Synthese (2009) 169:539–555
DOI 10.1007/s11229-008-9429-0

The explanatory potential of artificial societies

Till Grüne-Yanoff

Received: 31 October 2006 / Accepted: 13 October 2007 / Published online: 18 November 2008

Abstract It is often claimed that artificial society simulations contribute to the explanation of social phenomena. At the hand of a particular example, this paper argues that artificial societies often cannot provide full explanations, because their models are not or cannot be validated. Despite that, many feel that such simulations somehow contribute to our understanding. This paper tries to clarify this intuition by investigating whether artificial societies provide potential explanations. It is shown that these potential explanations, if they contribute to our understanding, considerably differ from potential causal explanations. Instead of possible causal histories, simulations offer possible functional analyses of the *explanandum*. The paper discusses how these two kinds explanatory strategies differ, and how potential functional explanations can be appraised.

Keywords Agent-based simulations · Complex systems · Explanation

1 Introduction

Artificial societies are often claimed to be explanatory (Axtell et al. 2002; Cederman 2005; Dean et al. 1999; Epstein 1999; Sawyer 2004; Tesfatsion 2007). Often these claims are ambiguous about how agent-based simulations are explanatory, and what they explain. In this paper, I show that an important class of agent-based simulations cannot fully explain a phenomenon. I further argue that agent-based simulations do not contribute to our understanding of a phenomenon by presenting its possible causal histories. Instead, I develop an account of possible functional explanations, and show how

T. Grüne-Yanoff (✉)
Collegium of Advanced Studies, Helsinki University, Helsinki, Finland
e-mail: till.grune@helsinki.fi

agent-based simulations can provide such potential explanations by offering possible functional analyses of a phenomenon.

Artificial societies simulate social phenomena. Phenomena are things in the world that are identifiable by the data they produce, but which are rarely observable themselves. For example, the history of a tribe is a large-scale social phenomenon that is evidenced by all sorts of documents: written record, eyewitness reports, pottery shards, ruins, etc. To simulate such a phenomenon is to construct a process whose output in relevant ways imitates the 'target' data that represents this phenomenon.[1]

Artificial societies simulate social phenomena with agent-based models. In such models, an aggregate state of the simulating system is determined by the states of individual agents. Each agent (which may represent people, firms, nation-states, etc.) is characterised by a number of attributes and a set of behavioural rules. Agents are *heterogeneous*, because the model can specify different attributes for different agents.[2] Agents are *autonomous*, because their interactions are determined by their individual behavioural rules (e.g. when to migrate, or how to estimate a future parameter), not by any global rule covering all simultaneously. Agents influence the environment through their actions, but are in turn influenced by the environment they and their peers create. The simulation imitates the target data by computing the individual agents' behaviour in response to some input environmental data, by computing the effects of the individual behaviours on the environment, and by computing the repercussions these environmental effects have on individual agents.

Epstein and Axtell (1996), who popularised the term 'Artificial Societies', showed how manipulating the attributes and behaviour rules of the model agents allows the generation of patterns akin to migration, markets, wars, etc. However, the similarity is fleeting and can be seen only by abstracting from many features of real-world phenomena. Because these simulations do not imitate the target data of any particular phenomenon, it seems implausible to claim that they would explain any such phenomenon.

This changed with the publication of papers that explicitly purported to simulate particular real-world phenomena by imitating their target data. Such simulations, it is claimed, *explain* the phenomena or essentially contribute to their explanation. By essential contribution, it is meant that generation is necessary for explanation, according to the motto 'If you didn't grow it, you didn't explain its emergence' (Epstein 1999, p. 43).

Section 2 presents an example of such a purported explanation. Section 3 argues that the example, as well as simulations of its kind, lacks the evidential support necessary for full causal explanations. Section 4 discusses the claim that simulations offer potential explanations. It argues that they do not contribute to our understanding of the phenomenon by providing possible causal histories; but instead may contribute to our understanding by providing possible functional analyses. The difference between

[1] The underlying aim of the simulation is therefore to imitate the real-world process that produced this data (cf. Hartmann 1996; Humphreys 2004).

[2] Of course, even homogeneous agents may be in different states at any given time: for example, they will be at different spatial locations. Heterogeneity of agents, in contrast, implies that agents differ in their fundamental propensities—e.g. they rate of fertility, fission or death.

potential functional and potential causal explanation is investigated, and criteria for the appraisal of the right possible functional analyses for potential functional explanations are given. Section 5 concludes.

2 An example of generative explanation

The chosen example purports to generatively explain the history of a pre-historical settlement of Ancestral Puebloans (often called Anasazi) in Long House Valley, northern Arizona, from 800 to about 1300 AD. The computation takes as input paleo-environmental data, including meteorological, groundwater and sediment deposition and fertility estimates for the reconstructed kinds of farmland. On the basis of this input, it reproduces the main features of the settlement's actual history, as witnessed by archaeological evidence—including population ebb and flow, changing spatial settlement patterns, and eventual rapid decline.

The computation from input to output is performed through two kinds of intervening variables. First, a dynamic resource landscape of the studied area is theoretically reconstructed from the paleo-environmental data. In particular, annual potential maize production per hectare is estimated for five different categories of potential farming land. Secondly, annual decisions of (re-)settlement, land cultivation and procreation, as well as annual deaths of household-agents are computed on the basis of the estimated maize crop, agents' attributes and behavioural rules. Agents' attributes (like lifespan, vision, movement capacities, nutrition requirements and storage capacities) are

> derived from ethnographic and biological anthropological studies of historic Pueblo groups and other subsistence agriculturalists throughout the world (Dean et al. 1999, p. 187).

Agents' behavioural rules, governing movement and selection of farming and settling sites are modelled as 'anthropologically plausible rules' (Dean et al. 1999, p. 180)—in effect optimization under very limited information.[3]

The original model (Dean et al. 1999) employs fairly homogenous agent attributes. It reproduces 'the qualitative features of the history', but yields populations that were substantially too large. Attempts to reduce the population in that model by changing agent attributes result in premature population collapse.

In a follow-up paper (Axtell et al. 2002), greater levels of both agent and landscape heterogeneity are incorporated. Individual agents' onset of fertility, household fission and death, and harvest per hectare are drawn from uniform distributions. Increasing heterogeneity improves the 'fit' of the model to the historical record. Fit is measured by calculating the differences between simulated households and historical record

[3] In particular, the behavioural rules are: Agents cease to exist if they cannot secure 800 kg of maize for themselves annually, or if they reach a threshold age. Food intake is determined by harvest yields from farmed plot, and storage from previous years. Households choose to change their farmed plots when harvest estimates (based on current year harvests) and storage combined are insufficient for survival. Households choose most productive available (unfarmed & unsettled) plots that are within 1.6 km of a water source. Households settle on available (unfarmed) locations closest to farmed plots. Households procreate annually (after a maturing period of 16 years) with probability of 0.125.

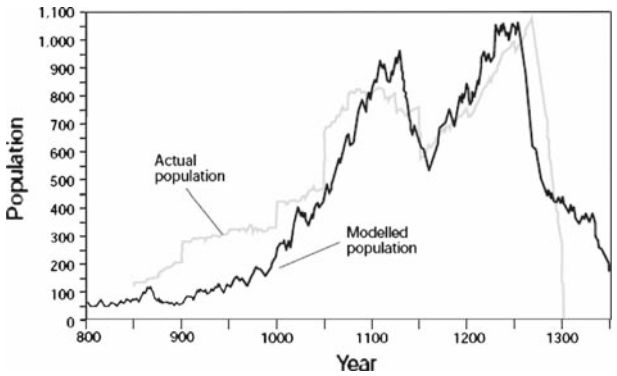

Fig. 1 Best single run of the model according to the $L1$ norm. (c) Nature 2002

for each period. Differences are cumulated according to a stochastic norm (a variant of the standard deviation measure). Depending on which norm is used, optimizing the model with respect to the distribution parameters yields a 'best-fitting' model. The 'best fitting' single run of the model is depicted in Fig. 1.

As shown, this 'best fit' still does not accurately replicate the historical findings. In particular, it simulates a higher population early on, and does not replicate the complete eclipse of the settlement in around 1300. The authors point out that better fits can be achieved by increasing the number of household attributes and their heterogeneity, possibly introducing non-uniform distributions.

The authors of both papers are convinced that their research contributes to the explanation of Anasazi population dynamics:

> Close fit [of the generated data to the observed data] indicates explanatory power (Dean et al. 1999, p. 180).

> Ultimately, "to explain" the settlement and farming dynamics of Anasazi society in Long House Valley is to identify rules of agent behaviour that account for those dynamics (Dean et al. 1999, p. 201).

> To "explain" an observed spatiotemporal history is to specify agents that generate—or grow—this history. By this criterion, our strictly environmental account of the evolution of this society during this period goes a long way toward *explaining* this history (Axtell et al. 2002, p. 7278).

According to these quotes, generating the history of the Ancient Puebloan settlement in an agent-based simulation either explains it or at least contributes to its explanation. Crucially, the simulation itself is claimed to carry the central explanatory role: it is the fit of the generated data, or the identification of generating agents and their rules of behaviour, that purportedly does the explaining. In the following, I investigate a number of possible accounts for this explanatory potential of artificial societies. The Anasazi example is helpful in this, because it lacks, as will be shown in Sects. 3 and 4, certain features that make other kinds of models explanatory.

3 Causal explanation

There are some indicators that the simulation researchers believe they are striving for causal explanation. First, the authors of the Anasazi project suggest that the simulation explains what it generates: a singular event, or a series of singular events in time (i.e. a history). The view that the *generandum* is the *explanandum* is expressed in the above Axtell et al. (2002, p. 7278) quote that growing the history of this society goes a long way toward explaining *that history*. One of the co-authors is even more explicit in another paper:

> This data set [the settlement's history] is the target—the *explanandum* (Epstein 1999, p. 44).

It is widely accepted that to explain an event requires identifying its predominant causes. Hence, the claim that the *generandum* is the *explanandum* implies that artificial societies strive for causal explanation.

Second, some proponents of generative explanation have explicitly claimed that social scientists do and should employ agent-based simulations to

> seek *causal explanations* grounded in the repeated interactions of agents operating in realistically rendered worlds (Tesfatsion 2007, p. 9, my emphasis).

This view is shared by some philosophers:

> The parallels [of artificial society simulations] with causal mechanism approaches in the philosophy of science are striking (Sawyer 2004, p. 222).

While striving for causal explanations with artificial societies is a legitimate goal, the chances of reaching this goal are small. To clarify why, let's compare the present case to a kind of simulation that does provide causal explanations: vehicle crash simulations. These analyze an actual vehicle 'system' into its components, by imposing a three-dimensional grid onto the vehicle and by measuring the relevant properties of each grid cell. Postulating specific impacts, they then calculate the behaviour of these components on the basis of the equations of motion. The macro-effect of the impact on the whole car is thus constituted by its micro-effects on the individual cells. Because the computation of these micro-effects is strictly governed by (well-confirmed) causal regularities, the simulation offers a good causal explanation of specific crash deformations: given the impact, it accounts for how the mechanical forces travelled through the vehicle to the specific area, and what effects they witnessed there. Further, it details the material properties of the specific area, so that it accounts for the fracturing of the relevant area of the windshield, given the impacting forces.

Now, doesn't the same account apply to the Anasazi model? No. Any account of causal explanation requires that the causal regularities included in the *explanans* must be true, or at least well-confirmed. The above car crash simulation bases its explanatory potential on the laws of kinematics, which are well-confirmed and widely believed to be true. Further, it precisely measures the actual vehicle properties. In analogy, agent-based simulations would have to derive their explanatory potential from the agents' behavioural rules applied in a precisely specified environment. The decisive question is what evidence one has to judge these rules to be true.

I think we have little evidential support for them. The fact that they generate the *explanandum* doesn't count much, as many other rules generate it similarly well. For example, similar results are obtained by using individuals of varying ages instead of households as the agents in the model (Axtell et al. 2002, p. 7278). Hence evidential support has to come from sources different than the simulation itself. I consider three potential sources: direct observation, well-confirmed theory, or results from externally valid behavioural experiments. The Ancient Puebloan society has long ceased to exist, and no documents concerning the behavioural rules of their members have been preserved. Direct observation of Ancient Puebloans' behaviour is therefore impossible. The authors instead claim that 'detailed regional ethnographies provide an empirical basis for generating plausible behavioural rules for the agents' (Dean et al. 1999, p. 181). Unfortunately, they do not detail the nature of these ethnographical studies, so that it remains unjustified why the results from these studies may be transferable to the agents under study. Recent research has shown that behavioural rules vary widely among small-scale agricultural societies (Henrich et al. 2004). In particular, this research shows that agents of different contemporary small-scale societies have widely differing attitudes towards mutual help, cooperation and sharing. Behavioural dispositions of this sort may well have significant influences on variables included in the simulation, like fertility, migration and death, particularly in times of crisis. It is therefore questionable whether the similarity in settlement features (e.g. its 'small-scale' property) justifies the transfer of behavioural regularities found in other tribes to the Ancient Puebloans.

This leaves behavioural experiments as a source of evidence for the required causal regularities. Some researchers indeed advocate using experiments this way:

> If we took two microspecifications as competing hypotheses about individual behaviour, then … behavioural experiments might be designed to identify the better hypothesis (microspecification) and, in turn, the better agent model (Epstein 1999, p. 48).

Obviously, there is again a problem of external validity here. The Ancient Puebloans are dead, and who could stand in for them in experiments so that the experimental results would legitimately cover this historical people as well? For the moment, let's bracket this issue in order to see another issue with experimental validation that applies to all artificial societies. Behavioural experiments are performed under strict control of the agent's environment. While this feature insures the exactness of the experimental results, it also limits the applicability of the results to agents in environments different from those controlled for in the experiment.

To ensure the external validity of the relevant experiments, one has to have good grounds to believe that the differences between the experiment and the target system do not create an error in the transfer of results from one to the other. This is a problem for agent-based simulations, because they employ the same behavioural rules under extremely changing environments. Take again the Anasazi model. The agents' behavioural rules are assumed to remain stable throughout (at least) four fundamentally different environments: (i) in a situation where a small group of settlers colonises an unpopulated valley; (ii) in a situation of rapid population increase, where farming density forces new households to occupy low-fertility lands or migrate; (iii) in a

situation of stagnation and slow decline, where environmental factors (draught, strong winters) are perceived as a threat and cultivated plots are given up; and finally (iv) in a situation of cataclysmic decline, where most of the population leaves the settlement area or dies. To transfer results of the experiment to the target system, it would have to be shown that none of these differences mattered.

Beyond the considerable practical problems of performing such experiments, this constitutes a methodological problem. Results from behavioural experiments have up to today not been synthesised to anything like a grand theory with regularities of large scope. Rather, experiments 'contribute to the library of phenomena that the applied scientists will borrow and exploit on a case-by-case basis' (Guala and Mittone 2005, p. 511). However, such piecemeal insights, while instructive for specific cases, do not provide decisive evidence for behavioural regularities required for artificial societies.

For the sake of the argument, let's imagine that experiments could provide decisive evidence for such broad-scoped regularities. What sort of experiments would that have to be? Experiments that would differentiate environments 'finely enough' and test the behavioural rules under all these environmental conditions. But such a gigantic test series, while providing the necessary evidence, would also trivialise the role of agent-based simulations: Because the experiments would have to be run in the all the relevant social environments, experimental design would construct in vivo what simulation would reproduce in silico. All the interesting information could then already be gleaned off the experiments, and there would be no need for simulations anymore at all. Hence, there is little evidential support for the behavioural rules of the Anasazi model at present, and there even are some reasons to believe that such evidence may not be available in principle.

4 Potential explanation

If an agent-based simulation cannot be a full explanation for the reasons spelled out above, it may still *contribute* to an explanation. Some proponents suggest as much:

> If a microspecification, m, generates a macrostructure of interest, then m is a *candidate* explanation (Epstein 1999, p. 43)

This suggestion gives a new meaning to the claims about the simulations' explanatory potential reviewed in Sect. 2. That projects like the Anasazi simulation have 'explanatory power', or that they 'go a long way toward explaining' then does not mean anymore that they provide an explanation. Instead, it is now suggested that they offer a contribution towards an explanation.

It is important to be very clear about this distinction. An explanation does very important things for us: it answers our question about relevant causes, it increases our belief in the *explanandum* in the right way, or it provides a deductive argument for the *explanandum*, etc. To be sure, it is sometimes difficult to adequately describe what exactly an explanation does; but in each particular case, most of us will be able to identify whether a certain cognitive procedure gives an explanation or not. If it does, then the procedure does something that is important to us and therefore merits

our attention. However, once one admits that a certain procedure only contributes to an explanation, or provides a candidate explanation, it is not clear anymore that this procedure merits our attention. The contribution, after all, may be insignificant, or the candidate not worthy of further thought. Interpreting agent-based simulations as only providing contributions or candidates, instead of full explanations, therefore raises the question: why bother? At least for explanatory purposes, these simulations may be insignificant, and their explanatory potential is equal to nil. It is therefore important to clarify what sort of contributions agent-based simulations like the Anasazi model make, and what kind of candidates they offer.

One way to interpret the above claims sympathetically is to see a candidate explanation as an incompletely developed full explanation. This interpretation matches well with the concept of a *potential explanation*, as it is sometimes used in the philosophy of science. Unfortunately, what makes a procedure a potential explanation is either not investigated at all; or, where proposals are made, they remain controversial. I will therefore try to clarify this notion to the extent that it can be made useful for the present discussion.

Hempel provided the first and best-developed notion of potential explanation. He defined a potential explanation as a set of propositions having all the characteristics of an explanation except, possibly, for their truth (Hempel 1965, p. 338). This definition leant on his deductive-nomological account of explanation: a cognitive procedure is a potential explanation, if the *explanandum* is deducible from a set of *lawlike* statements. Statements are lawlike if they are (i) exceptionless, (ii) if they contain purely qualitative predicates, and (iii) if they have a very wide scope. The problems with this account are well known and need not be rehearsed here (for a concise sketch, see Woodward 2003, pp. 154–161). But its rejection leaves us with the problem that it takes away the formal condition for a potential explanation.[4]

The obvious alternative is to account for the simulations' contribution as providing potential *causal* explanations. Modifying Lewis (1986), one may say that agent-based simulations contribute to the explanation of a social phenomenon by providing information about its possible causal histories—specifically, about the possible causes that operate on the micro-level: agents' properties and their behavioural rules. Simulations, one could argue, are particularly good at such a task, because they force researchers to be explicit about all factors and conditions, and because many inconsistencies in the model will become obvious when writing the code.

According to this interpretation, simulations are rigorous practices of articulating the ways a phenomenon could have possibly been produced. Following Lipton (2001, pp. 59–60), such articulations may contribute to our understanding of the phenomenon. Thus, agent-based simulations may be explanatorily worthwhile projects.

[4] In any case, the Anasazi model would satisfy neither criterion (i) nor criterion (iii). Regarding criterion (i), there is no reason to believe that any of these rules are exceptionless. For example, additional criteria like kinship proximity may have been an important criterion of farmland choice. Regarding criterion (iii), the purported scope of the behavioural rules is narrow: it only applies to small-scale subsistence maize agriculturalists in an arid region of the American continents. According to the D-N account, therefore, the Anasazi model would not provide potential explanations, which is explicitly acknowledged by some of the artificial society researchers (e.g. Epstein 1999, fn. 12).

However, from an explanatory point of view, such an articulation has shortcomings. Any collection of such possible histories will be very large. As Axtell et al. (2002, p. 7278) point out, for example, substituting random variables for the current fixed parameters of nutrition needs, birth and death rates, etc., yields simulation results with a fit as close as the original model. Just by varying the parameters, one obtains a large set of possible causal histories. Variation of the agents' behavioural rules further enlarges this set. But the larger the pool of potential explanations, the smaller the contribution to a full explanation. Singling out two or three ways an event could have been produced gets us a big step closer to actually explaining it—all that is needed is to decide between these options, may be by empirical evidence, or by the explanatory virtues they have. Identifying thousands of ways the event could have been produced, however, doesn't get us closer to full explanation at all—all the explanatory work is still left to be done by making a selection from this huge set. The generative richness of agent-based models is thus not an asset, but an embarrassment, as it in fact reduces their explanatory potential.

One may wonder whether there are ways to pre-select potential explanations from the vast pool of possibilities generated by the simulation. The use of empirical research may help in some cases, but as argued in Sect. 3, our capacity to perform the necessary research in cases like the Anasazi simulation is very limited. Instead, what is needed is a 'filter' that selects possible causal histories through criteria that are independent from our evidence for certain causes. If such a filter existed, the resulting small set of alternative possible causal histories might significantly contribute to our understanding of the phenomenon. Alas, the most natural places to look for such a filter turn out to be barren.

Lipton (2001, pp. 83–84) has argued that sometimes the pragmatics of the question to be explained may yield such a selection criterion. Most of our why-questions explicitly or implicitly come with a class of contrastive cases. When answering the question 'why did you shout?', it is important to know whether the inquirer implies '... and not whistle?', or '…instead of remaining quiet?'. To explain the contrast in which the inquirer is interested, one has to identify in which causes the contrasting events differ. Only these differentiating (possible) causes have explanatory relevance for explaining the contrast. From the set of all possible causal histories of the contrasting events, all those histories that do not contain these differentiating possible causes can therefore be eliminated; what remains is a refined set of the potential causal explanations of this specific contrast.

The problem with this selection technique is that it requires the explanatory project to be at least implicitly contrastive. Most why-questions have that form, but the researchers who developed the agent-based simulations commonly do not ask such questions. Rather, as shown in Sect. 2, they want to explain the settlement and farming dynamics, the history and the archaeological data. They use their simulations to answer the question *how* that history developed, *how* the data was generated, and they do not have any contrast in mind beyond the 'how *so,* and not in *any* other way?'. This renders the Lipton's selection technique inapplicable here.

Another approach would invoke *formal* criteria for potential causal explanations. In the style of Hempel, we may hope to describe what causal explanations are, and then specify potential causal explanations as causal explanations, minus, possibly, truth.

However, this approach is fraught with various problems. First, we do not have an uncontroversial descriptive account of causal explanation. Various proposals exist (for example, Salmon's mark-transmission account, and Woodward's counterfactual account), but each of them has its shortcomings, and, importantly, there are many cognitive procedures that fall under none of the theoretical accounts but are widely accepted intuitively as causal explanations.

But even if one could agree on some such conditions, a second problem arises—namely that these conditions are either too wide to perform any selection, or too narrow to allow any possible causal histories to be selected. A common if controversial claim is, for example, that causal explanations identify relevant causal mechanisms. Early attempts to characterise genuine mechanisms are the mark-transmission account (Salmon 1984) and the preserved-quantity account (Salmon 1998). These characterisations, however, use criteria most adept for physical processes. Although the behaviour of agents is realised by physical processes, the agent-based simulations do not describe these physical processes, but instead describe processes on a behavioural and intentional level. It is therefore unclear whether any possible history generated by the simulation satisfies the proposed criteria; hence these criteria are not helpful for the selection task at hand.[5]

More recent accounts of causal-mechanical explanation adopt a much wider account of mechanism. Machamer et al. (2000, p. 3) for example, define mechanism as organised collections of entities and activities that produce regular changes. Under such an account of mechanism, it seems that *all* possible histories generated by the simulation would pass the selection task. Thus, such accounts are not useful for the selection task at hand, because they are too permissive.

Woodward's counterfactual account characterises causal explanation as a matter of exhibiting systematic patterns of counterfactual dependence. Counterfactuals describe the outcomes of interventions: not only do they show that the *explanandum* is to be expected given the initial conditions, but they also show how these *explananda* would change if the initial conditions were changed (Woodward 2003, p. 191). Whether a set of propositions is a potential causal explanation depends on the *invariance* of the counterfactual statement. A generalising statement is invariant across certain changes if it holds up to some appropriate level of approximation across these changes. As Cartwright (2002) showed, such a condition must not be expected to hold universally. Instead, we need independent evidence for the invariance of the relevant counterfactual statements in order to say whether they function as potential explanations. Given that such evidence is hard to come by—as argued in Sect. 3—Woodward's counterfactual account is not useful for the selection task, either.[6]

[5] Salmon explicitly acknowledged this difficulty, but gave it a particular twist. In 'Explanation in Archaeology', for example, he argues that causal explanation in archaeology may be difficult because getting to the details of causal mechanism is a problem—in particular, because 'causal explanations often appeal to entities such as atoms, molecules or bacteria' (Salmon 1998, p. 359). So he interprets the inapplicability of his account to archaeology as a sign that archaeology does not offer causal explanations. This would hold similarly for the Anasazi simulation (which essentially deals with archaeological data), and each and every one of its possible causal histories. Salmon's account, thus, seems far too narrow for the purpose at hand.

[6] In addition, we have good reasons to believe that in the Anasazi simulation, the modelled behavioural rules are not invariant. Recent research into social norms shows that agents' choices strongly depend on

Of course, other accounts of causal explanations may exists or may be developed in the future that would provide better selection criteria for possible causal histories. But in the current state of agent-based simulations, no attempts are made to justify any selection procedure—neither by the discussed nor by any other criteria. Instead, the possible causal histories that are generated by agent-based simulations are little more than 'Just So Stories' with little or no explanatory potential.[7]

In accordance with this conclusion, some authors see the role of simulation in 'explor[ing] the theoretical structure of the data' (Küppers and Lenhard 2005, p. 9), or in 'computational theorising' (Axtell, quoted in Epstein 1999, p. 46). From that vantage point, of course, agent-based simulations are but sophisticated ways of formulating hypotheses, and are not in the business of explanation or potential explanation. But closer investigation of simulation practice shows that this is not its commonly pursued goal. Pursuing the formulation of hypotheses with the help of simulations would require identifying *all* the models that simulate the target data. In particular, as Axelrod has argued, researchers should seek to replicate one model's simulation results with another model (Axelrod 1997, pp. 33–34). But, as he points out further, this is not at all common practice amongst researchers in the field. Instead, they provide a *single* simulation of a data set, and argue—as shown in Sect. 2—that this one simulation contributes to explanation.

Instead of rejecting this practice as simply misguided, I will now try to develop a (non-causal) account of simulations' explanatory potential. Let's start with another simulation example (from climate research), where the authors deliberately falsify a specific causal relation in their simulation models (cf. Küppers and Lenhard 2005). The relevant model was first built using only six basic equations, which express well-accepted laws of hydrodynamics. It reproduced the patterns of wind and pressure of the entire atmosphere for a simulation period of about four weeks. After that period, the system 'exploded'—the stable flow patterns dissolved into chaos. Consecutive attempts to correct supposed 'errors' of the model—inaccurate deviations of the discrete model from the true solution of the continuous system—remained fruitless. Consequently, the modellers gave up on modelling the causal process. Instead, they focussed on imitating the dynamics alone, trying to find a stable simulation procedure. Assumptions were introduced that partly contradicted experience and physical theory. For example, it was assumed that the kinetic energy in the atmosphere would be preserved. This is definitely not the case in reality, where part of this energy is

Footnote 6 continued
the social context in which they are made. Different social norms will be activated depending on how a situation is understood (Bicchieri 2006, pp. 93–96). Bicchieri's research indicates that some of the variable changes which the simulation performs on are likely to influence the activation of social norm scripts. Take for example the rule of farm plot choice, which specifies that households choose available plots if available, and otherwise migrate. It is, however, plausible that under dense cultivation conditions, households disregard the availability condition and fight over land plots. In these cases, a change in availability will affect the choice rule itself, thus undermining its invariance. Hence Woodward's invariance criterion would be violated.

[7] 'Just So Stories' are fanciful origin stories by Rudyard Kipling, first published in 1902. They are fantastic accounts of how various natural phenomena came about, for example how the elephant got its trunk or the Leopard got its spots.

transformed into heat by friction. Moreover, dissipation is an important factor for the stability of the real atmosphere. In assuming the preservation of kinetic energy, the blow-up of instabilities was 'artificially' limited, for the purpose of reproducing the data over a longer period than in the original model.

Clearly, this simulation does not improve our understanding of the causes that produced the climate, because it incorporates at least one relevant causal relationship that we know is not true. Therefore, it does not provide a potential *causal* explanation. However, I think that one still can attribute explanatory power to this and similar simulations, if one uses a different notion of potential explanation.

The trick is in seeing these simulations not as providing possible causal histories, but possible *functional analyses*. In the climate model, using the relevant causal regularities alone did not yield a successful simulation of the actual climate data. Instead, some well-supported causal regularity had to be falsified in order to achieve generative success. That move damaged the simulation's causal explanatory power. But it did not damage the simulations contribution to a functional analysis of the climate system. The simulation showed that for some reason (e.g. omission of factors, measurement errors, etc.) the included causal regularities did not suffice to dampen the dynamic instabilities of the system. By including an artificial 'instability-dampener', the simulation introduced a functional component into the simulation system that in the real climate system is fulfilled by one or many separate causal factors. The simulation model does not identify these factors (for all we know, the lack of dampening may be the result of slight misspecifications of *all* of the included factors). Instead, it identifies a functional component missing in the existing model, and it specifies the role of this element in the generation of the target data, in the context of the existing model. The simulation therefore cannot be interpreted as providing a possible causal history of the target data. However, it can be interpreted as providing a possible functional analysis of its production process.

This argument can be made clearer with the help of Cummins' account of functional analysis. Functional analysis proceeds by analysing a capacity ψ of a system into a number of other capacities φ of the system or its parts such that their organisation amounts to the manifestation of ψ (Cummins 1975). Cummins' account differs substantially from standard views on functional explanation, which purport to explain the presence of an entity by reference to its effects (Hempel 1965; Little 1991; Kincaid 1996). Cummins claims that functional analysis explains a capacity ψ of a system by reference to the capacities φ of the system's components. The *explanandum* of the analysis is thus the system's ψing. The *explanans* consists of three parts:

i. An analytical account A of the system's ψing
ii. The claim that A involves a component x's φing
iii. The claim that x can φ

To employ the above example again, the climate researchers constructed a computational system that performed ψ. They built this system from a number of components x, y, z, each of which they designed with a specific capacity φ in mind (e.g. 'instability-dampener'). They wrote a program such that the capacities φ_x, φ_y, φ_z, when interacting properly, resulted in the system's ψing. The program then could be used as a possible functional analysis of the real-world climate system. It suggests

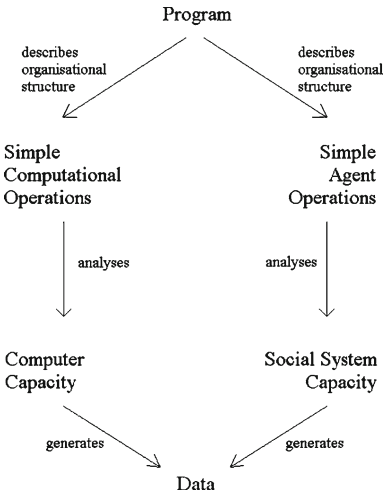

Fig. 2 Computer and target system share the same organisational properties specified by the computer program

an analogy between the organisational structure of the simulator and the real-world system. This analogy claims that a computational process, which imitates a system's behaviour, also shares its organizational properties. Due to their different constitutions (symbols and functions vs. human agents and institutions) the two systems' dispositions will analyse into different simpler operations. But on some level of description, both systems' simpler operations may be governed by the same organizational properties in order to constitute the same dispositions, as depicted in Fig. 2.

It is correct, as Kincaid (1996, p. 106) has pointed out, that Cummins-style functional explanations—if they are *full* explanations—are just a kind of causal explanation. To validate the organisation of the system and the effects its components have would be to validate a causal relation between a component and its effect. But as a *potential* functional explanation—improving our understanding without giving a full explanation—providing a possible functional analysis differs from providing a possible causal history in at least three aspects.

First, functional analysis individuates not according to possible factors or mechanisms, but according to possible functions. In the climate simulation, for example, the dampening of the accumulating instability is performed by a single component. By suggesting the simulation as a possible functional analysis of the real-world climate system, the researchers do not suggest that the stability of the real-world system is produced by a single component, factor or mechanism. Rather, when attributing the function to the system, they admit that there are many ways the real-world system could realise this function.

Potential causal explanations, in contrast, purport to give possible individuations of the relevant causal factors producing the *explanandum*. Causal explanation often requires getting into the details of the causal mechanisms involved that produced the event to be explained. This puts tighter constraints on potential causal explanations than on potential functional ones: given what is known about the causal relationships in the real-world climate system, a single component that preserves kinetic energy in

the atmosphere (and hence dampens dynamic instabilities) can be excluded as a possible causal factor. Thus, while the climate simulation provides a potential functional explanation that contributes to our understanding of the functional organisation of the real-world climate system, it does not provide a potential causal explanation of it.

Second, possible functional analyses are transferable across different causal contexts. To illustrate this point, let me give another example—the Ising model, which is often employed in simulations both of the natural and the social sciences. The Ising model is used both for analysing ferromagnetic systems—with reference to the behaviour of interacting atom magnetic moments—as well as to analyse market dynamics—with reference to socially influenced individual decisions (Brock and Durlauf 2001). Presumably, a ferromagnet and a financial market do not behave according to the same causal mechanisms. However, their possible functional organisation (on some level of description) can be analysed with the same model, and this model may improve our understanding of how each system acquires the capacities it has through the interactions of its subsystems.

Third, the driving power behind potential function explanations is the constitutive relationship between capacities on different levels. Functional analysis shows how lower-level capacities *constitute* higher-level capacities. The capacity of the Anasazi population to disperse in times of draught, for example, is constituted by the capacities of the household agents to optimise under constraints, and their capacity to move. The dispersion *is* nothing but the individual movings. Thus it is wrong to claim that the movings *cause* the dispersion. A functional analysis of the population dynamics is a potential explanation because it identifies these constitutive relationships, not because it identifies any causal relationships. Of course, the simulation always has to make causal assumptions about the influence on the lower-level variables as well; otherwise it cannot generate a dynamic. This is why any full functional explanation, Cummins-style, is a variant of a causal explanation. But *potential* functional explanations propose only constitutional relationships between capacities of different levels, while potential causal explanations propose causal relationships between capacities of the same level.

With the notion of potential functional explanation just developed, I can now clarify the explanatory potential of the Anasazi simulation. The Anasazi modellers constructed a computational system that generated the data set 'population dynamic' from the data set 'meteorological and soil conditions' (the system's ψing). The model on which the simulation is based specifies its subsystems x, y, z (the households, settlement areas and farming plots) and their capacities φ_x, φ_y, φ_z (movement, fertility, housing, crop yields, etc.). It organises these capacities in a specific 'program' (the behavioural rules of the households, the yield functions of the farming plots) so that their combined operation, when fed with the meteorological and soil data, produce the population data. Thus, the program could provide a possible functional analysis of the Anasazi settlement system.

However, as discussed in Sect. 2, the program of the 1999 and 2002 simulations alone did not yield a perfect fit; in particular, they did not replicate the complete eclipse of the settlement in around 1300. The authors therefore concluded that a further functional component had to be introduced into the model:

> The fact that environmental conditions may not have been sufficient to drive out the entire population suggests that *additional push and pull factors* impelled the complete abandonment of the valley after 1300. (Axtell et al. 2002, p. 7278, my emphasis)

The authors argue for 'push and pull factors' from a functional perspective: they do not cite independent causal regularities demanding such factors, but rather argue that the capacities of the system components alone are not sufficient to produce the system capacity.

Because they do not actually provide a simulation that includes such a functional 'pull' component, and that generates results close enough to the observation data, I conclude that the Anasazi simulations do not provide potential functional explanations.

Had the 'pull' factors been included, and had the simulation then been successful, it would have provided a potential functional explanation. But would any form of inclusion have provided equally good functional explanations? If that were the case, one could object that potential functional explanation suffered from the same deficit as potential causal explanations: there would be a large number of possible functional analyses, and the provision of such a large set of possibilities would not significantly increase our understanding of the *explanandum*. Hence providing possible functional explanations would not amount to potential explanations, either.

Fortunately, this conclusion is unwarranted, as we have criteria for the quality of functional analyses. It is useful to go back once more to Cummins, who argues that:

> the explanatory interest of an analytical account is roughly proportional to (i) the extent to which the analyzing capacities are less sophisticated than the analysed capacities, (ii) the extent to which the analysing capacities are different in type from the analyzed capacities, and (iii) the relative sophistication of the program appealed to. (Cummins 1975, p. 764)

The original Anasazi models do quite well on all three counts. The agents' behavioural rules are very simple and few, but they nevertheless create a complex population dynamic. Most of this difference is attributable to the particular way the simulation has them interact. However, simply plucking in a 'pull' component (e.g. assuming that the number of emigrants pulls with them an exponentially related number of other agents) would deteriorate the explanatory quality considerably, as it would be too close in kind to the population dynamic itself. Instead, some simple behavioural rule must be found that accounts for this component. This is where the difficulty of finding a good potential functional explanation lies.

Thus, the quality of its functional analyses can be assessed by the formal properties of the simulation. This gives us a good handle for selecting the best possible functional analyses, which in turn will constitute potential functional explanations.

5 Conclusion

Most full explanations elucidate the causes of the *explanandum*. On the way towards such full explanations, however, scientists use different strategies to build their explanations. Often, the way to full explanations is delayed or even blocked. This is

the case with the Anasazi simulation and similar examples: their models are not and may never be sufficiently validated. Therefore, they may never mature to a full explanation. Despite this, many feel that such simulations contribute to our understanding. They provide potential explanations of some sort, which identify possible *explanantia*. Because of the differences in explanatory strategies, these potential explanations may differ considerably, and may have to be appraised in different ways, too. I argued that the Anasazi simulation and similar models do *not* provide potential causal explanations. Instead, simulations of the Anasazi kind contribute to our understanding because they provide potential functional explanations. These differ from potential causal explanations in at least three ways. Understanding this difference will help to explain how simulations *qua* simulations can contribute to our understanding, even if their underlying models are not validated; and it will help to apply the right appraisal criteria, and hence to weed out good from deficient potential functional explanations derived from agent-based simulations.

References

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In R. Conte, R. Hegselmann & P. Terna (Eds.), *Simulating social phenomena* (pp. 21–40). Berlin: Springer.

Axtell, R. L., Epstein, J. M., Dean, J. S., Gumerman, G. J., Swedlund, A. C., Harburger, J., Chakravarty, S., Hammond, R., Parker, J., & Parker, M. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences, 99*(3), 7275–7279.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York: Cambridge University Press.

Brock, W. A., & Durlauf S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies, 68*(2), 235–260.

Cartwright, N. (2002). Against modularity, the Causal Markov Condition and any link between the two: Comments on Hausman and Woodward. *British Journal for the Philosophy of Science, 53*(3), 411–453.

Cederman, L.-E. (2005). Computational models of social forms: Advancing generative process theory. *American Journal of Sociology, 110*(4), 864–893.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy, 72* (20), 741–765.

Dean, J. S., Gumerman, G. J., Epstein, J. M., Axtell, R. L., Swedland, A. C., Parker, M. T., & McCarrol, S. (1999). Understanding Anasazi culture change through agent-based modeling. In T. A. Kohler & G. J. Gumerman (Eds.), *Dynamics in human and primate societies: Agent based modeling of social and spatial processes* (pp. 179–205). New York and Oxford: Oxford University Press.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity, 4*(5), 41–57.

Epstein, J. M., & Axtell, R. L. (1996). *Growing artificial societies: Social science from the bottom up.* Cambridge, MA: MIT.

Guala, F., & Mittone, L. (2005). Experiments in economics: Testing theories vs. the robustness of phenomena. *Journal of Economic Methodology, 12*, 495–515.

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In R. Hegselmann et al. (Eds.), *Modelling and simulation in the social sciences from the philosophy of science point of view* (pp. 77–100). Dordrecht: Kluwer.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: Free.

Henrich, J., Robert, B., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies.* New York and Oxford: Oxford University Press.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method.* Oxford: Oxford University Press.

Kincaid, H. (1996). *Philosophical foundations of the social sciences*. New York: Cambridge University Press.

Küppers, G., & Lenhard, J. (2005). Validation of simulation: Patterns in the social and natural sciences. *Journal of Artificial Societies and Social Simulation, 8 (4)*. Retrieved February 22, 2006, from http://jasss.soc.surrey.ac.uk/8/4/3.html.

Lewis, D. (1986). Causal explanations. In D. Lewis (Ed.), *Philosophical Papers* II (pp. 214–240). Oxford: Oxford University Press.

Lipton, P. (2001). *Inference to the best explanation*. London and New York: Routledge.

Little, D. (1991). *Varieties of social explanation*. Boulder, CO: Westview.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. (1998) *Causality and explanation*. New York and Oxford: Oxford University Press.

Sawyer, R. K. (2004). Social explanation and computational simulation. *Philosophical Explanations, 7*(3), 219–231.

Tesfatsion, L. (2007). Agent-based omputational economics: A constructive approach to economic heory. In L. Tesfatsion & K. Judd (Eds.), *Handbook of computational economics* (Vol. 2, pp. 1–50). North Holland: Elsevier.

Woodward, J. (2003). *Making things happen*. New York and Oxford: Oxford University Press.