because changing parameters within the model is analogous to applying different policy options in the real world, the model could be used to examine the likely real-world effects of different policy options before they were implemented.

As will be seen in Figure 1, the first contact with "the real world" had already occurred in the definition phase of the project. What do the stakeholders want to know in terms of policies for a certain research or innovation network? Identifying relevant issues, discussing interesting aspects about them, forming questions and suggesting hypotheses for potential answers was a first important step. It aimed to conclude with a finite set of questions and concrete designs of experiment with which to address them with the model. This was an interactive and participative process between the study team, which knew about the possibilities and limitations of the model, and the stakeholders, who could be assumed to know what are the relevant issues in their day-to-day practice of policymaking.

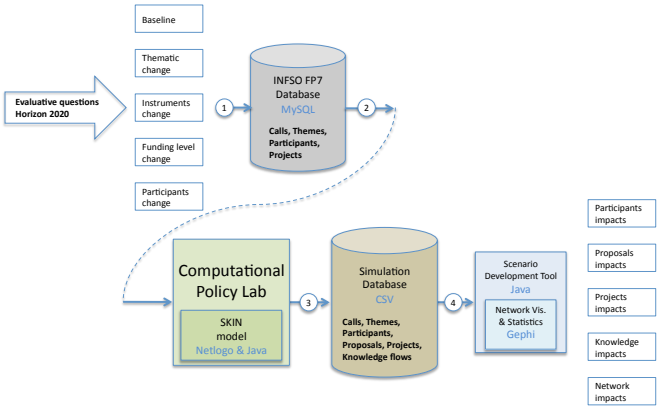

Fig. 1 Horizon 2020 study workflow [34]

After discussing the evaluative questions for the ex-ante evaluation part of this study with the stakeholders from DG INFSO, the following questions were singled out for experiments:

1. What if there are no changes, and funding policies of DG INFSO continued in Horizon 2020 as they were in FP7?
2. What if there are changes to the currently eight thematic areas funded in the ICT domain prioritising certain areas in Horizon 2020?
3. What if there are changes to the instruments of funding and fund larger/smaller consortia in Horizon 2020 than in FP7?
4. What if there are interventions concerning the scope or outreach of funding providing much more / much less resource to more / fewer actors?
5. What if there are interventions concerning the participation of certain actors in the network (e.g. SMEs)?

The next step was to collect relevant data to address these questions and hypotheses. This step is not different from the one every empirical researcher is confronted with. To identify relevant variables for operationalising hypotheses, to be as simple as possible but as detailed as necessary for description and explanation, is in line with the requirements of all empirical social research. For SKIN, the most important type of data is about knowledge dynamics (e.g. knowledge flows, amount of knowledge, diversity of knowledge) and its indicators (e.g. publications, patents, innovative ideas etc.), and about dynamics concerning actors, networks, their measures, and their performance (e.g. descriptive statistics on actors, network analysis measures, aggregate performance data).

These data were used to calibrate the initial knowledge bases of the agents, the social configurations of agents ("starting networks"), and the configuration of an environment at a given point in time. DG INFSO provided the data needed to calibrate the knowledge bases of the agents (in this case the research organisations in the European research area), the descriptive statistics on agents and networks and their interactions (in this case data on funded organisations and projects in ICT under FP7).

The time series data were used to validate the simulations by comparing the empirical data with the simulation outputs. Once we were satisfied with the model performance in that respect, experiments were conducted and the artificially-produced data analysed and interpreted. The stakeholders were again invited to provide their feedback and suggestions about how to fine-tune and adapt the study to their changing user requirements as the study proceeded.

The last step was again stakeholder-centered as it involved visualisation and communication of data and results. We had to prove the credibility of the work and the commitment of the stakeholders to the policy modelling activity.

We worked from an already existing application of the SKIN model adapted to the European research area [35], implemented the scenarios according to the evaluative questions and produced artificial data as output of the simulations. The results are reported in the Final Report, which presented to the European Cabinet, and were communicated to the stakeholders at DG INFSO.

C. The INFSO-SKIN example as seen by the Standard view

The standard view refers to *verification*, namely whether the code does what it is supposed to do, and *validation*, namely whether the outputs (for given inputs/parameters) sufficiently resemble observations of the target.

In terms of verification, the Horizon 2020 application has passed the test as far as this can go. Without claiming that realistic policy modelling always has to employ the standard view perspective, our study of course relies on a realist perspective because it refers to the observability of reality in order to compare the 'real' with artificial data produced by the simulation.

For addressing the evaluative questions of the stakeholders, we needed to create a simulation resembling their own world as observed as "empirical reality". The simulation needed to create the effect of similar complexity, similar structures and processes, and similar objects and options for interventions. To be under this *similarity threshold* would have led to the rejection of the model as a "toy model" that is not realistic and is under-determined by empirical data. In the eyes of these stakeholders, the more

features of the model can be fed with and validated against empirical data points the better. Of course, there will be always an empirical "under-determination" of the model due to the necessary selection and abstraction process of model construction, empirical un-observables, missing data for observables, random features of the model and so on. However, to find the "right" trade-off between empirical under-determination and model credibility was a crucial issue in the discussions between the study team and the stakeholders.

D. The INFSO-SKIN example as seen by the Constructivist view

The strength of a modelling methodology lies in the opportunity to ask what-if questions (ex-ante evaluation), an option that is normally not easily available in the policy-making world. INFSO-SKIN uses scenario modelling as a worksite for 'reality constructions', in line with Gellner's statement quoted above about the constructivist approach: "In dealing with experience, in trying to explain and control it, we accept as legitimate and appropriate to experiment with different conceptual settings, to combine the flow of experience to different 'objects'" [17]. Scenario modelling was employed in the study both for the impact assessment of existing funding policies, where we measured the impact of policy measures by experimenting with different scenarios where these policies are absent, changed or meet different conditions, and for ex-ante evaluation, where we developed a range of potential futures for the European Research Area in ICT by asking what-if questions.

These are *in silico* experiments, which construct potential futures. Is this then a relativist approach where "anything goes", because everything is just a construction? For the general aspects of this question we refer to Part I of this article. There we talk about the "reality requirements" of the constructivist approach, which mediates its claims.

E. The INFSO-SKIN example as seen by the User Community View

The user community view is the most promising, and in our eyes, the most work-intensive mechanism to assess the quality of this policy modelling exercise.

1) Identifying user questions

In our example, SKIN is applied to a tender study with a clear client demand behind it, where the questions the simulation needs to answer was more or less pre-defined from the onset of the project. Enough time should, however, be dedicated to identifying and discussing the exact set of questions the stakeholders of the work want to see addressed. We found that the best way to do this is applying an iterative process of communication between study team and clients, where stakeholders learn about the scope and applicability of the methods, and where researchers get acquainted with the problems policy makers have to solve and with the kind of decisions, for which sound background information is needed. This iterative process will result in an agreed set of questions for the simulation, which will very often decisively differ from the set proposed at the start of the study. For our example, a so-called "Steering Committee" was assigned to us by the European Commission consisting of policy makers and evaluation experts of DG INFSO.

Evaluative questions can address both, the knowledge and the network level. For example, the agreed set of evaluative questions for the INFSO-SKIN application only contained one question for the knowledge level (the first one) and various questions for the agents/networks level (see list of evaluative questions above).

There are various difficulties and limitations to overcome in identifying user questions. In the case of the DG INFSO study, though the questions under study were outlined in the Tender Specifications in great detail, this was a complicated negotiation process where the stakeholder group

- Had to find out about the exact nature and direction of their questions while they talked to the study team
- Had questioned the original set of the Tender Specifications in the meantime and negotiated among each other for an alternative set
- Did not share the same opinion about what questions should be in the final sample, and how potential questions should be ranked in importance
- Did not share the same hypotheses about questions in the final sample

The specification of evaluative questions might be the first time stakeholders talk to each other and discuss their viewpoints.

What is the process for identifying user questions for policy modelling? In the INFSO-SKIN application, the following mechanism was used by the study team and proved to be valuable:

- Scan written project specification by client (in this case the Tender Specifications of DG INFSO) and identify the original set of questions
- Do literature review and context analysis for each question (policy background, scope, meaning etc.) to inform study team
- Meet stakeholders to get their views on written project specifications and their view on context of questions; inform the stakeholders about what your model is about, what it can and cannot do; discuss until stakeholder group and study team is "on the same page"
- Evaluate meeting and revise original set of questions if necessary (probably an iterative process between study team and different stakeholders individually where study team acts as coordinator and mediator of the process)
- Meet stakeholders to discuss final set of questions, get written consent on this, and get their hypotheses concerning potential answers and potential ways to address the questions

- Evaluate meeting and develop experiments that are able to operationalise the hypotheses and address the questions
- Meet stakeholders and get their feedback and consent that experiments meet questions/hypotheses
- Evaluate meeting and refine experiment set-up concerning final set of questions

This negotiation and discussion process is highly user-driven, interactive, and iterative. It requires communicative skills, patience, willingness to compromise on both sides, and motivation to make both ends meet – the formal world of modellers and the narrative world of policymaking in practice. The process is highly time-consuming. In our example, we needed about six months of a 12-month contract research study to get to satisfactory results on this first step, before the simulation had even started.

F. Getting their best: users need to provide data

The study team will know best what type of empirical data would be supportive to inform the policy modelling activity. In SKIN, data availability is an important issue, because the findings have to be evidence-based and realistic. This is in the best interest of the stakeholders, who need to trust the findings, which will be the more the case when the simulated data resembles the empirical data known to the user. However, the study team might discover that data as desired is not available, either not existing or not willingly released by the stakeholders or whoever holds it.

In our example, the stakeholders were data collectors on a big scale themselves. The evaluation unit of DG INFSO employs a data collection group that provides information about funded projects and organisations at a detailed level. Furthermore, the DG often provides data to study teams of the tender projects they contract for their evaluation projects. This is why our example we had a luxurious and clean database concerning all issues the study team was interested in.

However, it was still an issue to confirm the existence, quality and availability of the data and check for formats and database requirements. Even if the data are there in principal, enough time should be reserved for such issues. The quality of the simulation in the eyes of the user will very much depend on the quality of the informing data and the quality of the model calibration.

What would have been the more common process if the study team had not struck lucky as in our example? In other SKIN applications, the following mechanism was used by the study team and proved to be valuable (the ones with asterisks also apply to our INFSO-SKIN example):

- Identify the rough type of data required for the study from the project specifications
- Estimate financial resources for data access in the project proposal to stakeholders (this can sometimes happen in interaction with the funding body)
- After the second meeting with stakeholders, identify the relevant data concerning variables to answer the study questions and address/test hypotheses*
- Communicate exact data requirements to stakeholders, who are usually experts on their own empirical data environment*
- Review existing data bases including the ones stakeholders might hold or can get access to*
- Meet stakeholders to discuss data issues; make them understand and agree on the scope and limitations of data access*
- If needed and required by stakeholders, collect data
- Meet stakeholders to discuss the final database
- Evaluate the meeting and develop data-to-model procedures*

G.Interacting with users to check the validity of simulation results

The stakeholders put heavy demands on the study team concerning understanding and trusting the simulation findings. The first and most important is that the clients want to understand the model. To trust results means to trust the process that produced them. Here, the advantage of the adapted SKIN model is that it relies on a narrative that tells the story of the users' every-day world of decision-making. In the SKIN model, a good example for "reality" requirements is the necessity to model the knowledge and behaviour of agents. Blackboxing the knowledge of agents or creating merely reactive simple agents would not have been an option, because stakeholders do not think the world works that way.

As mentioned, the SKIN model is based on empirical quantitative and qualitative research in innovation economics, sociology, science and technology studies, and business studies. Agents and behaviours are informed by what we know about them; the model is calibrated by data from this research. We found that there is a big advantage in having a model where stakeholders can recognise the relevant features they see at work in their social contexts. In setting up and adapting the model to study needs, stakeholders can actively intervene and ask for additional agent characteristics or behavioural rules; they can refine the model and inform blackbox areas where they have information on the underlying processes.

However, here again, we encountered the diversity of stakeholder preferences. Different members of the DG INFSO Steering Committee opted for different changes and modifications of the model. Some were manageable within the given time constraints and financial resources; some would have outlived the duration of the project if realised. The final course of action for adapting the model to study needs was the result of discussions between stakeholders about model credibility and increasing complexity and of discussions between stakeholders and the study team concerning feasibility and reducing complexity.

Once the stakeholders were familiar with the features of the model and had contributed to its adaptation to study

requirements, there was an initial willingness to trust model findings. This was strengthened by letting the model reproduce FP7 data as the baseline scenario that all policy experiments would be benchmarked against. If the networks created by real life and those created by the agent-based model qualitatively correspond closely, the simulation experiments can be characterized as 'history-friendly', which reproduce the empirical data and cover the decisive mechanisms and resulting dynamics of the real networks (see the *standard view*).

In presenting the results of the INFSO-SKIN study, however, it became clear that there were, again, certain caveats coming from the user community. The policy analysts did not want to look at a multitude of tables and scan through endless numbers of simulation results for interesting parameters; nor did they expect to watch the running model producing its results, for example during a presentation, because one run would last 48 hours. Presenting results in an appealing and convincing way required visualisations and interactive methods where users could intuitively understand what they see, had access to more detailed information if they wanted, e.g. in a hyperlink structure, and could decide for themselves in which format, in which order and in what detail they wanted to go through findings. This part of the process still needs further work: new visualisation and interactive technologies can help to make simulation results more accessible to stakeholders.

This leads to the last issue to be discussed in this section. What happens after the credibility of simulation results is established? In the INFSO-SKIN study, the objective was policy advice for Horizon 2020. The stakeholders wanted the study team to communicate the results as "recommendations" rather than as "findings": They required a so-called "Utility Summary" with statements about what they should do in their policy domain according to study results. Here the study team proved to be hesitant – not due to a lack of confidence in their model but due to (i) an understanding of its predictive limitations and (ii) an apprehension about normative statements, which were seen as a matter of political opinion and not as part of the scientific advisor role. The negotiations of wording in the Utility Summary again afforded an intense dialogue between stakeholders and study team. Nevertheless, the question whether the results had an influence on or were somehow useful in the actual political process of finalising Horizon 2020 policies was not part of the stakeholder feedback after the study ended. The feedback consisted of the formal approval of having fulfilled the contract of the policy advice project.

III. Conclusions

To trust the quality of a simulation means to trust the process that produced its results. This process is not only the one incorporated in the simulation model itself. It is the whole interaction between stakeholders, study team, model, and findings.

The first section of this contribution pointed out the problems of the Standard View and the Constructivist View in evaluating social simulations. We argued that a simulation is good when we get from it what we originally would have liked to get from the target; in this, the evaluation of the simulation would be guided by the expectations, anticipations and experience of the community that uses it. This would make the user community view the most promising mechanism to assess the quality of a policy modelling exercise.

The second section looked at a concrete policy modelling example to test this assumption. It showed that the very first negotiation and discussion with the user community to identify their questions was highly user-driven, interactive, and iterative. It required communicative skills, patience, willingness to compromise on both sides, and motivation to make both ends meet – the formal world of modellers and the narrative world of policymaking in practice.

Often, the user community is involved in providing data for calibrating the model. It is not an easy issue to confirm the existence, quality and availability of data and check for formats and database requirements. As the quality of the simulation in the eyes of the user will very much depend on the quality of the informing data and the quality of the model calibration, much time and effort need to be spent in coordinating this issue with the user community.

Last but not least, the user community has to check the validity of simulation results and has to believe in their quality. Users have to be enabled to understand the model, to agree with its processes and ways to produce results, to judge similarity between empirical and simulated data etc.

Summarising, in our eyes, the User Community view might be the most promising, but definitely is the most work-intensive mechanism to assess the quality of a simulation. It all depends on who the user community is and its composition. If there is more than one member, the user community will never be homogenous. It is difficult to refer to a "community" if people have radically different opinions.

Furthermore, there are all sorts of practical contingencies to deal with. People might not be interested, or they might not be willing or able to dedicate as much of their time and attention to the study as is needed. There is also the time dimension: the users at the end of a simulation project might not be the same as those who initiated it, because of job changes, resignations, promotions and organisational restructuring. Moreover, the user community and the simulation modellers may affect each other, with the modellers helping in some ways to construct a user community in order to solve the practical contingencies that get in the way of assessing the quality of the simulation, while the user community may in turn have an effect on the modellers (not least in terms of influencing the financial and recognition rewards the modellers receive).

If trusting the quality of a simulation indeed means trusting the process that produced its results, then we need to address the entire interaction process between user

community, researchers, data, model, and findings as the relevant assessment mechanism. Researchers have to be aware that they are co-designers of the mechanisms they need to participate in with the user community for assessing the quality of a social simulation.

REFERENCES

[1] Doran, J. and N. Gilbert (1994): Simulating Societies: an Introduction. In: J. Doran and N. Gilbert (eds.): *Simulating Societies: the Computer Simulation of social Phenomena*. London: UCL Press, pp. 1-18.

[2] Norris, C. (1992): *Uncritical Theory*. London: Lawrence and Wishart.

[3] Baudrillard, J. (1988): *Jean Baudrillard Selected Writings*. Cambridge: Polity Press.

[4] Ahrweiler, P. and Gilbert, N. (2005) 'Caffe Nero: the Evaluation of Social Simulation'. *Journal of Artificial Societies and Social Simulation*, 8 (4), 14.

[5] Gilbert, N. and K. Troitzsch (1997): *Simulation for the Social Scientist*. Buckingham, Ph.: Open University Press.

[6] Quine, W. (1977): *Ontological Relativity*. Columbia: Columbia University Press.

[7] Harbodt, S. (1974): *Computersimulationen in den Sozialwissenschaften*. Reinbek: Rowohlt. (Computer Simulations in the Social Sciences), pp. 258f.

[8] Balzer, W., C.U. Moulines und J.D. Sneed (1987): *An Architectonic for Science. The structuralist Program*. Dordrecht etc. Reidel.

[9] Carrier, M. (1994): *The Completeness of scientific theories. On the Derivation of empirical Indicators within a theoretical framework: The Case of Physical Geometry*. Dordrecht etc.: Kluwer.

[10] Axelrod, R. (1984): *The Evolution of Cooperation*. New York: Basic Books.

[11] Deffuant, G., Neau, D., Amblard, F. and Weisbuch, G. (2000): Mixing beliefs among interacting agents. Advances in Complex Systems. In: *Adv. Complex Syst*. 3, pp. 87-98.

[12] Ben-Naim, E., Krapivsky, P. and Redner, S. (2003): Bifurcations and Patterns in Compromise Processes. In: *Physica D* 183, pp. 190-204.

[13] Weisbuch, G. (2004): Bounded confidence and social networks. In: *Eur. Phys. J. B,* Special Issue: Application of Complex Networks in Biological Information and Physical Systems volume 38, pp.339-343.

[14] Cole, O. (2000): White-box testing. *Dr. Dobb´s Journal*, March 2000, pp. 23-28.

[15] Chalmers, D., R. French and D. Hofstadter (1995): High-Level Perception, Representation, and Analogy. ID. Hofstadter (ed.): *Fluid Concepts and Creative Analogies*. New York: Basic Books, pp. 165-191.

[16] Droste, W. (1994): *Sieger sehen anders aus*. Hamburg: Schulenburg. (Winners look different)

[17] Gellner, E. (1990): Pflug, *Schwert und Buch. Grundlinie der Menschheitsgeschichte*. Stuttgart: Klett-Cotta. (Plough, Sword and Book. Foundations of Human History)

[18] Searle. J. (1997): *The Construction of Social Reality*. Free Press.

[19] Kértesz, A. (1993): *Artificial Intelligence and the Sociology of Scientific Knowledge.* Frankfurt: Lang.

[20] Nickles, T. (1989): Integrating the Science Studies Disciplines. S. Fuller, M. de Mey, T. Shinn and S. Woolgar (eds.): *The Cognitive Turn. Sociological and Psychological Perspectives on Science*. Dordrecht : Kluwer, pp. 225-256.

[21] Benenson, I. (2005) The city as a Human-driven System. Paper presented at the workshop on Modelling Urban Social Dynamics, University of Surrey, Guildford, UK, April 2005.

[22] Bruch, E. (2005) Dynamic Models of Neighbourhood Change. Paper presented at the workshop on Modelling Urban Social Dynamics, University of Surrey, Guildford, UK, April 2005.

[23] Glasersfeld, E. von (1987): Siegener Gespräche über Radikalen Konstruktivismus. In: S.J. Schmidt (ed.): *Der Diskurs des Radikalen Konstruktivismus*. Frankfurt/M.: Suhrkamp, pp. 401-440. (Siegen Diskussions on Radical Constructivism)

[24] Gilbert, N., Ahrweiler, P. and Pyka, A. (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York.

[25] Ahrweiler, P., Pyka, A. and Gilbert, N. (2014b, forthcoming): Simulating Knowledge Dynamics in Innovation Networks: An introduction. In: Gilbert, N., Ahrweiler, P. and A. Pyka (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York.

[26] Pyka, A., Gilbert, N. and Ahrweiler, P. (2007) 'Simulating Knowledge Generation and Distribution Processes in Innovation Collaborations and Networks'. *Cybernetics and Systems*, 38 (7):667–693.

[27] Gilbert, N., Ahrweiler, P. and Pyka, A. (2007) 'Learning in Innovation Networks: Some Simulation Experiments'. *Physica A: Statistical Mechanics and Its Applications*, 378 (1):667–693.

[28] Ahrweiler, P., Pyka, A., Gilbert, N. (2011) 'A New Model for University-Industry Links in Knowledge-Based Economies'. *Journal of Product Innovation Management*, 28:218–235.

[29] Gilbert, N. (1997) A simulation of the structure of academic science, Sociological Research Online, 2(1997), 3. http://www.socresonline.org.uk/socresonline/2/2/3.html.

[30] Ahrweiler, P., Schilperoord, M., Pyka, A. and Gilbert, N. (2014a, forthcoming): Testing Policy Options for Horizon 2020 with SKIN. In: Gilbert, N., Ahrweiler, P. and A. Pyka (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York

[31] Korber, M. and M. Paier (2014): Simulating the Effects of Public Funding on Research in Life Sciences: Direct Research Funds versus Tax Incentives. in: Gilbert, N., Ahrweiler, P. and Pyka, A. (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York.

[32] Castelacci, F., A. Fevolden and M. Blom (2014): R&D Policy Support and Industry Concentration: A SKIN Model Analysis of the European Defence Industry. in: Gilbert, N., Ahrweiler, P. and Pyka, A. (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York.

[33] Pyka, A., Gilbert, N., Ahrweiler, P. (2003), Simulating Innovation Networks, in: Pyka, A. and Küppers, G. (eds.), *Innovation Networks – Theory and Practice*, Edward Elgar, Cheltenham, UK, 169-198.

[34] Schilperoord, M. and Ahrweiler, P. (2014, forthcoming): Towards a prototype policy laboratory for simulating innovation networks. In: Gilbert, N., Ahrweiler, P. and A. Pyka (eds.) (2014, forthcoming): *Simulating Knowledge Dynamics in Innovation Networks.* Springer: Heidelberg / New York.

[35] Scholz, R., Nokkala, T., Ahrweiler, P., Pyka, A., Gilbert, N. (2010), The agent-based NEMO model (SKEIN): simulating European Framework Programmes. Ahrweiler, P. (ed.), *Innovation in Complex Social Systems*, Routledge Studies in Global Competition, 300-314.