Are We Sims? How Computer Simulations Represent and What this Means for the Simulation Argument

Abstract

N. Bostrom's simulation argument and two additional assumptions imply that we likely live in a computer simulation. The argument is based upon the following assumption about the workings of realistic brain simulations: The hardware of a computer on which a brain simulation is run bears a close analogy to the brain itself. To inquire whether this is so, I analyze how computer simulations trace processes in their targets. I describe simulations as fictional, mathematical, pictorial, and material models. Even though the computer hardware does provide a material model of the target, this does not suffice to underwrite the simulation argument because the ways in which parts of the computer hardware interact during simulations do not resemble the ways in which neurons interact in the brain. Further, there are computer simulations of all kinds of systems, and it would be unreasonable to infer that some computers display consciousness just because they simulate brains rather than, say, galaxies.

1. Introduction

There is a fairly high probability that this paper is only a simulated paper, that I am only simulated and that you, honorable reader, are simulated, too. We all likely live in a computer simulation or are sims, for short.[1] This at least follows from the so-called simulation argument and two fairly plausible assumptions. The simulation argument has first been formulated by Bostrom (2003) and since then received much attention.[2]

The main idea behind the argument is as follows. Suppose that mankind continues to exist for at least another 100 years or so. Assume further that the available computing power continues to increase quickly.

"Are We Sims? How Computer Simulations Represent and What this Means for the Simulation Argument" by Claus Beisbart, *The Monist*, vol. 97, no. 3, pp. 399–417.

Copyright © 2014, THE MONIST, Peru, Illinois 61354.

We or our descendants will then enter a "posthuman" age and become "posthumans," as Bostrom puts it. Posthumans have the computing power to run large-scale computer simulations (CSs, for short) that imitate the lives of thousands of humans in great detail. In particular, the simulations imitate the processes in the brains of the people at the levels of neurons or below in a very realistic manner. Bostrom calls such simulations ancestor simulations. If such simulations are run, it seems likely that those parts of the hardware that underlie the simulations of the brains must become conscious and be subject to experiences of the kind we have. This at least is plausible if consciousness is not a matter of the underlying substrate, but rather of its structure. Bostrom assumes that, at some stage, the experiences associated with the computers on which ancestor simulations are run are qualitatively indistinguishable from our own experiences. Suppose now that many ancestor simulations are run. Then most human-like experiences are associated with computer hardware and thus belong to simulated persons. But why then should we think that we are real rather than simulated?

Whether the argument is successful depends, amongst other things, on how successful CSs represent their targets. The argument presumes that the programmed computer on which a brain simulation is run is very similar to the brain. This assumption may be justified by saying that the computer is a detailed material model of the brain and thus represents the brain on the basis of a close similarity. The aim of this paper is to examine whether this is true.

The question of how computer simulations represent their targets and how they help us gain knowledge is central to the philosophy of computer simulations. As a couple of authors (e.g., Stöckler 2000, 366; Barberousse, et al. 2009, 558) have urged, this branch of philosophy should explain how computer simulations produce knowledge. A couple of answers have been put forward, e.g., by Stöckler (2000), Barberousse, et al. (2009) and Beisbart (2012), but most of them do not focus on the hardware, which is crucial for my project. Contrary to Barberousse, et al. (2009), I will argue that a programmed computer can be used as a material model of its target. This does not suffice, however, for the simulation argument, or so I shall suggest.

For the purposes of this paper, I assume that computer simulations provide solutions or approximate solutions to equations that represent the dynamics of a target system. This notion of CS is in rough accordance with the definitions given by Hartmann (1996, Sec. 2.2) and by Humphreys

(2004, 110–14). Some part of my argument is restricted to deterministic simulations, see below for details.

The paper begins with a presentation of Bostrom's argument. Sec. 3 identifies the central assumption about computer simulations in the argument. That assumption turns on the question of how CSs represent. This question is discussed in Sec. 4. The results are applied to the simulation argument in Sec. 5.

2. The Simulation Argument

The simulation argument (Bostrom 2003) is supposed to show that at least one of the following three statements is true.

S1 Human-level civilizations in the universe are likely to become extinct before they enter a posthuman stage.

S2 Posthumans are unlikely to run many ancestor CSs.

S3 It is likely that we live in a CS.

The statements do not only refer to our human civilization and a related posthuman civilization, but also to other possible human-level civilizations and their posthuman successors. The idea is that a cosmic census of human-like experiences is taken.

Suppose for a moment that the disjunction of S1, S2 and S3, i.e., (S1 v S2 v S3) is true. Then, if S1 and S2 are false, S3 must be true. Therefore, if certain contingent assumptions about posthumans, viz. the negations of S1 and S2, hold true, we live in a CS with a high probability.

Does the disjunction hold? The answer turns on the success of the simulation argument. The latter proceeds from an assumption called "substrate-independence":

> Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences. It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks inside a cranium: silicon-based processors inside a computer could in principle do the trick as well. (Bostrom 2003, 244)

We may summarize this by saying that matter with suitable structural properties will be associated with consciousness.[3] More on computation will follow below.

The simulation argument is now supposed to work as follows. Assume that both S1 and S2 are false. Human-level civilizations reach a posthuman stage with a fairly large probability p_1 and posthumans run many ancestor simulations with a fairly large probability p_2. Let the expected number of humans that are simulated by a posthuman society be N_{sim} and the average number of human-level beings in a civilization be N_{real}. Here, N_{sim} is much larger than N_{real} because the expectation value of simulated humans N_{sim} is conditioned on societies in which many ancestor simulations are run. As posthuman simulations of humans are very realistic, the hardware parts that simulate humans are associated with consciousness and human-like experiences according to the assumption of substrate-independence. The expected number of sims under all streams of human-like experiences then is $p_1 \times p_2 \times N_{sim}$, while the expected number of humans is N_{real}. The former number is much larger than the latter because N_{sim} is much larger than N_{real} and $p_1 \times p_2$ is not negligible. Since we expect there to be many more sims than humans, the probability that conscious experience like our own belongs to sims is high, and this just is S3.[4] All in all, S3 seems to follow from a denial of S1 and S2, which means that the disjunction is true. But is it in fact?

3. Computer Simulations

The simulation argument draws on the following crucial assumption. If a computer simulates brains in a very realistic manner, then hardware parts of the computer become associated with human-like experiences. This assumption can be split up into substrate independence and a second assumption as follows:

SI: Matter with suitable structural properties will be associated with human-like experiences.

Sim: If a computer simulation successfully represents the dynamics of a brain in a very realistic manner, the computer hardware will have suitable structural properties.

Here, suitable structural properties are not just properties of computer and brain states; they also concern the dynamical behaviors of both systems. More on this will follow below.

Let us take substrate independence for granted for the sake of argument and focus on Sim. Sim may seem plausible at first sight. How can computer simulations represent the dynamics of a brain in a very realistic way without there being a thorough-going similarity? There are nevertheless doubts about Sim. To discuss Sim, we have to answer two questions:

Q1: What sorts of structural properties does a computer hardware need to successfully simulate brains realistically?

Q2: Are these sorts of structural properties sufficient to grant conscious experiences according to SI?

Turn first to Q1. The answer depends on what exactly we mean by computer simulations. Different types of simulations of the brain are conceivable. It may be possible to build special computer hardware that provides a material model of the human brain. That is, the workings of the hardware closely resemble those of the brain. For instance, microprocessors or electronic devices may be connected as are neurons in the brain. By definition, the hardware of such a computer is very similar to the brain, and the similarity may be so tight that substrate-independence can be brought to bear on the case.

In the following, I will bracket such CSs in favor of simulations run on all-purpose machines. The simplest example of such a computer is a von Neumann machine,[5] but other types of hardware such as those used for parallel computing are available, too. My restriction is legitimate for a number of reasons. A number of brain simulations are in fact run on such machines.[6] Simulations on special computers are more difficult to realize than simulations running on all-purpose machines. Finally, ancestor simulations in the simulation argument are supposed to imitate not only human brains but also their environments, and it is clear that at least these parts of the simulations have to be simulations run on all-purpose machines. Since the simulations of the brains have to be coupled to the simulations of the brains' environments, it seems promising to simulate the brains on all-purpose machines too. Indeed, an extremely ambitious project for brain simulations, the Blue Brain Project[7] uses a Blue Gene/P supercomputer, which is used for other purposes also.

If we restrict ourselves to all-purpose machines in this way, what will be important for the success of the simulation is the software and not so much the hardware. The argument below in Sec. 4.4 will make this plain.

The answer to question Q1 turns also on the question of what is meant by saying that the simulations of the brains are very realistic. I take this to mean that the simulations trace a lot of characteristics, i.e., that they are detailed; that the values of these characteristics are traced with high temporal resolution, and to high accuracy. The latter is to say that the quantitative predictions of the simulations come out true to a very good approximation.

It may be argued that very powerful computers display conscious states because they perform difficult cognitive tasks, e.g., because they perform computations (see Chalmers [2011] for discussion). But this does not seem to be the point of Bostrom's simulation argument because he would then not need to refer to simulations *of the brain* specifically. My question thus is whether the hardware is associated with conscious experiences *because it simulates a human brain*, and not whether it is *because it performs computations*. For simplicity, I will assume that the target of the computer simulation consists in one single brain; this assumption can easily be relaxed.

4. What Does Successful Representation Through Computer Simulations Imply?

Suppose then, that a computer simulation successfully models a brain in a very realistic manner. What does this mean? What are the implications for the relationship between the computer and its target, i.e., the brain? To answer this question, I will characterize computer simulations as models that represent their targets.

To begin it is useful to recap a few things about modeling quite generally.[8] Modelers are interested in a target and take an indirect route to learn about it: They investigate substitutes of their targets. For instance, they use a scale model of a car instead of a real car to learn about the behavior of the latter. The substitute is often called the source (Suárez 2003, 225). Once results about the source have been established, they are transferred back to the target. The sources of representational models vary greatly. They include mathematical equations, merely imagined systems, pictures and movies, and material objects. Depending on the type of source, we obtain *mathematical, fictional, pictorial, and material models*. As it happens, CSs are or contain models of all four types. There may even be additional types of models, but they do not matter for our purposes.

4.1 Computer simulations as mathematical models

Whatever computer simulations do, they perform calculations and thus provide solutions to mathematical equations. The equations and their solutions are empirically interpreted and used to learn about a target. This is to say that each CS includes a *mathematical model.*

Mathematical models represent in that their equations contain information about the target. For instance, under a common interpretation, the equation:

$$\mathbf{F} = -k\mathbf{x} \qquad (1)$$

constrains combinations of the magnitudes of force **F** and of some displacement **x** in some system.[9] Ideally, the constraint expressed by the equation holds exactly in the target system. In this case, it is immediately clear how the equations can be used to learn about the target: From the displacement and the value of k, we can infer how large the force is. But typical mathematical models abstract from many details and are highly idealized. For instance, Eq. (1) may abstract from an additional force due to friction. In this case, one has to be more careful about interpreting the solutions to the equations in terms of the target.

Often, computer simulations do not solve the equations that working scientists first write down to model a system. For instance, CSs do not provide exact solutions to differential equations, but rather use e.g. finite difference methods to approximate the true solutions for a finite set of times or grid points. Nevertheless, the computer does evaluate some equations. It is thus common to distinguish between the conceptual model, i.e., the equations working scientists start with, and the computerized model, i.e., the set of equations that the computer actually evaluates (Schlesinger, et al. 1979).[10] It is still possible to infer information about the target from the computer simulations if the differences between solutions to the conceptual and the computerized models are known to be small.

4.2 Computer simulations as fictional models

Since the equations that are evaluated using CSs do not often provide literally true descriptions of the target, it is useful to say that they refer to a merely imagined system distinct from the target. In fact, CSs are often described in terms of a system of which even the ontology differs from

that of the target. For instance, so-called N-body simulations in physics are naturally said to describe a system of artificial particles and their behavior.[11] But many N-body simulations assume artificial particles that do not exist in reality because there are no particles with roughly the properties of the artificial particles. The artificial particles constitute a merely imagined system. That system is used to represent the target system because it is in some way similar to it. Values of some physical characteristics of the fictional system are used to infer the values of physical characteristics in the target or at least its qualitative behavior. As a consequence, the imagined system that the simulations directly refer to provides a *fictional* model of the target of the simulations.

When I call such a model fictional, as do Frigg and Hartmann (2012), I do not mean to imply that related modeling amounts to storytelling. The point is only that modelers imagine a source distinct from the target, but somehow similar to it. The source then is considered in thought only. Typically, the source is constructed using a lot of idealizations such that it lends itself to a straight-forward description in the terms of a theory. The theory then is used to run inferences about the source.[12]

As with the mathematical equations constitutive of computer simulations, it is useful to distinguish conceptual and computerized fictional models behind computer simulations. The conceptual model is typically constructed using components to which well-known physical laws apply. The computerized model instead follows the dynamics that is defined in terms of the computerized mathematical model, e.g. suitable difference equations.

4.3 Computer simulations as pictorial models

Many simulations visualize their outputs using movies or pictures and thus define *pictorial* models. Pictures and movies allow for insights that would not be possible otherwise (Humphreys 2004, 113). The pictures describe imagined systems that are similar to the target. Due to the analogies, we can learn about the target by looking at the picture. For instance, a picture produced by a simulation contains information about the imagined source, which in turn looks similar to the target.

4.4 Computer simulations as material models

Finally, the computer is a physical object and may thus be thought to form a *material model* of the target of a simulation. Other famous material models include scale models of cars. Scientists observe the source or

experiment on it to learn about the target. Most, if not all material models represent in virtue of being similar to their target. For instance, they resemble their targets in shape or in the way they are composed of parts. Often, the similarity between the source and the target can be described using an isomorphism.[13]

But are computers really material models of the targets of CSs in this sense? On the face of it, this seems implausible. It is true that simulation scientists interfere with the computer to learn about their targets, but a programmed computer does not seem to bear analogies to its target in the way a scale model does. A computer, it may be said, is an instrument that performs certain computations to solve equations from a model. Why should it be similar to the objects to which the computations refer?

Nevertheless, the computer is used to learn about the target system and this is impossible if there is no close relationship between the computer and the target. That the processes in the programmed computer bear a similarity to those in the target is also suggested by Hartmann's claim that a simulation "*imitates one process by another process*" (Hartmann 1996, Sec. 2.2).

In the following, I will show that there is in fact an analogy between the processes in the hardware and in the target. This task is difficult because computers are very complicated material systems. I will thus rely on an idealized description of the computer. In obtaining this description, I will be favorable to the idea that there is an analogy between the programmed computer and the target of a CS.[14]

My argument is restricted to deterministic simulations of a target that is deterministic at the modeled level of description. I assume that, whenever an instance of "simulated" time has lapsed and the values of the relevant characteristics have been evaluated, they are printed. My argument is based upon formal work by Burks (1975) and Norton and Suppe (2001), but I will greatly simplify matters.[15] The formalism has been referred to in the philosophical literature, e.g., in Barberousse et al. (2009), but has not been well explained (Norton and Suppe [2001] refer to yet unpublished work by F. Suppe).

As the CS is run on the computer, the latter sequentially prints outputs. The outputs consist of numbers which are in turn the values of certain functions. These functions solve the dynamical equations from the computerized model. In the terms of Chalmers (2011, 327), this is to say that the computer hardware implements computations. For simplicity, we

assume that all numbers that are values of the functions for the same time t are printed together in one line. Thus one line of the output specifies a *state of the computerized fictional model.*

The printing of one line is proximally caused by a process that takes place within the computer and immediately precedes the printing. At a coarser level of description, we may say that the printing is preceded by one particular computer state that is causally responsible for it.

Consider now the states of the computerized model and the computer states just defined. We have effectively defined a mapping between the states: For each computer state, there is exactly one state of the computerized model. That is the state defined by the values of characteristics printed by the computer immediately after the computer state. Suppose now that the computer goes through a sequence of states, call them c1, c2 etc. We map these states to states in the computerized model in the way just explained. This defines the so-called simulation mapping φ. If the computer does in fact provide solutions to the equations from the computerized model, and we assume so, then $\varphi(\mathrm{c2})$ follows after $\varphi(\mathrm{c1})$ and so on. Computer states that succeed each other are mapped to states of the computerized model that succeed each other too. The simulation mapping thus preserves the temporal order of the states. We have constructed a homomorphism between the dynamics of both systems.

The states of the computerized model in turn are stand-ins of the states of the target system. Thus, states of the computerized models can be mapped to states of the target system. Typically, the values of certain characteristics that define a state of the computerized model are taken to be values of certain characteristics of the target. This leads to a description of the target. In general, the description will be coarse because the computerized model abstracts from many features of the target. So properly speaking, each state of the computerized model corresponds to a class of target states. There is thus a second mapping ψ from states of the computerized model to classes of target states.

If the computerized model successfully traces the dynamics within the target, then ψ again preserves the dynamics. That is, if $\varphi(\mathrm{c2})$ follows $\varphi(\mathrm{c1})$ in the computerized model, then states of class $\psi(\varphi(\mathrm{c2}))$ follow a state from class $\psi(\varphi(\mathrm{c1}))$. If this condition were not fulfilled, then the computerized model could not be used to predict and explain the dynamics of the target. Every prediction or explanation of the dynamics of the target works as follows: We start with an initial state of the target and

obtain a coarse description of it. This description is used to define a state of the computerized model. The computerized model is evaluated for later times. The states of the computerized model are then translated back to coarse state descriptions of the target. They provide information about how the target evolves only if ψ preserves the dynamics.

We can now compose φ and ψ to form a map $\varphi \circ \psi$ that maps states of the computer to classes of target states. Since both maps φ and ψ preserve the dynamics in the way described, so does the composed map $\varphi \circ \psi$. Two computer states that succeed each other are mapped to two classes of target states that succeed each other. Again, this map is crucial for the inferential power of the simulations. Using the computer simulation, we can draw inferences about the target in the following way. A coarse description of the target's initial state is used to define an initial state of the computerized model. ψ maps this state to the class of states to which the initial state of the target belongs. The computer is brought into a state that makes it first print the numbers that correspond to the initial state of the computerized model. As the computer program is run, it goes through a series of computer states c_i. These states correspond to states of the computerized model, and the latter are mapped to classes of target states. Running the computer simulations, we can obtain information about, or coarse descriptions of, the next states to follow in the target.

Thus, states of the programmed computer can be brought in a one-to-one correspondence with classes of target states in such a way that the dynamic order is preserved. This provides a dynamic analogy between the behaviors of the target and of the programmed computer, and this analogy is used to run inferences about the target. In this sense, the programmed computer does form a material model of the target. This result also makes good on Hartmann's claim that computer simulations imitate processes in their targets.

The construction of our maps is based upon coarse descriptions of the computer and the target. At a finer level of description, there are many different computer states that lead to the same printout (Barberousse, et al. 2009, 565). This is not a problem though. For our purposes, we need only a homomorphism at some level of detail.

Barberousse et al. (2009) think that the hardware of the programmed computer would only provide a material model of the target if the computer always stored the values of the variables in the computerized model at the same places. Suitably paired states of the computer and target would

then be similar, and one could map computer states to states of the computerized model by just reading the values off the relevant parts of the memory of the computer. For instance, there would be one place for storing the value of the position of the ith particle in the computerized model, another place for storing that of the $(i + 1)$th particle and so on. According to Barberousse et al. (2009, 564–65), all this does not work, however, because, as a matter of fact, the values of one variable are not always stored at the same place of the computer hardware.

This argument is too quick. First, for the computer to be a material model of the target it is not required that the target states and the computer states resemble each other. The dynamic analogy pointed out above is sufficient to learn about the target using the simulation. Second, there is no reason to require that the map φ be constructed as the authors require, i.e., that the states of the computerized model can be read off in the way they think.

But it is true that the analogy mediated via the mapping $\varphi \circ \psi$ is extremely thin and in a way peculiar. $\varphi \circ \psi$ maps global states to classes of global states and preserves their dynamics, but does not relate the ways in which the parts of the computer interact with each other to those ways in which the parts of the target interact with each other. And we know of course that computers work quite differently than do all sorts of targets of computer simulations.[16]

Why doesn't the mapping $\varphi \circ \psi$ establish an analogy between the inner workings of the target and the computer? The map ψ is unproblematic in this respect because the computerized model qua an imagined system will most often be similar to the target, and the inner workings of the computerized model and the target will be analogous if the simulation is successful. The question rather is whether the simulation map φ that maps computer states into states of the computerized model can in some way be thought to map the inner workings of the computer to those of the computerized model. Let us thus concentrate on φ and regard the computerized model as an adequate stand-in for the target.

φ maps global states to each other, but it induces mappings that relate parts of the target to the computer.[17] Assume that the target is composed of subsystems T1 – Tn. A subgroup of the characteristics that define the computerized model and that figure in its equations refer to T1, others to T2, and so on. Hold those characteristics fixed that belong to T2, . . . , Tn and hold T1 fixed too, apart from one aspect (the value of one natural characteristic) which is varied very slightly. Varying this aspect leads to a

set of global states of the target. Consider those states of the computer that are mapped to the variety of states via φ. If these are global states that differ with respect to a natural subsystem of the computer, the inner workings of the target would be mapped to those of the computer. This is not the case, however. The set of computer states that arise are those that lead to almost the same output, in which only small variations with respect to the value of one printed number are allowed. But if we describe the computer in a natural way, there are a lot of very different states of the programmed computer that can lead to one of the states. A reason is that such an output can arise on the basis of different hardware states. If the values of the same variables are not always stored at the same fixed places, as Barberousse et al. (2009) rightly claim, then the computer states that are mapped to our variety of target states do not only differ in terms of a certain local characteristic. For instance, if we run the same programmed computer from different simulated times to arrive at exactly the same final simulated time and at the same printout, the computer states responsible for the printouts will likely differ from each other. All this is to say that the inner workings of the computer are not mapped to those of the target.[18]

The same point can be made by saying that natural properties of the target do not correspond to natural properties of the computer. A natural property can be regarded as the set of states with that property. Natural properties of the computer and its parts are not mapped to natural properties of the target, and vice versa.

Even if the same physical part of the computer always corresponds to one characteristic of the target, proximate causes of phenomena in the computer and in the computerized model are not mapped to each other. Consider, for instance, a particle i in an N-body simulation. That the particle takes such and such a position at a certain time can proximally be explained in terms of the positions of all particles at the previous time and the forces that the other particles exert on i. The proximal causes for a register state corresponding to the position of the particle at that time are a small number of other register states being such and such. For instance, the register state may result from an addition, thus the previous state of the register and that of another register provide a proximal explanation of the state. When we trace back the causal history of the register state a bit more, we note further that hardware parts that encode parts of the program are crucial for the causal order in the hardware. But the program thus

"materialized" has no counterpart in the target system. This shows that the causal orders of the inner workings in the computer and in the computerized model are very different. This is reflected by the fact that the computer hardware and the target are not in general subject to the same dynamic equations (cf. Barberousse, et al. 2009).[19]

In sum, even though there is a dynamic analogy between the computer and its target, the parts of the computer hardware do not interact in a similar way as do the building blocks of the computerized model (and thus those of the target).

5. Consequences for the Simulation Argument

We are now clear about the representation of a target in a computer simulation. What are the implications for consciousness if a brain is simulated in a very realistic manner? This is to raise Q2. The answer to this question is of course not a matter of armchair philosophy. However, since at least present simulations of the human brain are still quite limited and since it will be difficult to find out whether parts of the hardware become conscious anyway, the only way to answer the question right now is to judge the plausibility of the possible answers.

We need not much consider the first three modes of representation. That the equations of the mathematical model are solved using the computer doesn't tell us anything relevant about the computer (recall that we are asking whether computers resemble a brain because they *simulate brains*). Second, that the computerized model represents the brain in a sufficiently realist manner to display the right sort of structure according to SI implies that the model is structurally equivalent to the brain, but this similarity by itself leads us nowhere in the physical world. Third, movies on the monitor may look very similar to what's going on in slices of the brain. But the information in movies is not detailed enough to make us think that the monitor and the brain are sufficiently analogous for the monitor to display consciousness, granted SI.

However, on top of this the programmed computer serves as a material model of the target in the sense described above. There is a similarity between the dynamics of the computer and the target. The crucial question then is whether the computer as a physical system and the brain are sufficiently similar that the hardware parts are associated with consciousness.

The problem is that the analogy concerns only global states and their dynamic order. The parts of the target and the parts of the computer do not

interact in similar ways to produce the dynamics of the global states, respectively. This means that the programmed computer and the brain do not instantiate the same spatiotemporal pattern.

It seems obvious, then, that our dynamic analogy between the hardware of the computer and the brain is too thin to warrant an application of substrate-independence. A system may be associated with consciousness according to SI if its inner workings are similar to those of brains. But this condition is not fulfilled in our case. There may be other reasons to think that the hardware of a computer is associated with consciousness, but we are here concerned with reasons that turn on a close similarity.

Let me support this conclusion by considering brain simulations on all-purpose machines in more detail. I have above suggested that the map $\varphi \circ \psi$ does not preserve natural properties. This has important implications. Natural properties figure in the formulation of natural laws and we run inductive inferences about such properties. "Grue," by contrast, does not refer to a natural property, and we would not run inferences about the grueness of objects (Goodman 1983, sec. III.4, particularly 74).

This is a problem for the simulation argument. Not every possible brain state will be associated with human-like experiences. Some activities of certain regions of the brain are necessary for conscious experiences. Group together those states that are associated with consciousness and form the set C of all computer states that are mapped to this set. The question now is whether the computer has to be in a state of this set to be associated with consciousness. Those who think that parts of the hardware simulating a brain have consciousness face a dilemma at this point. They can either say that the computer has to be in one of the states from C to be associated with human-like experiences. But since C cannot be defined using simple combinations of natural characteristics of the computer hardware, this is to say that extremely artificial properties have empirical significance and that they make a difference as to whether a system has certain other properties. This is to run inferences along very artificial properties, and this is not plausible, as the example of grueness shows. The other alternative is to deny that the artificial properties on the computer side have any significance for other properties. This is in effect to say that only natural properties of the computer count for consciousness. But from the viewpoint of the natural properties of the computer, it seems incredible that the computer *simulating a brain* should be associated with

consciousness, whereas computers doing all other kinds of things should not. When we look at the natural properties on the side of the computer, it does not make a qualitative difference whether the computer simulates a galaxy or a cell. Thus, if a computer simulating a brain is associated with consciousness, then so should computers simulating other things or computers running other programs. The computer would be subject to conscious states qua doing calculations and not qua running brain simulations. This goes against the spirit of the simulation argument, which considers the computer not qua calculator, but rather as simulating brains.

A possible objection against my argument is that the similarity between computers and brains is much closer in brain simulations because the computers do not only simulate the brain via the computerized model, but also run computations that are similar to computations rooted in the brain. In his statement of substrate independence quoted above, Bostrom indeed refers to two systems that both run computations. Also, building blocks of computers may resemble neurons. So isn't there a much closer analogy between computers and brains?

The problem with this is that analogies do not generally "add up" to form tighter analogies. If there is one homomorphism between the behaviors of two systems with respect to some levels of descriptions, and another with respect to others, it is a nontrivial question whether there are third levels of descriptions that integrate the available information and allow for a third, more detailed homomorphism.

The upshot for the simulation argument is clear enough. The analogy between the hardware and the brain is too thin to warrant the application of substrate-independence. It seems unlikely that higher-level properties of the brain carry over to the computer just because of the map between computer and brain states. And this is to say that the simulation argument fails.[20]

Claus Beisbart

University of Bern

NOTES

1. I take this term from Weatherson (2003).

2. See Bostrom (2005), Bostrom (2009) and Bostrom and Kulczycki (2011) for further explication and elaboration of the argument. See Weatherson (2003) and Brueckner (2008) for criticism and Bostrom (2005) and Bostrom (2009) for replies. For further references

consult www.simulation-argument.com. As far as I am aware, the simulation argument has not yet been discussed from the perspective of the philosophy of computer simulations.

3. Cf. Chalmers (2011, 338–41).

4. There is obviously more to be said about this argument and its use of probabilities. See Bostrom (2003, Sec. V), Weatherson (2003), Bostrom (2005) and Bostrom and Kulczycki (2011) for further clarification and discussion about this point. Related details don't matter for our purposes.

5. Very roughly, von Neumann machines are built of processors (viz. CPUs) each of which executes the commands from a program sequentially using a small number of registers. See Rechenberg (2000, ch. 3) for details.

6. See e.g., Izhikevich and Edelman (2008).

7. See http://bluebrain.epfl.ch/.

8. See e.g., Swoyer (1991), Suárez (2004), and Weisberg (2007). In the following, the term "models" refers to what Frigg and Hartmann (2012, Sec. 1) call representational models of phenomena, and not necessarily to models as structures that fulfill axioms.

9. I will not discuss what it means that characteristics such as force have certain magnitudes because this question is not specific to models and simulations. One option is to be realist about physical characteristics and their values.

10. Winsberg (1999) even distinguishes five types of models in the construction of a simulation, but this is not necessary for our purposes.

11. See e.g., Peebles (1980, Part II) and Dolag et al. (2008) for N-body simulations.

12. See e.g., McMullin (1985), Frigg and Hartmann (2012, Sec. 1.1), Humphreys (2004, Sec. 5.2) for abstraction and idealization.

13. See e.g. van Fraassen (1980, 43–46). My point here is not that representation is equivalent to similarity or isomorphism in these models, an idea that has rightly been criticized (e.g. Suárez 2004). My claim is only that similarity is decisive for the inference between source and target.

14. Note that the analogy can only hold between the *programmed* computer and the target. For simplicity though I will often drop the qualification "programmed."

15. See also Zeigler (1976) and Chalmers (2011).

16. See e.g. Rechenberg (2000, 262–63).

17. Cf. Chalmers (2011, 329) for the following argument.

18. Chalmers (2011, 329) distinguishes between the physical implementations of finite-state and of combinatorial-state automata. Since the latter implementation requires that the states of the physical device are vectorized in components to which components of the automaton states are mapped, it is much more demanding (ibid.). My point here is that the brain and the programmed computer do not implement the same combinatorial-state automaton, at least if we require that the brain and the computer and their states are broken up into natural components.

19. Matters differ in so-called analog simulations. See Trenholme (1994), particularly 118–20.

20. I am very grateful to an anonymous referee and to P. Humphreys.

REFERENCES

Barberousse, Anouk, Sara Franceschelli, and Cyrille Imbert. 2009. "Computer Simulations as Experiments," *Synthese* 169: 557–74.

Beisbart, Claus. 2012. "How Can Computer Simulations Produce New Knowledge?" *European Journal for Philosophy of Science* 2: 395–434.

Bostrom, Nick. 2003. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53: 243–55.

———. 2005. "The Simulation Argument: Reply to Weatherson," *Philosophical Quarterly* 55: 90–97.

———. 2009. "The Simulation Argument: Some Explanations," *Analysis* 69: 458–61.

Bostrom, Nick and Marcin Kulczycki. 2011. "A Patch for the Simulation Argument," *Analysis* 71: 54–61.

Brueckner, Anthony. 2008. "The Simulation Argument Again," *Analysis* 68: 224–26.

Burks, Arthur W. 1975. "Models of Deterministic Systems," *Mathematical Systems Theory* 8: 295–308.

Carrier, Martin, Gerald J. Massey, and Laura Ruetsche, eds. 2000. *Science at the Century's End: Philosophical Questions on the Progress and Limits of Science*, Pittsburgh, PA: University of Pittsburgh Press.

Chalmers, David J. 2011. "A Computational Foundation for the Study of Cognition," *Journal of Cognitive Science* 12: 323–57.

Dolag, Klaus, Stefano Borgani, Sabine Schindler, Antonio Diaferio and Andrei M. Bykov. 2008."Simulation Techniques for Cosmological Simulations," *Space Science Reviews* 134: 229–68.

Edwards, Paul and Clark A. Miller, eds. 2001. *Changing the Atmosphere,* Cambridge, MA, MIT Press.

Frigg, Roman and Stephan Hartmann. 2012. "Models in Science," in Zalta (2012, http://plato.stanford.edu/archives/fall2012/entries/models-science/).

Goodman, Nelson. 1983. *Fact, Fiction, Forecast*, Cambridge, MA: Harvard University Press, fourth edition.

Hartmann, Stephan. 1996. "The World As a Process: Simulations in the Natural and Social Sciences," in Hegselmann, et al. (1996: 77–100), quoted from the revised version at http://philsci-archive.pitt.edu/archive/00002412/.

Hegselmann, Rainer, Ulrich Mueller and Klaus G. Troitzsch, eds. 1996. *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View,* Dordrecht: Kluwer.

Hughes, Richard I.G. 1997. "Models and Representation," *Philosophy of Science (Proceedings)* 64: S325–S336.

Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*, New York: Oxford University Press.

Izhikevich, Eugene M. and Gerald M. Edelman. 2008. "Large-scale Model of Mammalian Thalamocortical Systems," *Proceedings of the National Academy of Science of the U.S.A.* 105: 3593–98.

McMullin, Ernan. 1985. "Galilean Idealization," *Studies in the History and Philosophy of Science* 16: 247–73.

Norton, Stephen D. and Frederick Suppe. 2001. "Why Atmospheric Modeling Is Good Science," in Edwards and Miller (2001, 67–106).

Peebles, Philip J.E. 1980. *The Large Scale Structure of the Universe*, Princeton (NJ): Princeton University Press.

Rechenberg, Peter. 2000. *Was ist Informatik? Eine allgemeinverständliche Einführung*, München: Hanser, third edition.

Schlesinger, Stewart, et al. 1979. "Terminology for Model Credibility," *Simulation* 32: 103–104.

Stöckler, Manfred. 2000. "On Modeling and Simulations As Instruments for the Study of Complex Systems, in Carrier, et al. (2000, 355–373).
Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation," *Philosophy of Science* 71: 767–79.
Swoyer, Chris. 1991. "Structural Representation and Surrogative Reasoning," *Synthese* 81: 449–508.
Trenholme, Russell. 1994. "Analog Simulation," *Philosophy of Science* 61: 115–31.
van Fraassen, Bas. 1980, *The Scientific Image*, Oxford: Clarendon Press.
Weatherson, Brian. 2003. "Are You a Sim?" *Philosophical Quarterly* 53: 425–31.
Weisberg, Michael. 2007. "Who Is a Modeler?" *British Journal for Philosophy of Science* 58: 207–33.
Winsberg, Eric. 1999. "Sanctioning Models: The Epistemology of Simulation," *Science in Context* 12: 275–92.
Zeigler, Bernard P. 1976. *Theory of Modelling and Simulation*, New York: J. Wiley.
Zalta, Edward N. ed. 2012, *The Stanford Encyclopedia of Philosophy,* fall 2012 ed., http://plato.stanford.edu/archives/fall2012/.