Simulation Models as Opaque Thought Experiments

Ezequiel A. Di Paolo[1], **Jason Noble**[2] and **Seth Bullock**[3]

[1]GMD—German National Research Center for Information Technology (AiS)
[2]Center for Adaptive Behavior and Cognition, MPI für Bildungsforschung, Berlin
[3]Informatics Research Institute, School of Computer Studies, University of Leeds
Ezequiel.Di-Paolo@gmd.de, noble@mpib-berlin.mpg.de, seth@scs.leeds.ac.uk

Abstract

We review and critique a range of perspectives on the scientific role of individual-based evolutionary simulation models as they are used within artificial life. We find that such models have the potential to enrich existing modelling enterprises through their strength in modelling systems of interacting entities. Furthermore, simulation techniques promise to provide theoreticians in various fields with entirely new conceptual, as well as methodological, approaches. However, the precise manner in which simulations can be used as models is not clear. We present two apparently opposed perspectives on this issue: simulation models as "emergent computational thought experiments" and simulation models as realistic simulacra. Through analysing the role that armchair thought experiments play in science, we develop a role for simulation models as *opaque thought experiments*, that is, thought experiments in which the consequences follow from the premises, but in a non-obvious manner which must be revealed through systematic enquiry. Like their better-known transparent cousins, opaque thought experiments, when understood, result in new insights and conceptual reorganisations. These may stress the current theoretical position of the thought experimenter and engender empirical predictions which must be tested in reality. As such, simulation models, like all thought experiments, are tools with which to explore the consequences of a theoretical position.

Introduction

Imagine that you have constructed an artificial life system in which the interactions of many simple agents give rise to complex patterns at the global level. Suppose that these complex patterns remind you of some real-world phenomenon, such as termite nest construction or the behaviour of human investors on the stock market. How do you go about demonstrating the scientific value of your work? There exists a range of answers to this question. At one extreme is the "strong artificial life" position, which suggests that your work is not a *model* of nest construction or investment behaviour, but an *instantiation* of the phenomenon (and hence more than just a simulation). Accepting this view means seeing your piece of work as a new data point to be added to those found in the natural world; scientific investigation proceeds as a search for common features across natural and artificial versions of the phenomenon. An opposing position states that what you have done can have no scientific value, as it is ultimately just a computer program that rearranges symbols in a logical fashion, and as such cannot arrive at new knowledge. This is the idea that if the premises or input are already known, then the conclusions or output cannot constitute an empirical discovery.

We are unhappy with both of these extremes. However, as will be argued below, we are also not content with some of the intermediate positions that have been advanced by artificial life researchers over the past decade. Our goal in this paper is to clearly spell out one way in which the type of systems characteristic of artificial life can make a contribution to science as simulation models — note that we are concerned only with artificial life research conducted in a scientific mode, and will have nothing to say about work directed towards other goals, such as engineering or education.

After reviewing previous attempts to describe the scientific role of artificial life simulations, we will develop our own position through an extended comparison with thought experiments. Our view, in brief, is that although simulations can never substitute for empirical data collection, they are valuable tools for re-organising and probing the internal consistency of a theoretical position. Because simulations are complex, their internal workings are opaque: it is not immediately obvious what is going on or why. This opacity means that researchers must spend time developing and testing a theory of the simulation's operation, before relating this internal theory back to theories about the world, and, ultimately, to the world itself through empirical investigation. Links in this chain are often missing in current artificial life research.

Previous Suggestions: Babies and Bathwater

Several authors have attempted to carve out a niche for the style of simulation modelling pioneered within artificial life (Bonabeau & Theraulaz, 1994; Fontana, Wagner, & Buss, 1994; Ray, 1994; Taylor & Jefferson, 1994; Miller, 1995; Sober, 1996; Bedau, 1999; Maley, 1999). Two basic approaches can be identified. First, the worth of artificial life models is sometimes located in their unique ability to explore some important class of subject matter. Second, artificial life modelling is sometimes claimed to offer new and perhaps superior techniques with which to attack problems that would traditionally be dealt with using existing formal logico-mathematical approaches. These two lines of argument (which are put forward by strongly overlapping groups of authors) will be reviewed in caricature below. Subsequently, two perspectives on the role of evolutionary simulation models in scientific enquiry will be presented. The first takes such models to be "emergent thought experiments" (Bedau, 1999), whilst the second considers their use to form part of a conventional cycle of hypothesis generation and testing dubbed the "physics model" (Kitano, Hamahashi, Kitazawa, Takao, & Imai, 1997).

Unique Object of Enquiry

Artificial life researchers have sometimes claimed that the simulation models which they construct enable them to explore phenomena which lie beyond the ambit of more traditional modelling techniques. The argument runs that since simulations are built from low-level mechanisms (e.g., simulated organisms) which instantiate low-level behaviours (e.g., locomotion), they have the potential to explore the nature of phenomena which although not straightforwardly instantiated by these low-level mechanisms, are nonetheless robust aspects of their high-level, aggregate behaviour (e.g., flocking). This intuition underlies the assertion by Miller (1995), Bonabeau and Theraulaz (1994) and Taylor and Jefferson (1994) that the strength of artificial life simulation models lies in their ability to model natural phenomena which are complex, emergent, and/or self-organising, and that it is through such modelling that artificial life simulations will prove to be most useful since these phenomena are hard to model using previously existing techniques. Indeed some claim that "analytic approaches are certainly doomed" (Bonabeau & Theraulaz, 1994, p. 315).

However, there are potential dangers involved in closely associating the utility of a modelling technique with its application to a specific set of new concepts (e.g., the role of self-organisation in evolution). Such an association may lead to the growing conviction that an important class of phenomena can only be modelled using one particular approach (other approaches being "doomed"), and a reduced tendency to engage with alternative modelling enterprises will be the result. This could seriously impede our ability to construct unifying explanations of the phenomena, since doing so requires that we reconcile the conflicting suggestions of alternative modelling approaches and not merely dismiss them.

Furthermore, the validity of a modelling practice should not stand or fall on the worth of a particular theoretical idea to which it is applied. Recently this issue has been the topic of debate in ecology: to what degree is the methodology of individual-based modelling wedded to a philosophical position regarding the nature of the systems being modelled? Whereas Judson (1994) asserts that a growing awareness of ecosystems as chaotic systems has led directly to the adoption of individual-based simulation modelling practices (which are more able to capture the complexities of such systems), James Bullock (1994) has countered that accepting the utility of these modelling techniques is in no way dependent on accepting this particular perspective on ecosystems. He offers an alternative benefit to this kind of model by suggesting that individual-based modelling can augment more traditional modelling techniques within orthodox theoretical frameworks (see below).

In summary, by claiming that evolutionary simulation modelling is a new technique which should properly be applied exclusively to a new class of problems, modellers incur a risk of scientific isolation and an attendant lack of rigour.

Unique Method of Enquiry

A parallel route has been to claim not that evolutionary simulation models are associated with particular novel phenomena, but that they have properties which make them better than or at least different from existing modelling techniques (e.g., Taylor & Jefferson, 1994; Miller, 1995) in their application to existing phenomena of interest.

We agree with Miller (1995) and Taylor and Jefferson (1994) when they note that mathematical assumptions which are made in the construction of tractable equational models may be relaxed under a simulation-based regime. The infinite, random-mating, unstructured populations often assumed within evolutionary models based upon differential equations may be replaced with finite, structured populations in order to highlight effects of genetic drift, frequency dependent selection, extinction, and other evolutionary phenomena. In addition, the difficulties faced by equational models in capturing non-linearities, or increasingly complex inter-dependencies between the actions of agents, are largely absent from simulation-based models. Further characteristics of natural phenomena which prove difficult to incorporate within equational models include the representation of spatially distributed phenotypes, and repeated interactions between individuals (Nowak & May, 1992; Lindgren & Nordahl, 1994). The difficulty in constructing equational models of many verbal arguments is highlighted by Miller (1995) and Di Paolo (1996). Both suggest that simulation models of such arguments might prove easier to construct.

However, some authors go further, claiming that equational models and individual-based simulation models can be distinguished on the basis of general considerations such as clarity, explicitness, and intersubjectivity. For example, Miller (1995) claims that simulations are more explicit models than those built from differential equations, since, during simulation design, the processes which govern the evolution of the system must be rendered as particular pieces of computer code. Miller also claims that simulations may be passed easily between researchers allowing more effective peer-validation of simulations than of equational models. Similar claims are made by Taylor and Jefferson (1994) who state that the "explicit" representation of behaviour within a simulation compares favourably with the "implicit" representation of an organism's behaviour within equational models. The authors also maintain that simulations are a more direct "encoding" of behaviour, and that this facilitates their design, use, and modification, to such an extent that these processes are necessarily easier to carry out for simulations than equational models.

These unqualified claims are misleading. For example, it is equally admissible to claim of equational models that they capture theoretical assumptions *more* explicitly than simulations since they do not involve extraneous processes which are necessary in order to implement the model as an unfolding, automated process, but which are not spoken to by the theory being tested, and are thus the source of potential artefacts. Similarly, the claim that simulations can be exchanged by modellers in order for their validity to be checked, can be made more forcefully for equational models, which can be presented in their entirety within an academic paper, rather than requiring an additional exchange of computer code. In general, unqualified claims of the superiority of one style of modelling over another are not compelling. Clarity, ease of design, ease of presentation, etc., will vary from model to model to a greater extent than they vary from modelling style to modelling style.

Whilst we agree that simulation models sometimes offer advantages over equational models, the view of individual-based evolutionary simulation models as merely augmenting existing modelling efforts is also overly limiting. It fails to acknowledge that a new tool does not merely increase the number of ways to attack old problems, but also changes the nature of these existing problems, and, in an extreme case, may reveal whole new classes of problem to systematic enquiry. Evolutionary simulation models are not merely a trivial addition to the arsenal of modelling techniques at the disposal of, for instance, the theoretical biologist, but offer a chance to reconsider and explore the theoretical commitments made within existing modelling paradigms and compare them to those made within the new modelling paradigm (Di Paolo, 1996).

In summary, evolutionary simulation models may sometimes offer advantages over traditional methods. However, this is not true as a general rule, and must be assessed on a case-by-case basis. Furthermore, in claiming that, when they are advantageous, evolutionary simulation models are merely a new tool for an old job, there is a risk of undue conservatism and a failure to fully exploit the potential of a novel modelling paradigm.

Emergent Thought Experiments vs. The Physics Model

The positions outlined in the previous two sections address the issue of what challenge artificial life models are best suited to meet — extending old models or modelling new phenomena. We argue that artificial life simulation models have the potential to meet both of these challenges. However, there is a more fundamental question — how can simulations be models at all? In this section we consider two perspectives on *how* simulation models of the kind developed within artificial life can be made to subserve either of the scientific projects identified in the last two sections.

Maynard Smith (1974) maintains that, in the context of ecology, the difference between models and simulations is that whereas models strive for a minimum of detail, simulations strive for a maximum — models are general whereas simulations are specific, gaining validity and scientific worth to the extent that they accurately capture as much about a particular real system as possible. Does the notion of a simulation model imply a departure from this understanding of the role of a simulation? If so does this departure mean that the standards by which we judge the worth of a simulation must also change? Should a simulation model be judged on the same considerations as a model (i.e., generality, parsimony, coherence, etc.) or a simulation (i.e., fidelity, realism, resolution, etc.)?

Two extreme positions on this issue are apparent in the artificial life literature. One position takes the role of simulation models to be essentially in line with that proposed by Maynard Smith for simulations — they are maximally faithful replicas — whereas the other understands simulation models to be more like thought experiments: unrealistic fantasies which nevertheless shed light on our theories of reality.

Kitano et al. (1997) propose that detailed simulations of particular biological systems can serve as the source of novel hypotheses. As such the use of simulation models fits into what they term the "physics model" of scientific enquiry (see figure 1c), in which theories give rise to predictions, which are cast as hypotheses and are tested through experimentation, the results of which have implications for the generation of new theories, giving rise to new predictions, and so on. Under this reading, the attraction of simulation models lies in the claim that within disciplines studying systems comprised of many interacting components, such as those concerning biologists, particular kinds of hypotheses may be unattainable through conventional mathematical analysis.

Proponents of this position claim that in order for a simulation to be a useful source of hypotheses it must be "valid", that is, the behaviour of the simulation must square with data available from real experiments. Only once a simulation has been validated in this manner can one have any interest in any novel insights that might be suggested by it. Kitano et al. liken the process of obtaining these insights to "virtual experiments" designed to provide "decisive evidence" on specific biological questions. Once these virtual experiments have been carried out, the resulting hypothetical answers to these questions can be corroborated, or challenged, by experimentation in the real world. The results of these real experiments will add to biological knowledge and may require accommodating changes in the design of the simulation, new "virtual experiments", and so on.

In contrast, Bedau (1998, 1999) has presented a radically different perspective on artificial life simulation models, claiming that they are best understood as "emergent, computational thought experiments". Far from being authentic reproductions of existing natural systems, simulation models may provide explanations possessing both "simplicity and universality" if their designers "*abstract away* from the micro-details in the real system" in order to construct "models which are as *unrealistic* as possible", but suffice, nevertheless, to provide instances of the phenomena of interest (Bedau, 1999, p. 20).

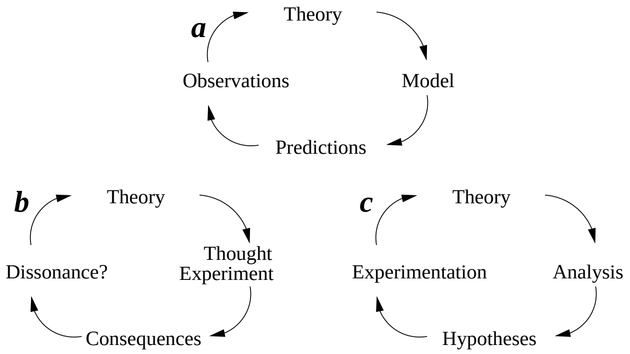

Figure 1: The thought experiment (b) and the physics model (c) can be understood as examples of a more general cycle of scientific enquiry (a).

With the advent of such simulation models, some claim that philosophers have gained a valuable tool which can be brought to bear on contentious thought experiments (e.g., Dennett, 1994; Bedau, 1998, 1999). For instance, those thought experiments involving complex biological systems are often hard to apprehend unaided. For example, if we rewound the geological clock and started evolution from scratch again, would the same history unfold (Gould, 1989)? Would any differences caused by random events be minor ones, or would radically different states of affairs come about as this imaginary terrestrial history diverged from our own? Bedau claims that until a computer version of this thought experiment is constructed and run "all guesses about its outcome — including Gould's [(1989) guess that the evolution of some intelligent lifeform like ourselves would be very unlikely] — will remain inconclusive" (Bedau, 1998, p. 145). What is required in order to settle this matter, Bedau claims, is a powerful computer simulation to aid us in discovering the implications of the premises postulated in the original thought experiment. The prefix *emergent* is appended to the label *thought experiment* in such a situation presumably because it is precisely the emergent properties of the complex systems implicated in some thought experiments which are hard to intuit about. For these situations, in which our natural reasoning apparatus stumbles, Bedau proposes the computer simulation as a philosophical crutch.

How much credence should we give to each of these two visions of individual-based artificial life simulation models, and how might their differences perhaps be reconciled? In the next section, we consider these two proposals and how they contrast with our own.

Simulations as Opaque Thought Experiments

One of the key methodological questions about the scientific use of simulations — what knowledge can be gained from one if all that is 'fed into it' is already known — mirrors exactly the question of what use a thought experiment can be if, unlike a real experiment, it cannot bring to light any new information about a natural phenomenon. It may be of some use to examine briefly how this question has been answered for thought experiments and then see if the same answer works for abstract simulation models as well. In doing so we will uncover one possible use of simulation models and reveal the presence of internal tensions within the roles that Bedau and Kitano et al. attribute to them. However, it should be borne in mind that the analogy between thought experiments as practiced in the armchair and simulation modelling as practiced in the computer laboratory is not complete. There are important differences between the two which demand discussion.

Kuhn (1977, p. 241) poses the following questions about thought experiments. First, to what conditions of verisimilitude must a thought experiment be subject? Second, given that a successful thought experiment involves the use of prior information which is not itself being questioned, how "can a thought experiment lead to new knowledge"? Finally, "what sort of knowledge can be so produced"? Kuhn says that these questions have a set of rather straightforward answers which are important but "not quite right". These answers suggest that all the understanding that can be gained from thought experiments will be understanding about the researcher's *conceptual apparatus*; eliminating previous confusions or revealing inconsistencies within a theory, for instance. If one employs thought experiments to this end, the only requirement of verisimilitude that must be fulfilled is that "the imagined situation must be one to which the scientist can apply his concepts in the way he has normally employed them before" (ibid., p.242)[1].

Kuhn describes a well known thought experiment in which Galileo demonstrates the Aristotelian concept of speed (something similar to the present day idea of average speed) to be paradoxical. An immediate interpretation of this thought experiment can be given along the lines mentioned above. However, by carefully analysing how the Aristotelian concept of speed was used by Aristotle and his followers, Kuhn finds that it "displayed no *intrinsic* confusion" (ibid., p. 261). Contradictions arise when the scientist tries to apply this concept to previously unassimilated experience, as occurred when the 'corrected' Galilean concept of speed was itself confronted with situations where its application proved inappropriate (such as the additivity of the velocities of electromagnetic waves). In this way, Kuhn argues, the thought experiment is indirectly saying something about nature and has a historical role similar to empirical observation. But how is this possible when it was assumed that no new empirical information was 'fed' into the thought experiment? Kuhn answers: "If the two can have such similar roles, that must be because, on occasions, thought experiments give the scientist access to information which is simultaneously at hand and yet somehow inaccessible to him" (ibid., p. 261).

Scientists decide to pay attention to "problems defined by the conceptual and instrumental techniques already at hand" (ibid., p. 262). Therefore, some facts, although known, are pushed to the periphery of scientific investigation, either because they are thought not to be relevant, or because their study would demand unavailable techniques. A thought experiment will, on occasions, bring the relevance of these facts into focus, and therefore catalyse a re-conceptualization which may involve anything from an undramatic reorganisation of relationships between existing concepts to a scientific revolution.

This understanding of thought experiments suggests that they question a theoretical framework in the way depicted in figure 1b. It is important to contrast this view with a commonly held attitude towards artificial life simulation models as synthetic sources of empirical data. As we saw in the previous section, Bedau regards "emergent thought experiments" in a way that seems to imply the latter attitude. According to him, "[it] is worth emphasizing that a model can explain how some phenomenon occurs only if it produces *actual examples of the phenomena in question*, it is not sufficient to produce something that represents the phenomenon but lacks its essential properties", (Bedau, 1999, p. 21, our emphasis). Simulation models that, like mathematical or pen-and-paper variants, merely represent the phenomena of interest cannot serve Bedau's purposes. This is a courageously different understanding of thought experiments, and models in general, and the burden of proof falls on its proponents who must show how a simulation, which always starts from a previously agreed upon theoretical stance, could ever work like a source of new empirical data about natural phenomena.

[1] To this one could add Popper's further requirement that, in the case of an argumentative thought experiment, idealizations should always work in favour of the position that the experimenter is trying to debunk, (Popper, 1959, p. 444).

Bedau is right when he notes that the conclusions of armchair thought experiments may not be justified in the case of complex systems of many interacting elements. But, by suggesting that "emergent thought experiments" can provide the evidence that will settle such matters, he is making a category error which implicitly raises emergent computational thought experiments to the status of empirical, rather than conceptual, enquiries. Gould's argument concerning "replaying evolution's tape" would perhaps be better construed as a speculation about an empirical (albeit hypothetical) state of affairs rather than a thought experiment[2]. Even if scores of "emergent thought experiments" supported this speculation, Gould might still be wrong since there may always exist undiscovered phenomena that would invalidate his reasoning. But, such phenomena can never be *discovered* through building simulations because these are always based on existing theoretical knowledge and, as we have seen from Kuhn's arguments, can only reshuffle existing theoretical ideas, not deliver new facts. Thus Bedau's "emergent computational thought experiments" fall between two stools. If they are sources of empirical data then they are not thought experiments (only real empirical experiments are sources of empirical data). On the other hand, if they are not sources of empirical data then they cannot do the job he requires of them, that is, to provide decisive evidence one way or the other about the validity of our intuitions concerning an armchair thought experiment.

To see how a simulation model *could* function as a thought experiment, consider, as an example, a famous paper by Hinton and Nowlan (1987) in which a clear demonstration of the Baldwin effect is provided using an elegant evolutionary simulation scenario. The Baldwin effect, which stipulates that phenotypic plasticity can speed up an evolutionary process, had existed as a theoretical idea for nine decades at the time the paper appeared. According to Maynard Smith, the Baldwin Effect "has not always been well received by biologists, partly because they have suspected it of being lamarckist [...], and partly because it was not obvious that it would work. What Hinton and Nowlan have done is to answer these objections", (Maynard Smith, 1987, pp. 761–762). In other words, they have discovered nothing new, but have helped in changing an attitude towards an already known piece of information. This change is evidenced by the amount of literature related to the Baldwin Effect that followed, both in theoretical biology and evolutionary computing. Hinton and Nowlan's simulation model thus plays the role of a successful thought experiment, demanding a reorganisation of an existing theoretical framework.

The use of thought experiments sketched above parallels that of abstract models in general. There is, however, an important difference between thought experiments and abstract *simulation* models. A thought experiment has a conclusion that follows logically and clearly, so that the experiment constitutes in itself an *explanation* of its own conclusion and its implications. If this is not the case, then it is a fruitless thought experiment. In contrast, a simulation can be much more powerful and versatile, but at a price. This price is one of *explanatory opacity*: the behaviour of a simulation is not understandable by simple inspection; on the contrary, effort towards explaining the results of a simulation must be expended, since there is no guarantee that what goes on in it is going to be obvious.[3]

This difficulty in achieving an adequate *understanding* of a simulation model threatens to nullify the advantage that simulation models enjoy in terms of the ease with which they may be designed. Whereas many authors have claimed that in important situations the construction of a simulation model tends to be far less of a chore than devising an equivalent formal mathematical model, few have reported the flipside of this advantage — that effort must be made to reconstruct the *relationships between classes* (which mathematical treatments get for free and utilise in explaining the behaviour of analytically derived models) from the *instances* which the simulation model generates. Thus, although, under certain conditions, the construction of simulation models might prove more tractable than the construction of analogous equational models, the analysis of such simulation models often requires an additional effort which threatens to more than compensate for any increased ease of design. This situation has been dubbed the "law of uphill analysis and downhill invention" (Braitenberg, 1984), and has been given a more thorough exposition by Clark (1990), in terms of a general distinction between automatic and manual models.

[2] If the conclusions prompted by an armchair thought experiment are unclear or tendentious — if different thought experimenters reach different conclusions from the same premises — then it is just not a very good thought experiment. Thought experiments are successful to the extent that the conclusions follow trivially from the premises. This unanalysed straightforwardness may of course disguise general reasoning biases which, although shared by all thought experimenters, are in fact fallacious (for example holding that humans are superior to other animals, or that nothing can be smaller than an atom, and so on). These biases drive troublesome *intuition pumps*.

[3] However, although opaque, simulations are explicit and manipulable, and hence may be cognitively penetrable to a greater extent than thought experiments carried out 'in the head' where hidden assumptions may be harder to uncover.

To return to Hinton and Nowlan's simulation model, it is possible to say that, despite its clarity as a simulation, the model did not achieve the logical 'closure' of a good thought experiment. As a matter of fact, the model posed open questions that other researchers have investigated in subsequent work, and which have also turned out to be of theoretical importance (e.g., Harvey, 1993; Mayley, 1996). For instance, Harvey (1993) has investigated a subsidiary aspect of the model: the number of genotypic loci that fixate as non-plastic. The number of non-plastic alleles increases very rapidly during the first stages of the simulation, but then tends to stabilise at some high, but suboptimal value. Harvey shows that genetic drift (random fluctuations in finite populations) is the cause of this phenomenon, thus discovering further richness in the original model, which did not consider this factor, and at the same time demonstrating that unlike clear thought experiments, even simple and elegant simulation models may have a hidden explanatory structure.

Intuitively, we can expect this difference between simulations and thought experiments to become more accentuated as the complexity of the phenomena of interest increases. It is in precisely such cases, according to Kitano et al. (1997), that simulations are deemed most useful. However, lack of transparency in a simulation signifies an important problem for the strategy of validation against empirical data that Kitano et al. propose (see also Maley, 1998). Their "physics model" relies heavily on simulations being implementations of current theory, otherwise the truth or falsity of predictions generated by the model will have no implications for the theory being modelled, but only for the validity of the simulation being built. Specifically, in such a situation, a prediction made by the simulation model which fails to be borne out by experimentation in reality, may be attributed to the simulation's failure to adequately implement the theory, rather than a failure of the theory to adequately account for the relevant natural phenomena. But, despite one's best efforts, empirical validation cannot guarantee that the simulation will implement the theory one intends it to, since different theories could be neutral with respect to the data that one is validating it against. (Particularly if the validation involves some sort of parametric adjustment.) The explanatory opacity of simulations is a disarming problem for their proposed strategy. The lack of *a priori* certainty about what happens in a simulation may be something that we will have to learn to live with, if they are to be applied to the understanding of complex systems involving many interacting parts[4].

[4]Notice that this criticism does not invalidate the "physics model" when other formal approaches are used instead of simulations. In general, in a good mathematical model, everything is (or should be) spelled out, so that if it is not a good implementation of a theory this should eventually become apparent, if not to the researcher, then to critics of the model.

A Workable Methodology

We now turn to the question of how this difference between simulations and thought experiments can be reconciled within a workable methodology if we want them to perform a similar scientific role. There is no general answer to this question. The following paragraphs describe a *possible* way of using simulations as scientific tools in which the difference between them and thought experiments is made methodologically evident. However, this description is no prescription. In particular, not much will be said about how a simulation *should* be built, or when it is adequate for a particular job. Rather, some landmarks in its *use* will be pointed to which ultimately will help in ensuring that the simulation plays a scientific role without undermining its potential[5].

The first preconception that must be challenged is the idea that all that is required from a simulation model is the choice of a plausible mechanism and the replication of a certain pattern in order to claim that an explanation of a similar natural pattern has been achieved. This idea is based on the premise that successful replication implies understanding of how the pattern arises in the simulation (figure 2). It does not always work like that; in fact, it rarely does. Only in a limited number of cases will the researcher be concerned with *just* the basic mechanisms of the model — typically when she wants to present a proof of concept of the type: "it is commonly thought that M is needed to generate P, but here is a model in which M', which is simpler (more plausible, nicer, etc.) than M, reproduces something that looks like P".

Whether explanations are couched in terms of the atomic properties of the simulation or involve higher-level entities as well, it need not be directly obvious how the patterns of interest arise or which aspects of the model are involved and which are inconsequential. Simulations are opaque and must be explored. Relevant observables must be chosen and may lead to the discovery of non-obvious patterns, some of which may not have been suspected initially. That this may happen is a further consequence of the difference between thought experiments and simulations. In the former all relevant entities are already defined whereas in a simulation some of the entities that are interesting, at

[5]There is nothing particularly new about this description. This methodology has been applied successfully in many instances (Boerlijst & Hogeweg, 1991; Fontana et al., 1994, and others).

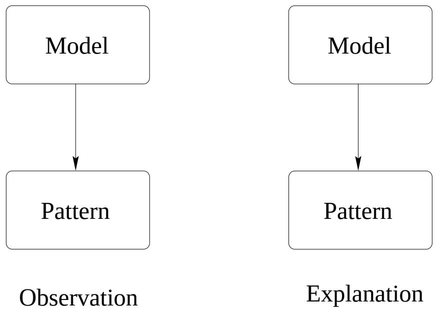

Figure 2: Direct explanation. Observed patterns are explained exclusively in terms of modelled entities and processes. This is to be contrasted with the indirect explanation shown in Figure 3.

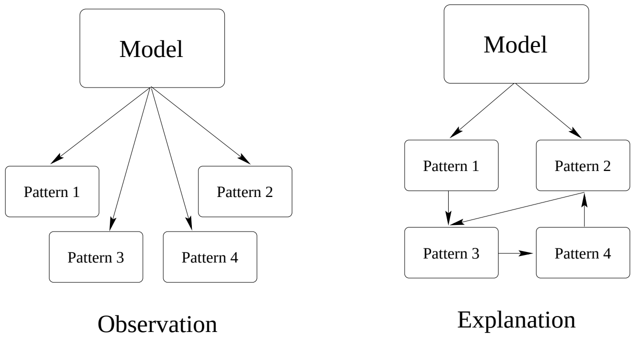

Figure 3: Indirect explanation. Some observed patterns are explained in terms of modelled entities and processes *and* by other observed patterns.

least from a descriptive (but also possibly explanatory) perspective, are not modelled explicitly and are discovered only after the simulation has been observed. An explanation of these entities is often germane to the task that the simulation sets out to achieve.

Notice that simply treating these non-obvious patterns or entities as 'emergent' is not an explanation at all, but rather the statement of a problem. For a simulation model to be of any use, both obvious and non-obvious patterns must be explained and not brushed under the carpet of emergence as this amounts to an admission of failure[6]. This is not an advocation of reductionism, since we are not implying that the explanation must proceed only from the micro- to the macro-structure of observed entities. Some observations can be explained in terms of the basic constituents of the simulation model but others may have to be explained in terms of higher order structures and patterns. Consequently, different observations have to be related through an *explanatory organisation* which, in general, can be more complex than that shown in figure 2 and more like the one depicted in figure 3 where the observed patterns play different explanatory roles, and only some patterns are explained exclusively at the micro-level. Ultimately, this explanatory organisation (which explains what happens *within* the simulation) must then be related to the corresponding theoretical terms which describe analogous phenomena in the natural world.

Consider as an example the model of spatially distributed catalytic reactions presented by Boerlijst and Hogeweg (1991). In this simulation model different chemical species in a two-dimensional lattice spontaneously form dynamic macroscopic patterns in the form of rotating spiral waves. This is one observation. It is also observed that hypercycles (closed loops of catalysing relations between species and reactions) are resistant to the invasion of parasites (chemical species that take advantage of catalysis without themselves catalysing any other reaction in the loop). This is a second observation. This latter observation is important because such resistance is not observed in mathematical models of catalytic reactions taking place in a mixed medium (i.e., those which do not consider the possible effects of spatial structure). Based on this evidence and crucial experiments to test relevant hypotheses, the authors conclude that the first pattern provides an explanation for the second one. They show that the rotational dynamics of spiral waves do not allow parasites to invade. In this simulation model, then, one high level pattern explains another. Now, the authors must relate this explanatory structure to the real case. Since spiral waves have also been observed in the analogous natural systems, it is possible for Boerlijst and Hogeweg to recast the explanatory relevance of this known natural phenomenon using the explanatory organisation developed to account for the behaviour of their simulation model. Whether the resulting explanation holds water will be discovered through actual empirical experimentation on the natural systems con-

[6]Marr (1977) makes a similar point using his distinction between theories of Type-1 (essentially explanations) and Type-2 (essentially descriptions) when he argues that "failure to find [a Type-1 theory] does not mean that it does not exist" (p. 135). Acceptance of a Type-2 solution to a problem in the belief that a more principled understanding of the nature of the problem is made impossible by its emergent, chaotic, non-linear, or context-sensitive nature is at best premature. Although some problems no doubt are of this type (Marr offers protein folding as one possibility), merely assuming that this is the nature of the beast is an unproductive strategy. "Such pieces of research have to be judged very harshly, because their lasting contribution is negligible." (ibid.).

cerned.

In general, we can distinguish three different phases for *using* simulation models in the way we propose:

1. *Exploratory phase:* After the initial simulation model is built, explore different cases of interest, define relevant observables, record patterns, re-define observables or alter model if necessary.

2. *Experimental phase:* Formulate hypotheses that organise observations, undertake crucial "experiments" to test these hypotheses, explain what goes on in the simulation in these terms.

3. *Explanatory phase:* Relate the organisation of observations to the *theories* about natural phenomena and the hypotheses that motivated the construction of the model in the first place, make explicit the theoretical consequences.

The first two phases concern the simulation itself. Here the practitioner is dealing with her own created system. The organisational hypotheses formulated during the second phase prevent random fact-gathering and provide a theoretical perspective proper to the simulation, laying the groundwork for the third phase in which those hypotheses that were developed and supported in the experimental phase can be meaningfully compared with existing theories or hypotheses about natural phenomena.

This final comparison involves a 'backward metaphorical step'. The first, forward use of metaphor occurs when the model is built. Entities in the model represent theoretical entities metaphorically or analogically. However, nothing guarantees that this same set of metaphors will be sufficient when one wants to project the observations made after running the simulation back onto existing theoretical entities relating to the natural world. This may be a trivial step if the observed patterns, or relationships between patterns, have corresponding counterparts in the existing theory. But it is possible to discover relationships between observations that are not easily accommodated by an *existing* theoretical framework. This is a tricky but interesting situation, because this tension may mean that the simulation model is flawed — that it is not modelling what it is supposed to. Alternatively, it may be the current theories that are at fault — the model may be pointing to genuinely new theoretical constructs, which perhaps deserve new names.

The organising theory achieved by the modeller in order to understand what is observed in the simulation may provide a new perspective with which to understand the analogous, existing theory of the natural phenomena being modelled. Conversely, this existing theory may prompt a re-consideration of the organising theory developed during the three phases of simulation modelling, prompting the modeller to explore where the crucial differences lie and make a possibly useful conclusion about them[7].

These possible outcomes of simulation modelling are directly analogous to the results of armchair thought experiments. When a thought experiment generates dissonance (i.e., the consequences of the thought experiment are not easily accommodated by our current understanding of the phenomena involved) we must question both the integrity of our current theories, and the validity of the intuitions which guided our thoughts during the thought experiment. A dialectic process must be invoked in order to reconcile this kind of theoretical impasse.

Conclusion

Thus it is reasonable to understand the use of computer simulations as a kind of thought experimentation: by using the relationships between patterns in the simulation to explore the relationships between theoretical terms corresponding to analogous natural patterns. Through this practice, theoretical terms may be shown to stand in different relationships than previously thought. However, this is an unusual kind of thought experiment. Due to their explanatory opacity, computer simulations must be observed and systematically explored before they are understood, and this understanding can be fed back into existing theoretical frameworks. The necessity of this systematic enquiry into the workings of computer simulations is not part of armchair thought experimentation. The irony here is that, although we advocate an understanding of simulations as tools of *theoretical* enquiry, working with simulations in the way proposed above does have an 'empirical' flavour precisely because complex simulations are not obvious; hence the aptness of the phrase 'computer experiment'.

An additional difference lies in the fact that it may indeed be possible to make a stronger case with simulations than with a 'naked' thought experiment since a simulation can also provide insights that could not be arrived at by thinking alone. As with traditional thought experiments, the information 'fed' into the computer model may not be controversial but, in the end, the researcher may be forced to focus on facts or processes that were at the periphery of her conceptual structure and place them in novel relationships with

[7] See Hesse (1980) for further discussion on the use of metaphors in science and particularly on the two-way conceptual dynamics which is generated when two domains are related metaphorically.

other theoretical terms.

Acknowledgments

Thanks to Walter Fontana, Mike Wheeler and three anonymous reviewers for their helpful comments.

References

Bedau, M. A. (1998). Philosophical content and method of artificial life. In Bynum, T. W., & Moor, J. H. (Eds.), *The Digital Phoenix: How Computers are Changing Philosophy*, pp. 135–152. Basil Blackwell, Oxford.

Bedau, M. A. (1999). Can unrealistic computer models illuminate theoretical biology?. In Wu, A. S. (Ed.), *Proceedings of the 1999 Genetic and Evolutionary Computation Conference Workshop Program*, pp. 20–23. Morgan Kaufmann, San Francisco.

Boerlijst, M. C., & Hogeweg, P. (1991). Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites. *Physica D*, *48*, 17–28.

Bonabeau, E. W., & Theraulaz, G. (1994). Why do we need artificial life?. *Artificial Life*, *1*(3), 303–325.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology.* MIT Press, Cambridge, MA.

Bullock, J. (1994). Letter. *Trends in Ecology and Evolution*, *9*(8), 299.

Clark, A. (1990). Connectionism, competence and explanation. In Boden, M. A. (Ed.), *The Philosophy of Artificial Intelligence*, pp. 281–308. Oxford University Press.

Dennett, D. (1994). Artificial life as philosophy. *Artificial Life*, *1*(3), 291–292.

Di Paolo, E. A. (1996). Some false starts in the construction of a research methodology for artificial life. In Noble, J., & Parsowith, S. R. (Eds.), *The Ninth White House Papers.* Cognitive Science Research Paper 440, School of Cognitive and Computing Sciences, University of Sussex.

Fontana, W., Wagner, G., & Buss, L. W. (1994). Beyond digital naturalism. *Artificial Life*, *1*(1/2), 211–227.

Gould, S. J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History.* W. W. Norton, New York.

Harvey, I. (1993). The puzzle of the persistent question marks: A case study in genetic drift. In Forrest, S. (Ed.), *Genetic Algorithms: Proceedings of the Fifth International Conference on Genetic Algorithms*, pp. 15–22. Morgan Kaufman, San Mateo, CA.

Hesse, M. B. (1980). The explanatory function of metaphor. In *Revolutions and Reconstructions in the Philosophy of Science.* Harvester Press, Brighton, UK.

Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, *1*, 495–502.

Judson, O. P. (1994). The rise of the individual-based model in ecology. *Trends in Ecology and Evolution*, *9*(1), 9–14.

Kitano, H., Hamahashi, S., Kitazawa, J., Takao, K., & Imai, S.-i. (1997). The virtual biology laboratories: A new approach of computational biology. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*, pp. 274–283. MIT Press, Cambridge, MA.

Kuhn, T. (1977). A function for thought experiments. In *The Essential Tension: Selected Studies in Scientific Tradition and Change.* Chicago University Press.

Lindgren, K., & Nordahl, M. G. (1994). Evolutionary dynamics of spatial games. *Physica D*, *75*, 292–309.

Maley, C. C. (1998). Models of evolutionary ecology and the validation problem. In Adami, C., Belew, R. K., Kitano, H., & Taylor, C. E. (Eds.), *Artificial Life VI*, pp. 423–427. MIT Press, Cambridge, MA.

Maley, C. C. (1999). Methodologies in the use of computational models for theoretical biology. In Wu, A. S. (Ed.), *Proceedings of the 1999 Genetic and Evolutionary Computation Conference Workshop Program*, pp. 16–19. Morgan Kaufmann, San Francisco.

Marr, D. (1977). Artificial intelligence — A personal view. In Boden, M. A. (Ed.), *The Philosophy of Artificial Intelligence*, pp. 133–147. Oxford University Press. Collection published in 1990.

Mayley, G. (1996). Landscapes, learning costs and genetic assimilation. *Evolutionary Computation*, *4*(3), 213–234.

Maynard Smith, J. (1974). *Models in Ecology.* Cambridge University Press.

Maynard Smith, J. (1987). When learning guides evolution. *Nature*, *329*, 761–762.

Miller, G. F. (1995). Artificial life as theoretical biology: How to do real science with computer simulation. Cognitive Science Research Paper 378, School of Cognitive and Computing Sciences, University of Sussex.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*, 826–829.

Popper, K. R. (1959). *The Logic of Scientific Discovery.* Hutchinson, London.

Ray, T. S. (1994). An evolutionary approach to synthetic biology: Zen and the art of creating life. *Artificial Life*, *1*(1/2), 179–209.

Sober, E. (1996). Learning from functionalism — Prospects for strong artificial life. In Boden, M. A. (Ed.), *The Philosophy of Artificial Life*, pp. 361–378. Oxford University Press.

Taylor, C., & Jefferson, D. (1994). Artificial life as a tool for biological inquiry. *Artificial Life*, *1*(1/2), 1–13.