©Copyright JASSS

Petra Ahrweiler and Nigel Gilbert (2005)

Caffè Nero: the Evaluation of Social Simulation

Journal of Artificial Societies and Social Simulation vol. 8, no. 4
<http://jasss.soc.surrey.ac.uk/8/4/14.html>

For information about citing this article, click here

Received: 02-Oct-2005 Accepted: 02-Oct-2005 Published: 31-Oct-2005

Abstract

This contribution deals with the assessment of the quality of a simulation by discussing and comparing "real-world" and scientific social simulations. We use the example of the Caffè Nero in Guildford as a 'real-world' simulation of a Venetian café. The construction of everyday simulations like Caffè Nero has some resemblance to the construction procedure of scientific social simulations. 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. After briefly discussing the 'ordinary' method of evaluating simulations called the 'standard view' and its adversary, a constructivist approach asserting that 'anything goes', we heed these similarities in the construction process and apply evaluation methods typically used for everyday simulations to scientific simulation and vice versa. The discussion shows that a 'user community view' creates the foundation for every evaluation approach: when evaluating the Caffè Nero simulation, we refer to the expert community (customers, owners) who use the simulation to get from it what they would expect to get from the target; similarly, for science, the foundation of every validity discussion is the ordinary everyday interaction that creates an area of shared meanings and expectations. Therefore, the evaluation of a 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).

Keywords:
Evaluation, Social Simulation, Standard View, Constructivist View, User Community

Introduction

1.1

This contribution deals with the assessment of the quality of a simulation. 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 Doran and Gilbert 1994).

1.2

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. Norris 1992), or as Baudrillard put it, "to simulate is to feign to have what one hasn´t" while "substituting signs for the real" (Baudrillard 1988). We use 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.

Figure 1. Interior view of Caffè Nero

1.3

The construction of everyday simulations like Caffè Nero has some resemblance to the construction of scientific social simulations (see Table 1):

Table 1: 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?

1.4

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.

1.5

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?

1.6

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". The standard view

2.1

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 outputs are not to be expected, as discussed in detail in Gilbert and Troitzsch (1997). 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.

2.2

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 a 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.

2.3

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.

The problem of underdetermination

2.4

Some philosophers of science argue that theories are under-determined by observational data or experience (for a summary of the discussion see Carrier 1994). 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.

2.5

However, the problem that theories are underdetermined by empirical data cannot be solved by cumulative data gathering: it is more general and therefore more serious. The underdetermination problem is not about a missing quantity of data but about the relation between data and theory. As Quine (1977) 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" (Harbodt 1974: 258f, own translation).

2.6

The proposal mentioned above to solve the underdetermination 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.

The theory-ladenness of observations

2.7

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" (Balzer et al. 1987) 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 misled from the beginning.

2.8

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" (1987: 48). 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. Carrier 1994: 1-19). These problems apply equally to simulations, which are just theories in process.

2.9

We can give examples of these difficulties in the area of social simulation. To compare Axelrod´s The evolution of cooperation (Axelrod 1984) 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. Deffuant, Neau, Amblard and Weisbuch 2000; Ben-Naim, Krapivsky and Redner 2003; Weisbuch 2004). 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?

2.10

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.

2.11

There are also more technical problems. Validity tests should be "exercised over a full range of inputs and the outputs are observed for correctness" (Cole 2000: 23). 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).

2.12

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.

The constructivist view

3.1

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" (Chalmers, French and Hofstadter 1995: 173).

3.2

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 divisable like the coat of Saint Martin; everybody gets a piece of it and everybody has a nice feeling" (Droste 1994:50).

3.3

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'" (Gellner 1990: 75).

3.4

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, rhetorics and arbitrary talk. Can we so easily dismiss the possibility of evaluation?

The user community view

4.1

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* (1997) how conventions are "real": they are not deficient for the support of a relativistic approach because they are constructed.

4.2

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".

4.3

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" (Kertész 1993: 32).

4.4

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?

4.5

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 be a model that demonstrates 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.

4.6

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 (2005) 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.

4.7

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 (2005) 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.

4.8

In this paper, 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 (1987:429) 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).

Notes

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

[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 underdetermination of scientific claims may be seen; in other words, to reveal the necessity of sociological analysis. Philosophy reveals the depths of the underdetermination problem, which has always been the central problem of methodology, but is powerless to do anything about it. Underdetermination now becomes the province of sociologists, who see the limits of underdetermination as the bounds of sociology. Sociology will furnish the contingent connections, the relations, which a priori philosophy cannot" (Nickles 1989: 234f).

References

AXELROD, R. (1984): *The Evolution of Cooperation*. New York: Basic Books.

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

BAUDRILLARD, J. (1988): *Jean Baudrillard Selected Writings*. Cambridge: Polity Press.

BEN-NAIM, E., Krapivsky, P. and Redner, S. (2003): Bifurcations and Patterns in Compromise Processes. In: *Physica D* 183, pp. 190-204.

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.

BRUCH, E. (2005) Dynamic Models of beighbourhood Change. Paper presented at the workshop on Modelling Urban Social Dynamics, University of Surrey, Guildford, UK, April 2005.

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.

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

COLE, O. (2000): White-box testing. In: *Dr. Dobb´s Journal*, March 2000, pp. 23-28.

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.

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.

DROSTE, W. (1994): *Sieger sehen anders aus*. Hamburg: Schulenburg. (Winners look different)

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

GILBERT, N. and K. Troitzsch (1997): *Simulation for the Social Scientist*. Buckingham.: Open University Press.

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)

HARBODT, S. (1974): *Computersimulationen in den Sozialwissenschaften*. Reinbek: Rowohlt. (Computer Simulations in the Social Sciences)

KÉRTESZ, A. (1993): *Artificial Intelligence and the Sociology of Scientific Knowledge*. Frankfurt: Lang.

NICKLES, T. (1989): Integrating the Science Studies Disciplines. In: 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.

NORRIS, C. (1992): *Uncritical Theory*. London: Lawrence and Wishart.

QUINE, W. (1977): *Ontological Relativity*. Columbia: Columbia University Press.

SEARLE. J. (1997): *The Construction of Social Reality*. Free Press.

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.

Return to Contents of this issue

© Copyright Journal of Artificial Societies and Social Simulation, [2005]