UMEÅ UNIVERSITET
Institutionen för filosofi och lingvistik

D-UPPSATS: SLUTGILTIG
Teoretisk filosofi D, Vt. 08

Anders Hammarström
Rullstensgatan 6
906 55 Umeå
Mobil: +46 (0) 705-50 58 69
E-post: anders.hammarstrom@hotmail.com

I, Sim

- an Exploration of the Simulation Argument

Handledare: Sten Lindström
Umeå universitet
Institutionen för filosofi och lingvistik

Abstract

Nick Bostrom argues that we have reason to believe that we are currently living inside a complex computer simulation of the world. This paper explores this *simulation argument*, to see if it's really plausible to assume that we live out our lives in a computer. What does it really mean to be living in a simulation? Will philosophy of mind allow these simulated people to be genuinely conscious, as you and I are? If they are conscious, what are their mental lives like? Do their mental states have causal powers? Could we have any empirical reasons to believe that our world is real rather than simulated? Relative to me, could it ever be true that I am living in a simulation, or is the discussion meaningless? The questions are many, and the answers might seem elusive.

The conclusion reached in this paper is that the argument is, at our current stage of technological development, in principle irrefutable. It all depends on whether or not consciousness can emerge from advanced computer simulations of the human brain, and the answer to this question is, unfortunately, out of our current reach.

Contents

1. Introduction

In the article entitled *Are You Living in a Computer Simulation?* Nick Bostrom argues, somewhat convincingly I might add, that there's a good chance that I, you, everyone you know and all the things that occupy this world are merely a part of a simulation run by a "posthuman" race. Ergo, we live out our entire lives in a world that isn't real in a classical sense. Of course this raises questions about whether or not computer programs can be conscious and whether or not we can tell if we are in a simulation. This paper seeks to explore the argument as set forth by Bostrom. Is it really a sound argument? It does raise some obvious questions that need to be answered before we can accept the possibility that we are simulated.

1.1 What Is This Paper About?

The following questions are to be treated in this paper:

1. Can computers be conscious?

Bostrom's argument presupposes that human-like conscious experiences can supervene on a multitude of physical substrates, more specifically on non-biological substrates. Is it plausible to suppose this? What are the arguments supporting this? Could the argument survive a rejection of this supervenience?

2. How likely am I to be living in a computer simulation?

The argument includes a "bland indifference principle" that has been given some critique. This principle claims that our credence in the hypothesis that we are living in a computer simulation should be equal to the fraction of human-like observers we believe to be living in simulations. Do we have any other information that might influence our credence in this hypothesis? Are our empirical studies of the world independent of the fact that we are or are not living in a simulation. This paper will discuss the arguments for and against this principle.

3. Could it be true that I am in a simulation?

In the modern classic philosophical text *Brains in a Vat* Hilary Putnam argues that if we really are brains in vats in a certain way, being fed with experiences from a computer, we could never say or even think that we are. Does this argument apply to Bostrom's line of reasoning? Could it be the case that if we are in a simulation in a certain way, we can never say or think we are?

Chapter 2 will present my interpretation of the strongest possible simulation argument, based on what is proposed in Bostrom's original paper. This does not mean that I will alter the argument in any way, only that I will be as kind as I can possibly be in my reading of the argument. Chapters 3-5 will discuss the questions that the argument raises. Chapter 6 will summarize and by then we'll see if Bostrom's argument has survived and if we should all concede to the possibility that we might not be nothing more than binary code.

1.1.1 Why Discuss These Matters?

On the face of it, the simulation argument seems to be nothing more than a skeptical metaphysical hypothesis, just like Descartes evil-demon scenario or a *Matrix* or *Truman Show* scenario. But Bostrom argues, in the FAQ section of the simulation argument webpage, that it is nothing like these skeptical challenges. On the contrary, the simulation argument does not start of in doubt about the reality of appearances but accepts most of what we know about the world to be true. And as a consequence, we seem to have empirical reasons (i.e. development of computational abilities and human tendencies to simulate all kinds of stuff) to believe that a certain disjunction, namely the one presented in the simulation argument, is true.

1.2 What Is This Paper Not About?

Rather a large portion of Bostrom's essay deals with the technological limitations of computation. One of the obvious questions raised by the argument is the question of whether or not it is physically possible to build a computer that has the computational power that would be necessary to simulate complete worlds with conscious inhabitants. Since this is not my area of expertise I will simply assume, for the sake of the argument, that this is possible in a manner that Bostrom argues.[1]

The possibility of living in a simulation raises some moral issues. Like for instance, it might change the outset of how we view Robert Nozick's classic thought experiment "The Experience Machine". If we are already living in a simulation, why wouldn't we want to switch to a simulation that is entirely controlled by ourselves? Wouldn't that simulation be as real as the one we were in before? What level of reality is real enough? Etc. Bostrom also argues that the argument might give us reason to act morally. Since we might be living in a simulation, our actions may be punished or rewarded by our simulators. But how would we know what sort of behavior would be punished or rewarded? And how do we know that we

[1] I can't really claim to be an expert on philosophy either, but at least I am a lot better oriented in this area than in the science of technology.

are punished or rewarded at all? These, and similar questions, are difficult and space demanding questions, and they will not be dealt with within the confines of this paper.

The double-aspect theory, the position in the philosophy of mind that claims that the physical and the mental are two aspect of the same neutral substance, will not be dealt with. This is because of the positions speculative nature and due to the fact that there is nothing in Bostrom's paper that indicates that this, or something like it, is his standpoint.

1.3 A Few Technicalities

The simulation argument, which states a certain disjunction, should not be confused with the *simulation hypothesis*, which says that I am now living in a computer simulation. To make this distinction clearer the simulation argument will be referred to by its full name and the simulation hypothesis will be called SIM. SIM, on the other hand, should not be confused with 'Sim'. A Sim is a simulated person living in a simulated world, as in the popular computer game *The Sims* by renowned game designer Will Wright. I.e. if I believe that the simulation argument is sound and give a high credence to SIM, then I will believe that there is a good chance that I might actually be a Sim.

Before we start going over the argument it might be a good idea to define 'computer simulation'. A simulation is an imitating model of whatever it aims to simulate, i.e. a simulated bull can be found in a lot of bars in Texas and a swimming pool at NASA could be a simulation of space. A computer simulation is a computational model that imitates whatever it aims to simulate. To see examples of computer simulations you can probably just look around your own computer. If your computer is like mine you're likely to find e.g. a simulation of a deck of cards that allows you to play a game of solitaire. Computational models can be used to simulate virtually everything, from the effects of CO_2 emissions on the polar ice cap to developments in economics. In this paper the type of simulation of interest is a computer model of the physics of the whole world as we know it.

What would it be like to live in such a simulation? Would it be like living in a fictional story, or a computer game? Well, it's hard to tell, but if the argument is sound and we are currently living in a simulation, then what it's like to live in a simulation is just like, well, our lives. Contemplate what it's like to live in the world you currently occupy, that is exactly what it's like to live in one of Bostrom's simulations.

2. Bostrom's Simulation Argument

Bostrom argues that at least one of the following propositions is true:

(1). "The fraction of all human-level civilizations that reach a posthuman stage is very close to zero;"

(2). "The fraction of posthuman civilizations that are interested in running ancestor-simulation is very close to zero;"

(3). "The fraction of all people with our kind of experiences that are living in a simulation is very close to one." [2]

To get the argument off the ground Bostrom needs two assumptions. The first is the assumption of *substrate-independence*. This position says that mental states supervene on physical states. More importantly, for Bostrom's case, mental states must be realizable in physical states that are non-biological, such as in a computer. However, Bostrom ads, this does not require any strong version of *functionalism* to hold true. He continues: "We need only the weaker assumption that it would suffice for the generation of subjective experiences that the computational processes of the brain are structurally replicated in suitably fine-grained detail[.]"[3]

Bostrom's second assumption is that it is physically possible to build computers that are powerful enough to run simulations of entire worlds. These simulations must be fine-grained enough to generate subjective experiences in each of the simulated beings inside the simulation. He estimates, roughly, that this sort of simulation would need a computer capable of carrying out somewhere between 10^{33} to 10^{36} operations per second and claims that this is well within the reach of "posthuman" computational power, even when leaving a large margin for error.[4]

Given the truth of these two assumptions, we now come to the heart of the argument. If we consider all the human-like civilizations that reach a sufficiently advanced technological level to run ancestor-simulations, this number might be large depending on how many human-like civilizations there are in the universe. But even considering that this number might be low, the number of ancestor simulations run by technologically advanced civilizations will probably

[2] Bostrom, Nick. *Are You Living In a Computer Simulation?* (2003) http://www.simulation-argument.com/simulation.pdf ?, 2007-11-11. p. 11.
[3] Ibid. p. 2.
[4] Ibid. p. 4f.

still be large, since each one of those simulations only takes up a small fraction of the civilizations total computing power. This aids us in drawing the conclusion that the total number of observers with human-like experiences might be extremely large compared to the number of actual human-like observers. All this adds up to form the initial disjunction. Either no human-like civilization survives to reach the sufficiently advanced technological level, or even though they attain this level of technology they choose not to make any significant number of ancestor simulations, or almost all conscious observers are living in one of these simulations.[5]

2.1 The Bland Indifference Principle

Bostrom claims that our credence in the hypothesis that I (i.e. you) am now living in a computer simulation, should be equal to *x*, when *x* is also equal to the fraction of observers with human-like experiences I believe to be living in computer simulations. This principle looks like this:

$$Cr(SIM \mid f_{sim} = x) = x$$

But this only applies if we don't have any information concerning whether or not our own experiences are more or less likely to be that of a simulated or real human, which we don't seem to have according to Bostrom. He calls this his *bland indifference principle*. So, for instance, if I were to believe that 98% of all observers with human-like experiences are living in computer simulations, following Bostrom's principle I would have to say that there is a 98% probability that I am also living in a computer simulation:

$$Cr(SIM \mid f_{sim} = 0.98) = 0.98$$

Of course, this principle is only applicable to the hypothesis given the truth of (3).[6]

Bostrom's conclusion is that it is very likely that the disjunction, (1) $\vee$ (2) $\vee$ (3), is true. If (1) or (2) is true then the possibility that we are in fact living in the real world right now is very large. But if (3) is true, then together with the bland indifference principle we can see that there is a great possibility that I am now currently living in a computer simulation. Please note that Bostrom does not argue specifically that we are living in a simulation. He merely states that the disjunction is true. But what we are interested in is whether it is plausible to assume

[5] Ibid. p. 5f.
[6] Ibid. p. 6.

that such a simulation is at all possible. If it isn't, then what is left in the disjunction would be (1) $\vee$ (2) $\vee$ [$\neg$(2) $\wedge$ $\neg$(3)], which is speculative at best, and quite trivial actually.[7]

2.2 Reconstructing the Simulation Argument

Needless to say (3) is the most interesting of the three disjuncts, this paper will therefore mostly be concerned with the plausibility of this. Let's try to reconstruct the line of reasoning that lead Bostrom to the plausibility of (3):

(a). Our world contains observers with human-like mental states.

(b). Assumption 1: Human-like mental states supervene on sufficiently complex physical states or processes in e.g. human-like brains and computers (or computer simulated human-like brains).

(c). Assumption 2: Physical or technological limitations don't stand in the way of running computer simulations of (several) entire worlds, in suitable detail, for any "posthuman" race.

(d). From (b) and (c) we get: Computer simulated worlds could contain observers with human-like mental states.

(e). From (a) and (d) we get: Our world could be a computer simulated world.

(f). From (c) and (e) we get: There could be many worlds just like ours.

(g). From (f), and the fact that the number of "real" worlds is very low, we get: The number of simulated worlds (and observers) could be a lot larger than the number of non-simulated worlds (and observers).

And

(h). From (g) and the *bland indifference principle*: It could be the case that there is a great chance that I am a simulated observer currently living in a simulated world.

[7] This disjunction states: (1) The human race will likely go extinct before reaching a posthuman stage; or (2) posthumans are unlikely to run large numbers of ancestor simulations; or $\neg$(2) $\wedge$ $\neg$(3) posthumans are likely to run significant numbers of ancestor simulations, but they will not contain any observers with human-like experiences.

What is expressed in (g) makes what (3) says plausible, and by appealing to the bland indifference principle we come to (h), which makes it rational to ascribe a high credence to the simulation hypothesis, or SIM.

A chain is never stronger than its weakest link and therefore we will examine the joints on which the argument ultimately hinges. The obvious weak point is (b), and in chapter 3 we will examine this assumption closer. (c) could also pose a problem, but for the sake of the argument I will accept it as it stands. The step from (g) to (h) is not entirely self-evident, this has been given critique and will be examined in chapter 4. Other than that, the argument looks valid. If (b), (c) and the bland indifference principle will hold, then we will have to grant that it's at least plausible that we are currently living in a simulation.

3. Could Computers Be Conscious?

It is clear that, in order for the simulation argument to work, there has to be a real possibility of computers, or computer programs, being conscious. I am more or less certain that I am conscious, and if computers or computer programs can't be, then I can properly deduce that I'm not living in a computer simulation.[8]

Bostrom assumes that any given mental state can supervene on different kinds of physical states.[9] This assumption is generally called the *multiple realizability thesis*. This thesis says that a mental state, such as pain, can be realized in different kinds of physical states, such as brain states in mammals, electronic states in properly programmed computers, green slime states in aliens, etc.[10] This argument was responsible for ending the reign of *type-physicalism* (the theory that each type of mental states are identical to a type of brain states) as the preeminent position in the philosophy of mind in the late 60s and early 70s. And in its wake emerged *functionalism*, a way of viewing the mind as analogous to a computing machine.[11] So, let's see if functionalism might fit the bill of what Bostrom is looking for, and if it will hold.[12]

3.1 Functionalism

Functionalism is a position in the philosophy of mind that defines mental states as functions or causal roles in cognitive systems. For instance, to be in *pain* is to be in a state that is typically caused by bodily damage, and that causes the belief that "it hurts" and a desire to stop the hurting, and causes one to wince and moan.[13] Thereby, mental states are explained in terms of what they do instead of what they are.

Two of functionalisms trademarks are that it says that mental states and processes are analogous to computational states and processes, and its claim that it's not the biological properties of our brains that constitute our mentality, but its computational powers.[14] "In short, our brain is our mind because it is a computing machine, not because it is composed of

[8] Or at least I can deduce that my mind is not a simulated mind.
[9] Ibid. p. 2.
[10] Bickle, John. "Multiple Realizability", *Stanford Encyclopedia of Philosophy* (2006). http://plato.stanford.edu/entries/multiple-realizability/, 2008-04-02.
[11] Kim, Jaegwon. *Philosophy of Mind* (2006). Westview Press, Cambridge, MA. p. 122ff.
[12] On page 2 in Bostrom's paper he does mention functionalism.
[13] Levin, Janet. "Functionalism", *Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/functionalism/, 2008-04-10.
[14] Kim. p. 137.

the kind of protein-based biological stuff it is composed of."[15] Jaegwon Kim says that the core of the functionalist conception of mind is this:

> "If minds are like computers and mental processes – in particular, cognitive processes – are, at bottom, computational processes, we should expect no prior constraint on just how minds and mental processes are physically implemented. Just as vastly different biological or physical structures should be able to subserve the same psychological processes."[16]

In principle any physical system that has the capacity to meet the above stated requirements for pain truly are capable to be in the mental state pain.[17] Since pain is nothing more than a computational process, its function can be carried out in vastly different physical systems, as long as they're able to compute.

But how can we know that different physical systems are in the same mental state? What does pain have in common in humans, octopuses and computers? Well it can't be a certain brain state, because differences in the brains (and the lack thereof) would rule that out. The common ground is instead that pain has certain inputs and outputs. The typical input is bodily damage and the typical output is pain behavior (wincing and moaning, etc.) However, functionalism should not be confused with *behaviorism*, which gives a similar account of mental states. For the functionalist, mental states are *real* internal states, and as such, they are "over and above" behavior.[18] Behaviorism, on the other hand, claims that mental states are nothing else than behavioral dispositions. On this account, to be in pain is in fact to demonstrate pain behavior, not to be in a state that causes pain behavior.[19]

3.1.1 Functional Reduction

If we introduce the concept of functional reduction, we end up with a functionalistic conception of mind that is on par with *ontological physicalism*, the theory that nothing exists that isn't physical.[20] Functional reduction simply says that to be in a mental state is to be in a physical state that has a certain causal role in a system. In humans we can say that the function of pain is carried out by C-fiber stimulation. If I were to prick my finger on a needle my C-fibers would be stimulated, causing me to say "Ouch!" and to hastily remove my finger

[15] Ibid.
[16] Ibid. p. 118.
[17] Bickle.
[18] Kim. p. 119ff.
[19] Graham, George. "Behaviorism", *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/behaviorism/, 2008-04-10.
[20] Stoljar, Daniel. "Physicalism" *Stanford Encyclopedia of Philosophy* (2001). http://plato.stanford.edu/entries/physicalism/, 2008-04-10.

from the vicinity of the needle. Therefore, pain is identical to C-fiber stimulation. But this only goes for humans. In other species and systems *the physical realizer* of pain might be something completely different.[21] On this view, we can see that functional reduction is consistent with multiple realizability, since a functional role can be played out by different physical states.[22] This conception of mind would be suitable to incorporate into the simulation argument as it permits computers to have mental states.

3.1.2 Can All Mental States Be Functionally Reduced?

Isn't there anything more too mental states than to have a specific functional role? The obvious answer is that mental states have phenomenological qualities, also known as *qualia*. For instance, in addition to being caused by injuries and causing pain behavior, pain is *painful.*[23] The functionalist conception of mental states doesn't seem to give any room for the feeling of *what it is like* to be in pain, to pack a suitcase for a two-week vacation, or to be angry, etc., these qualia and other qualia cannot be explained in functional terms.[24] There are a number of popular arguments that take advantage of functionalisms qualia-trouble, among these are arguments from inverted or absent qualia and the knowledge argument.

The argument from inverted qualia states that it is conceivable that someone could function exactly like me, but experience different qualia under the same circumstances. E.g. when we look at a ripe tomato in normal lighting conditions we both agree that it is red, but he experiences the *quale* that I would call green when he looks at the tomato. This would not yield any difference in behavior, even though he and I, strictly speaking, are not having the same experience.[25] A related objection to functionalism, the problem of absent qualia (also known as the zombie argument), states that it's conceivable that someone who functions exactly like me lacks all conscious experiences. Even though this person would be just as good as I am to identify tomatoes as red, to object to the Platonic conception of "ideas" and to cry at the ending of *E.T.*, it is possible that he doesn't experience any qualia, or that he doesn't have any subjective experiences at all. [26] The conceivability of both inverted and absent qualia shows that functionalism missed the phenomenological qualities of mental states. The

[21] Kim. p. 280ff.
[22] David, Marian. "Kim's Functionalism", *Philosophical Perspectives, 11, Mind, Causation and World* (1997). Blackwell Publishing, Malden, MA. p. 134f.
[23] Kim. p. 162.
[24] Levin.
[25] Kim. p. 162.
[26] Chalmers, David J. *Consciousness and its Place in Nature* (2003). http://consc.net/papers/nature.pdf, 2008-04-02. p. 5.

functionalist might reply that, since mental states are realized by neurobiological states, if I and the supposed zombie are in the same brain state then we necessarily experience the same qualia. But this reply seems to ignore qualia inversions among members of different species. Even though I see red when viewing a tomato, a functionally equivalent martian (or conscious robot) might see green (or nothing).[27]

The knowledge argument says that physical facts do not tell us all there is to know about conscious experiences. The standard version of this argument stems from Frank Jackson.[28] His version goes a little something like this: Imagine Mary, a girl who is confined to a black-and-white room, who reads a lot of black-and-white books and watches programs on a black-and-white television. In this way she learns all there is to know about the physical process of seeing colors. She knows exactly what happens in the brain of someone who is seeing a ripe tomato and could explain it in detail, but she has never seen the color red herself. When she is let out of this room and spots a red rose for the very first time, would she learn anything that she didn't already know? If she does, then we can conclude that the physical facts about seeing is not all there is to it, *what it is like* to see red cannot be described in physical terms.[29] A popular objection to this argument is to say that although she does learn something new, what she learns is first-person concepts, and not anything that isn't reducible to physics, she just learns old facts in new ways. But this seems to presuppose that mental states can be reduced to physics, and that is what the issue at hand is.[30] Another objection claims that what she learns is not propositional knowledge, but practical knowledge, namely how to recognize the color she has seen. But if we put her in front of roses of different shades of red, it's very likely that she won't be able to recognize the rose that has the exact same shade of red as the one she saw earlier if the red hues are close (as red-17 and red-18). So it seems that she really hasn't learned any ability, but this is a topic of debate.[31]

These arguments all rely on an *epistemic gap* between the physical and the mental domain for their success, in particular they claim that mental facts cannot be explained in terms of, or reduced to, physical facts.[32] The arguments make use of the fact that functionalism, and reductive physicalism in general, makes consciousness available to the third-person, even

[27] Kim. p. 163.
[28] Chalmers. p. 6.
[29] Jackson, Frank. "What Mary Didn't Know", *The Journal of Philosophy, Vol. 83, No. 5.* (1986). The Journal of Philosophy Inc, New York, NY. p. 291f.
[30] Levin.
[31] Tye, Michael. "Qualia", *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/qualia/, 2008-04-12.
[32] Chalmers. p. 7.

though one of the trademarks of consciousness is that it is essentially subjective.[33] After establishing this epistemic gap, the arguments go on to claim that this implies an ontological gap as well.[34] On this view there are things in the world that can't be reduced to physics, and if these arguments are correct then the phenomenological qualities of mental states must be over and above the physical domain. In addition to the problem of reducing qualia to functional states, no one has yet been able to give a functionalistic definition of intentional states, such as beliefs and desires, and it is, in the words of Kim, "rather unlikely that a full functional definition will ever be formulated".[35] So, it seems as if intentional states too are incapable of being functionally reduced. However, Kim also says that beliefs and desires still are understood in terms of their function, but that their causal role is open-ended in ways that doesn't allow them to be functionally explained.[36] We will explore the possibilities of a non-reductive physicalism a little later. We aren't quite done with functionalism yet.

3.1.3 Multiple Realizability Revisited

There seems to be a downside to regarding mental states as computational processes of the brain (or of any physical system). Since each computational process is made up of a series of internal states, in order for two physical systems to share a mental state their internal processes would have to be identical.[37] For example, we mentioned earlier that pain is a state that is typically caused by bodily damage, that causes the belief that "it hurts" and a desire to stop the hurting, and causes one to wince and moan. This means that for two physical systems to both be in pain, they both have to have had bodily damage that causes certain beliefs and desires and that causes certain behavior. But this means that physical systems that don't have any internal states that realize the belief "this hurts" are incapable of sharing this state.[38] According to functionalism, for two systems to be in the same psychological state, their whole psychology has to be identical to each other. But to claim that all who share the belief "snow is white" are identical to one and other when it comes to psychological regularities governing their behavior is, to say the least, counterintuitive.[39] In this way, multiple realizablity strikes back furiously at functionalism.[40] One could object and say that systems share mental states if

[33] Searle, John. *The Problem of Consciousness* (1994). http://cogsci.soton.ac.uk/~harnad/Papers/Py104/searle.prob.html, 2008-04-03.
[34] Chalmers. p. 8.
[35] Kim. p. 301.
[36] Ibid. p. 301f.
[37] Ibid. p. 138.
[38] Levin.
[39] Kim. p. 138.
[40] Ibid. p. 141.

they approximately realize the same functional states. But this only begs the question. What is it to approximately realize something? Where does one draw the line between approximate realizations and things that aren't?[41]

3.2 Non-Reductive Physicalism

So, functionalism (and reductionism in general) seems to miss the mark, even though it identifies mental states with computational states. Now we'll see if a *non-reductive physicalism*, also known as *property dualism*, or *emergentism*, might do the trick. This is the theory that claims that mental states are distinct from physical states, that mental states cannot be reduced to physical states, and that mental states *supervene* on sufficiently complex physical states.[42] I will simply refer to this position as *emergentism* from here on. This position doesn't have the problems that functionalism had with characterizing phenomenal states, since it claims they can't be reduced. On this account, phenomenal states are simply phenomenal states, nothing else.[43] Emergentists also hold that mental states are capable of influencing their supervenience base, i.e. physical states cause mental states, and mental states cause physical states. This is the principle of "downward" causation.[44] E.g. C-fiber stimulation causes pain, and pain causes certain beliefs, desires and certain physical behavior.

There is a common objection to emergentism, known as *the exclusion argument*, that states the obvious question: How can mental states (which are non physical) causally influence the physical world? This objection has been around in ages, at least since Elisabeth of Bohemia wrote her infamous letter to Descartes, asking how the non-extended soul could affect the extended body.[45] This is commonly referred to as the problem of downward causation.

3.2.1 The Problem of Downward Causation

The problem of downward causation arises from the combination of four premises that seem to be generally accepted.

(i). Principle of *impact*: Mental states sometimes cause physical states.

(ii). Principle of *antireductionism*: Mental states are not identical with physical states.

[41] Levin.
[42] Kim. p. 290.
[43] Ibid.
[44] Chalmers. p. 29.
[45] Schmidt Galaaen, Öisten. *The Disturbing Matter of Downward Causation* (2006). University of Oslo (Ph.D. Dissertation) https://webspace.utexas.edu/deverj/personal/test/disturbingmatter.pdf, 2008-04-02. p. 3.

(iii). Principle of *overdetermination*: Physical states are, in general, not *overdetermined*.

(iv). Principle of *causal closure*: Any physical event that has a sufficient cause has a sufficient physical cause.[46]

It seems that (i) is reasonable to accept. It merely states that my mental states cause some of my physical states. For instance, my belief that if I write this paper now, I won't have to do it later, and my desire to not write this paper in the summer, causes me to write this paper now.[47] We've already gone through some of the arguments for (ii), so that will not be necessary to do again. (iii) claims that physical events, in general, don't have more than one sufficient cause. For something to have more than one sufficient cause would be for it to be causally overdetermined. Think of a person being hit by two bullets, each of which were sufficient to kill the person. This person's death is causally overdetermined.[48] (iv) says that the physical domain is causally closed, i.e. that everything that happens in the physical world can be explained in terms of physical causes and effects.[49] We can now clearly see that the four premises are not compatible. (iii) and (iv) makes a powerful couple that seems to rule out (i) if we accept (ii), hence it is called the exclusion argument.[50] In light of the previous discussion concerning reductionism we're inclined to accept (ii). So it seems that the exclusion argument draws us towards denying (i) and embracing *epiphenomenalism*, which holds that mental states are causally impotent in the sense that they have no effect on the physical domain.[51] But epiphenomenalism is, without making any understatement, as counterintuitive as it gets. If it is true, then the sensation of pain has nothing to do with the fact that I remove my hand from the needle, and my desire to have a cheeseburger doesn't cause me to say "I'd like a cheeseburger please" to the guy behind the counter at [insert your favorite hamburger restaurant here].[52] Is this really a view we'd like to accept? Jerry Fodor put's it like this:

> "I'm not really convinced that it matters much whether the mental is physical; still less that it matters very much whether we can prove that it is. Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my

[46] Ibid. p. 11.
[47] Ibid.
[48] Kim. p. 196.
[49] Ibid. p. 194f.
[50] Ibid. 197.
[51] Chalmers. p. 32f.
[52] Ibid. 33.

> scratching, and my believing is causally responsible for my saying …, if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world."[53]

I think a lot of people share this sentiment with Fodor, so antireductionists will have to come up with a way of escaping epiphenomenalism. Recent developments in physics might suggest that we can avoid this by denying (iv), the principle of causal closure, instead.

3.2.2 Quantum Physics and Mental Causation

The strongest antireductionist objection to the exclusion argument is to be found in contemporary physics. A common conception in quantum physics is that the development of a quantum system is subject to two rules. Firstly, while unobserved, quantum states develop in accordance to a wave-function that allows several states to be *superposed*, in accordance with an equation developed by physicist Erwin Schrödinger. Secondly, when measured, the wave collapses into one of the superposed states. E.g. a particle can be located at P1, P2 and P3 at the same time in the wave-function, but when measured it assumes a definite position at either P1, P2 or P3.[54] There are disagreements on what constitutes a way of measurement, but one thing that everyone agrees is a type of measurement is *observation by a conscious observer*. In this way the dynamics of the collapse of the wave into a definite state fits the antireductionists like a tailored suit. It consists of one deterministic rule governing physical evolution and another rule governing a nondeterministic evolution that could plausibly be linked to the mental.[55] Renowned physicists Niels Bohr and Eugene Wigner both say that the collapse of a wave-function could be due to the interaction between quantum mechanics and consciousness.[56] However, this interpretation of quantum mechanics is not entirely uncontroversial, David Chalmers writes:

> "Many physicists reject it precisely because it is dualistic […]. There is some irony in the fact that philosophers reject interactionism [or emergentism] largely on physical grounds […] while physicists reject an interactionist [or emergentist] interpretation of quantum mechanics on largely philosophical grounds. Taken conjointly, these reasons carry little force[.]"[57]

There are questions concerning how a theory of this kind is to be formulated in detail, but it seems as if contemporary physics is not capable of ruling out the possibility of downward causation on the quantum level after all. Of course this doesn't entail the truth of

[53] Fodor, Jerry. "Making Mind Matter More", *A Theory of Content and Other Essays* (1990). MIT Press, Cambridge, MA. p. 156. (Qouted in both Kim (p. 181) and Schmidt Galaaen (p. 15).)
[54] Schmidt Galaaen. p. 139.
[55] Chalmers. p. 30f.
[56] Schmidt Galaaen. p. 139f.
[57] Chalmers. p. 31.

emergentism, but if it's this or epiphenomenalism, then this should indeed be a position to be reckoned with.[58]

You may have noticed that the discussion has strayed a bit from the original subject. The emergentist position from quantum mechanics might not seem to have anything to do with the simulation argument, but we have at least reached a position in the philosophy of mind that seems to fit what Bostrom is looking for. It states that mental states supervene on physical states, and combined with the thesis of multiple realizabilility it opens for the possibility of mental states supervening on simulated physical states in computer programs. In addition, if simulated minds could be conscious in this way, quantum physics might save their mental states from epiphenomenalism. Now we will turn our attention to one of the most debated arguments against the notion of conscious computers.

3.3 The Chinese Room

In 1980 John Searle introduced an argument to show that computers aren't capable of understanding. The argument is against "strong artificial intelligence" (strong AI). Strong AI is the view that suitably programmed computers can understand natural languages and are capable of having a mental life similar to that of humans.[59]

Searle tells us to think of a room. Confined within the room there is a man who has no understanding of the Chinese language. However, he has a book full of rules for systematically correlating strings of Chinese symbols with other strings of Chinese symbols. The rules for correlating the strings do not depend on the meaning of the symbols, but simply on how they look. Now, every time a string of Chinese symbols are sent in to the man in the room he consults the rulebook and sends the correlating string out. For someone outside the room who understands the Chinese language, the input, what is sent into the room, is questions in Chinese and the output, sent out by the man in the room, is appropriate responses.

Clearly, the man doesn't understand Chinese. But what if we swap the man for a computer? The computer would systematically answer the questions with appropriate answers in the same way that the man would. What happens inside a computer is essentially the same thing that goes on inside the Chinese room; symbols are handled in accordance to certain rules

[58] Ibid. p. 32.
[59] Cole, David. "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/chinese-room/, 2008-03-26.

based on their syntax. From this we can conclude that the computer does not really understand Chinese.[60]

The point of this argument is that programs in computers base their operations entirely on syntax. Computers respond only to the syntax of the symbols it handles without regard to the meaning of these symbols. Minds, in contrast, have *representational* and *intentional states*, states with *meaning*. A conscious language user associates words with their meaning, and respond to the meaning of the words. But, according to Searle, a computer could never do this because syntax is neither constitutive nor sufficient for *semantics*.[61] Many have criticized this view and claimed that computers may very well have states with meaning. For instance, on an externalist conception of meaning, the computer might have representational content caused by its programming. However, the computer would not be aware of the meanings of its states, and that is the important issue at hand.[62] But what about intentional states, could a computer have those? Searle argues that they cannot. First, he makes a distinction between *original* and *derived intentionality*. Original intentionality is what a mental state has, derived intentionality is what a written or spoken sentence has as it is interpreted by someone. Searle goes on to argue that we may interpret a computational state as having content, but never as having original intentionality. But critics have objected saying this distinction between original and derived intentionality is faulty. Dennett, for instance, says that there is only derived intentionality. He claims all intentionality is derived and all attributions of intentionality are tools of predicting behavior.[63] These issues are complex and far from resolved. But there is another strong objection here, more related to our topic. If physical states in computers aren't representational or intentional, how do neurobiological states in brains come to be about something? For Hilary Putnam this question led to the failure of functionalism.

> "[P]ropositional attitudes, as philosophers call them […] are not 'states' of the human brain and nervous system considered in isolation from the social and nonhuman environment. [….] *Functionalism, construed as the thesis that propositional attitudes are just computational states of the brain, cannot be correct.*"[64]

[60] Searle, John. *Minds, Brains, and Programs* (1980) http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-00/bbs.searle2.html, 2008-04-18.
[61] Kim. p. 146.
[62] Cole.
[63] Ibid.
[64] Putnam, Hilary. *Representation and Reality* (1988). MIT Press, Cambridge, MA. p. 73.

This is a challenge to any materialist conception of mentality.[65] Of course, we have already sidestepped this question with our appeal to the antireductionism of emergentism. But can't mental states emerge from physical states in computers then? There is a common objection to the Chinese room argument that asks precisely this question.

3.3.1 The Systems Reply

The so called *systems reply* to the Chinese room argument says that even though the man in the room doesn't understand Chinese, he is just a part of a bigger system, and the system does understand Chinese.[66] Seen in analogy to a computer, the man in the room is merely the CPU, the larger systems includes the memory and the rules for correlating string of symbols with other strings of symbols. What constitutes understanding of Chinese is not the CPU, but the whole system.[67] Searle's reply to this is simply this: Let the man memorize all the rules so that he doesn't need to look anything up to answer the Chinese questions. We can let him work outside the room. He would be doing exactly what the whole system did earlier, since he has internalized all the parts of the system, now *he is the system*. He still wouldn't understand Chinese.[68] Some have objected that Searle's reply isn't as good as it seems. According to John Haugeland the fact that the man doesn't understand Chinese is irrelevant, he is not the system, he's just the implementer. The larger implemented system could constitute understanding. This reply says that Searle presupposes that the mind that understands Chinese, if any, would have to be the mind of person in the room (or the computer). But if understanding is produced in the Chinese room, then the mind that understands would not be that of the person (or the computer), it would be distinct from him (or it). That the man doesn't understand doesn't prove that there is no understanding taking place.[69] This seems to suggest that "minds are more abstract than the systems that realize them".[70]

Searle says that if we were to accept the systems reply, this would have absurd consequences. For instance we might have to say that there is a chance that my stomach processes information at some level of description when it digests food, and that this might come to constitute understanding of this information.[71] Since computation is observer relative, one can

[65] Kim. p. 148.
[66] Searle (1980).
[67] Cole.
[68] Searle (1980).
[69] Cole.
[70] Ibid.
[71] Searle (1980).

pretty much ascribe computational properties, hence also understanding, to almost anything.[72] But we have already seen that a computational model of consciousness will fail, remember what Putnam said.[73] Ascribing computation is beside the point, because that is not what produces consciousness, complex physical states (like brain states) do. But how do we know which physical states are capable, or incapable, of producing consciousness? Why are certain physical states accompanied by phenomenal experiences? This problem is often referred to as "the hard problem" in philosophy of mind.[74] The systems reply sheds a lot of light on this mysterious relation between body and mind. Searle himself is a materialist, and his argument is, on the first hand, angled against the notion that the computer *is* a mind. But that is not what the systems reply is debating. Its concern is whether or not a sufficiently complex physical state in a computer could give rise to a conscious experience.[75] And given the truth of multiple realizability, it seems plausible that it could. But how would we ever know if it was conscious? Kim gives a good example of this problem. Imagine yourself as being an engineer and someone employs you to design a *pain box* that enables robots that have this box installed to feel pain. Where would you begin? You would probably program the box to cause certain behavior in the robot when it is damaged. But what about the phenomenal experience of pain? How would you go about programming the experience of pain, and how would you know whether or not you've succeeded?[76]

3.3.2 Duplication or Simulation?

There is another important point presented in Searle's original article, namely the distinction between duplication and simulation, or in other words, between real things and simulated things.

> "The idea that computer simulations could be the real thing ought to have seemed suspicious in the first place because the computer isn't confined to simulating mental operations, by any means. No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? [….] To confuse simulation with duplication is the same mistake, whether it is pain, love, cognition, fires, or rainstorms."[77]

[72] Searle (1994).
[73] See quote from Putnam on page 20.
[74] Chalmers. p. 3.
[75] Cole.
[76] Kim. p. 303.
[77] Searle (1980).

This makes a good point. Why would a simulation of a conscious being yield consciousness in a way that simulations of certain weather wouldn't yield wetness? It would seem absurd to suppose that even though the computer simulating rain itself isn't wet the larger implemented system could be. But it could be objected that we don't always know where to draw the line between simulation and duplication. For instance, is an artificial heart a simulation or a duplication of a real heart? Is a prosthetic leg a simulation or duplication of a real leg? If we are inclined to say that these examples are functional duplicates rather than simulations of hearts and legs, why wouldn't we say that artificial intelligence is a replication of a human mind, rather than a simulation?[78]

The simulation/ duplication distinction could also be used to formulate an argument against functionalistic minds in simulations. We saw earlier that functionalism is formulated in terms of inputs and outputs, e.g. in order for anything to be in pain it would have to have certain inputs and certain outputs. But a simulated mind in a simulation could not have the same inputs and outputs as a non-simulated mind, it could only have simulated inputs and simulated outputs. So, the simulated mind would not be in pain, because its inputs and outputs are not the certain ones that are needed to generate a state of pain.

The questions posed by the thought experiment presented by Searle are many, and they shed significant amounts of light on the prominent problems in the field of philosophy of mind. These problems are continuously subject for debate and it will probably be a long time before they are resolved. This discussion has as of now not given us any decisive answers to the question of whether or not computers could be conscious, mainly because there simply are no such answers to be found. If we accept emergentism we can just conclude that it is plausible, but not by any means certain.

[78] Cole.

4. What Do We Know About Ourselves?

Bostrom argues that, given the truth of (3) our credence in *the simulation hypothesis* (SIM), i.e. that you are living in a computer simulation, should be subject to a *principle of indifference*. [79] [80] What then, is a principle of indifference? Well, it is simply a principle that says that each possible outcome is *equally probable*. The canonical example is that of a coin toss. If I were to fling a normal coin up into the air, what is the probability of it landing heads up? If the coin is normal there is no reason to favor 'heads' in front of 'tails'. We have no evidence pointing either way, so we would have to ascribe a credence of 1/2 (0,5) to the *heads-hypothesis*. Further, if our hypothesis is that '6' will come up at the roll of a six-sided dice, then if the dice seems to be normal we would have to ascribe a credence of 1/6 (0,166…) to the *6-hypothesis*. In every case where we have symmetrically balanced evidence, or complete lack thereof, each possibility must be ascribed the same probability.[81] To conform our credence in the heads-hypothesis to Bostrom's notation we could express it like this:

$$Cr(Heads \mid f_{heads} = 0{,}5) = 0{,}5$$ [82]

This should make it pretty clear what Bostrom means by his bland indifference principle:

$$Cr(SIM \mid f_{sim} = x) = x$$

Our credence in SIM should equal the fraction of all human-like observers we believe to be currently living in a simulation, unless we have any evidence that could influence this credence.[83] So, given that it's true that some human-like civilizations will go on to become "posthuman" races who simulate their past, our credence in the simulation hypothesis should be very high. Let's say that the "real" world contains about 7 billion human-like agents, and let's say that posthumans run 10 000 simulations of a world like the one we live in. This would give us a chance of 1 to 10 000 to be living in the real world, and hence our credence in SIM should be 0,9999. But as we noted above, this principle only applies if we do not have any evidence for us being, or not being, in a simulation. So, the question here is: Do we have any such evidence?

[79] (3): The fraction of all observers that are currently living in a simulation is close to one.
[80] Bostrom (2003). p. 6.
[81] Hájek, Alan. "Interpretations of Probability" *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/probability-interpret/, 2008-04-21.
[82] Where f_{heads} is the fraction of all the sides of the coin that has 'heads' on them, i.e. one out of two.
[83] Ibid.

4.1 Evidence Against the Simulation Hypothesis

In Brian Weatherson's response to Bostrom's paper, he interprets Bostrom as supporting the following:

(P1). *A Priori*, our credence in SIM is *x*, given that $f_{sim} = x$

(P2). All our evidence is probabilistically independent of the truth of SIM

(C). Our credence in SIM is *x*, given that $f_{sim} = x$ [84]

The conclusion (C) ultimately rest on the truth of (P2), that our evidence does not affect our credence in SIM. Weatherson argues that we may have three reasons to deny (P2). The first is that our evidence consists of more than phenomenal experiences. On an (phenomenal) *externalist* conception our phenomenal states are influenced by their objects. For instance, there is a difference in perceiving a tree and *sim-perceiving* a *sim-tree*, and these different situations may give rise to different evidence even though the experiences themselves are very alike. On an *internalist* conception our evidence is in part constituted by sensory stimulation, e.g. visual perception involves stimulation of the retina in our eyes. But if we were simulated we would not have eyes, and thus we couldn't experience visual perception. We might have *sim-eyes*, but that is not the same as having eyes. Even though the experience of *sim-seeing* with *sim-eyes* is alike the experience of seeing with eyes, it may in fact result in qualitatively different evidence.[85]

Second, it may be the case that each of our experiences is independent of SIM, but that does not entail that the conjunction of all our experiences is independent. The totality of our evidence may give us reason to reevaluate our credence in SIM.[86]

Lastly we can just flat out deny (P2). Just because our empirical evidence does not entail that we are humans rather than simulated humans (or the other way around) it does not mean that our evidence is probabilistically independent.[87]

Bostrom replies that he never said that all our evidence is irrelevant to the assessment of credence in SIM, he merely stated that we do not seem to have any evidence that is "sufficiently strongly correlated" with SIM that could influence our credence in SIM given

[84] Weatherson, Brian. *Are You a Sim?* (2003) http://www.simulation-argument.com/weatherson.pdf, 2007-12-14. p. 6
[85] Ibid. p. 7.
[86] Ibid. p. 8.
[87] Ibid.

that (3) is true. It could for instance, says Bostrom, be the case that future simulators would insert a "window" into the visual field of the simulated beings that says 'You are living in a computer simulation'. If this were to occur we could be almost certain that SIM is true. In another scenario we may be utterly convinced that everyone who was living in a simulation was wearing black trench coats, then if I am not wearing a black trench coat I can deduce that SIM is false. But since we lack all such evidence we should assign a high credence to SIM.[88]

Bostrom also notes that (P2) is incorrectly formulated and exchanges it for (P2*), saying:

(P2*). All our evidence is probabilistically independent of SIM, after we conditionalize on $f_{\text{sim}} \approx x$

However, Bostrom says that we should take 'probabilistically independent' in a loose sense, it may be the case that we still have evidence that could influence our credence in SIM even after we have established that $f_{\text{sim}} \approx x$. E.g. we could expect future simulators to be especially interested in running simulations of important periods in time (say World War 2 or the aftermath of 9/11) or unusually interesting peoples' lives (like Jesus of Nazareth or Bob Dylan). So if we find ourselves living in a "special" time that could be of certain interest to our descendants we might interpret this as supporting SIM. But we should note that we don't know what our technologically advanced descendants would be interested in simulating, so these types of theories are highly speculative and largely unfounded.[89]

4.2 Can We Trust Perceptual Appearances?

To counter the attack from externalism, that there is a difference in seeing a tree and *sim-seeing* a *sim-tree*, Bostrom simply says that perceptual appearances should not be considered epistemic trumps. We often find ourselves in situations where we ought not to trust our perceptions, like when we see an ore half immersed in water that looks crooked or when we see the wheels of a car spinning backwards when the car is actually moving forward. In these situations it is more reasonable to assume that one is experiencing an illusion than to think that the world is exactly as it is perceived. If (3) is true, then it's reasonable to think that we are in such a predicament. This does not mean that simulated people don't have true beliefs about their world, what they're mistaken about is simply what really constitutes the world. Given the truth of (3) we should not consider our perceiving a tree as a particularly strong

[88] Bostrom, Nick. "The Simulation Argument: Reply to Weatherson" (2005), *Philosophical Quarterly, Vol. 55, No. 218.* Blackwell Publishing, Malden, MA. p. 91.
[89] Ibid. p. 93.

reason to think that we perceive a real tree rather than a simulated tree.[90] It could serve Bostrom's purposes to quote a good point made by Kant in this context:

> "The transcendental concept of appearances […] is a critical reminder that nothing intuited in space is a thing in itself, that space is not a form inhering in things in themselves as their intrinsic property, that objects in themselves are quite unknown to us, and that what we call outer objects are nothing but mere representations of our sensibility, the form of which is space. The true correlate of sensibility, the thing in itself, is not known, and cannot be known, through these representations; and in experience no question is ever asked in regard to it."[91]

The point here is that appearances don't give away the ultimate reality of things, for the underlying structure of the world, whether it is a physical world or a computer simulated world, cannot be known on the basis of perceptual appearances.[92] This, of course, goes for an internalist conception of epistemology as well. Our knowledge that we have eyes gives us no strong reason to conclude that we have real eyes rather than *sim-eyes*.

It could also be appropriate to remind ourselves of another classical point, made by Hume about the *self*:

> "For my part, when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe any thing but the perception. When my perceptions are remov'd for any time, as by sound sleep; so long am I insensible of myself, and may truly be said not to exist."[93]

We can introspect all we want, but when we do we never quite grasp what we call the *self*. And since we can't do this we have no way of knowing the nature of our selves.[94] It may be the case that I am a human mind implemented in a human body, and it may be the case that my mind supervenes on physical states in a computer, but no empirical investigation of myself can reveal my true nature, if I have one that is.

Aranyosi has objected that Bostrom's indifference principle implies that we need another assumption, in addition to that of *substrate independence* (multiple realizability) and future computational powers, to make the simulation argument work. Namely, we have to assume that we do not know for sure that we are living in the year 2008 (Or whatever year you think

[90] Ibid. p. 94f
[91] Kant, Immanuel, *Critique of Pure Reason* (1787). Macmillan, London (1929). B45.
[92] Ibid. B 66.
[93] Hume, David. "Of personal Identity", *A Treatise on Human Nature* (1739-40). http://www.class.uidaho.edu/mickelsen/ToC/hume%20treatise%20ToC.htm, 2008-04-20.
[94] Ibid.

that you may be living in right now) and Aranyosi means that this assumption is "too strong" and that it makes the simulation argument less convincing.[95] But Bostrom notes that our simulators are very much capable of faking historical records and making it seem as if we are living in the 21st century, and they could do this even if our simulation wasn't a computer simulation. We could imagine a *Truman Show*-scenario in which we would be unaware of the actual year.[96] If our simulators could deceive us regarding what year we live in outside of a computer simulation, they would certainly be able to deceive us regarding this if we were inside a computer simulation. In analogy to Bostroms reply to externalist epistemology we can claim that our being under the impression that we live in 2008 does not give us any strong reason for holding that we are living in a real 2008, rather than a simulated 2008.

[95] Aranyosi, István A. *The Doomsday Simulation Argument* (2004). http://www.personal.ceu.hu/students/03/Istvan_Aranyosi/Doomsday%20Simulation/The%20Doomsday%20Simulation%20Argument%20by%20I.A.%20Aranyosi.pdf, 2008-03-14. p. 7.
[96] Bostrom (2005). p. 96.

5. Brains in Vats vs. Minds in Simulations

In Hilary Putnam's modern classic "Brains in a Vat", from *Reason, Truth and History*, an argument is presented to show that a specific skeptical argument is unsound, namely the argument that we might be brains in vats.[97] The skeptical argument can be formulated like this:

[P1]. If I know that P then I know that I am not a brain in a vat.

[P2]. I do not know that I am not a brain in a vat.

[C]. I do not know that P.

Where P is a proposition which contains information about the external world, e.g. 'I am sitting on a chair' or '20 degrees centigrade is a comfortable temperature'. If the argument is sound then I can conclude that I don't really know that I am sitting on a chair or perhaps reading a philosophy paper, because there is a possibility that I am just a brain in a vat.[98]

5.1 Putnam's Brains

Putnam tells us to imagine the following logically possible case: The universe was created with nothing but brains in vats and machines that take care of these brains. All the brains are hooked up to a supercomputer who feed all the brains with correlated stimulations making it seem to them as if everything is just like we, human beings, experience the world. They would think that they have bodies, that the sun feels warm on the skin and that George W. Bush beat Al Gore in the presidential election in 2000. When one of the brains tries to speak it would send signals to the computer who in turn would cause the brain to experience herself as speaking and cause the surrounding brains to hear her speak, etc.[99] From here on, a brain in this situation will be called a 'BIV'. Now, what if you and I were BIVs, could we ever say or think that we were?[100]

If we now modify the skeptical argument to fit our new terminology we will see that I (the first person) cannot know P because I do not know that I am not a BIV. Putnam now argues

[97] Putnam, Hilary. *Reason, Truth and History* (1981). Cambridge University Press, Cambridge, MA. p. 1ff.
[98] Gallois, André N. "Putnam, Brains in Vats and Arguments for Scepticism", *Mind, New Series, Vol. 101, No. 402*. (1992) Oxford University Press, New York, NY. p. 275.
[99] Putnam (1981). p. 5ff.
[100] Ibid. p. 7.

that I can know that 'I am a BIV' is false, by appealing to the truth condition for the sentence 'I am a BIV' i.e. what it takes for the sentence to be true.[101]

5.1.1 Semantic Externalism

Putnam's argument rejects what he calls "magical theories of reference", in other words he denies that words, images and other forms of representation *intrinsically* represent what they are about. For instance, an ant who accidentally traces a line in the sand in a desert somewhere that looks exactly like Winston Churchill has not really drawn a picture of Winston Churchill. This is due to the fact that the ant has no concept of Winston Churchill, and its intention was not to depict anything. In order for anything to represent Winston Churchill there must be a causal connection between the representation and Winston Churchill. The same goes for any representation of anything. In order for it to be a representation there has to be a causal link to the thing that it represents. On this account, as far as we know, the line in the sand drawn by the ant doesn't represent anything at all.[102] This position is generally called *semantic externalism* and more specifically it states that the reference and meaning of the words we use are, at least partially, determined by our environment and not wholly by internal mental states.[103] On this account a word, for instance 'water' refers to something in our environment that we are causally connected to in a certain way, namely whatever it is that usually causes our water-perceptions and water-beliefs, rather than to magically refer to all water everywhere.

This position changes the truth conditions of propositions concerning the external world. For a BIV, a statement 'I am sitting on a chair' is true if the computer hooked up to the BIV makes it perceive its body positioned on top of a chair. This is because a BIVs word 'chair' does not refer to chairs, but to what usually causes its sense impressions of chairs, namely certain features of the computer program that regulates its stimuli. So, statements by BIVs, in contrast to statements by normal humans, should not be considered to have so called *disquotational* truth conditions (i.e. 'P' is true iff P). A statement by a normal human living in the normal world would be true if it were the case, a statement by a BIV is true if it is the case from the perspective of a BIV.[104]

---

[101] Brueckner, Anthony. "Brains in a Vat", *The Journal of Philosophy, Vol. 83, No. 3.* (1986). The Journal of Philosophy Inc, New York, NY. p. 149f.
[102] Putnam (1981). p. 1ff.
[103] Lau, Joe. "Externalism About Mental Content" *Stanford Encyclopedia of Philosophy* (2003) http://plato.stanford.edu/entries/content-externalism/, 2008-04-25.
[104] Brueckner (1986). p. 150f.

5.1.2 Why 'I Am a Brain in a Vat' Is False

Given semantic externalism, we now realize that a BIVs words do not refer to the same things as the words of a normal human being. As Putnam does, we too should make a distinction between the languages of BIVs and normal humans. We'll call the language spoken by BIVs *vat-English*, in contrast to English, spoken by humans. To get us running let's again consider the word 'chair'. In English it refers to chairs, but in vat-English it refers to certain program features that the BIV is experiencing when it thinks it's around chairs. Now, consider the word 'brain'. In English it refers to brains, but in vat-English it doesn't. And respectively 'vat' doesn't refer to vats in vat-English either. So when a BIV utters the phrase 'I am a brain in a vat' it doesn't refer to brains in vats, but to certain program features that represents what BIVs call brains and vats in their world. And, when a human utters the same phrase it does actually refer to brains and vats. Now we can clearly see that every time that either a BIV or a normal human being utters the phrase 'I am a BIV' this statement is necessarily false. A BIV is not program features, so its statement must be false, and likewise, a human is not a BIV, so its statement must be false too.[105] To make this even clearer, let's give this argument a little structure:

[1]. Either I am a BIV (speaking vat-English) or I am a non-BIV (speaking English).

[2]. If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.

[3]. If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV.

[4]. From [2] and [3] we get: If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are false.

[5]. If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.

[6]. From [5] we get: If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are false.

[7]. From [1], [4] and [6] we get: My utterances of 'I am a BIV' are false.[106]

[105] Putnam (1981). p. 14f.
[106] Brueckner (1986). p. 154.

So, there we have it. Whenever I say 'I am a BIV' I am wrong, even if it actually is the case that I am a BIV.

5.1.3 'I Am a BIV' Is False, But Still I Could Be a BIV

The skeptics naturally object. If you read the last sentence of the preceding section carefully you'll see that there's something fishy going on. The fact that my utterance of 'I am a BIV' is false does not rule out the metaphysical possibility of me being a BIV. Even though 'I am a BIV' is false relative to me, it might be true about me relative to some other (real or stipulated) observer that has a more objective perspective on the world. To prove that there is no metaphysical possibility that I am a BIV we must somehow go from [7] to:

[8]. It is not the case that I am a BIV.

But how can this step be sanctioned? We need a few more steps in between (7) and (8) to do this:

[7.1]. From [7] and natural assumptions about truth and negation we get: My utterances of 'I am not a BIV' are true.

[7.2]. From [7.1] and the device of disqoutation we get: My utterances of 'I am not a BIV' are true iff I am not a BIV.

Using [7.1] and [7.2] we get [8].[107] But is this reasoning sound? Compare [7.2] with [2]. Premise [2] says that, if I am a BIV speaking vat-English, the truth conditions of the sentence 'I am a BIV' are not disqoutational, in the manner expressed in [7.2]. Hence, if I am a BIV I cannot use the device of disqoutation to reach the conclusion expressed in [8] in the way that a non-BIV speaking English could. [7.2] does not employ the correct truth conditions for BIVs. For a BIV it would look like this:

[7.2^{BIV}]. From [2] and [7.1] we get: My utterances of 'I am not a BIV' are true iff I do not have sense impressions as of Being a BIV.

It is clear that [8] does not follow from [7.2^{BIV}]. So it seems that Putnam's argument fails in showing that we cannot possibly be BIVs, even though it does in fact show that whenever we utter the sentence 'I am a BIV' it will be false.[108]

---

[107] Brueckner, Anthony. "Brains in a Vat" *Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/brain-vat/, 2007-10-28.
[108] Brueckner (1986). p. 164f.

We have seen that the skeptic can adopt Putnam's truth conditions to show that his reasoning in *brains in vats* doesn't really refute the skeptical argument. But what happens to the skeptical argument itself if we apply the Putnamian truth conditions to it? If I am a BIV, then to me the skeptical argument would have to be formulated like this:

[$P1^{BIV}$]. If I know that I have sense impressions of P, then I know that I do not have sense impressions as of being a BIV.

[$P2^{BIV}$]. I do not know that I do not have sense impressions as of being a BIV.

[C^{BIV}]. I do not know that I have sense impressions of P.

This argument is, although valid, utterly absurd. Neither skeptics nor realists will hold [$P2^{BIV}$] or the conclusion to be true. The lesson learned here is that if the skeptic uses semantic externalism to counter Putnam's argument, she seriously undermines her own position. So a skeptic who is trying to convince Putnam that he could in fact be a BIV, although 'I am a BIV' is false, will not convince anybody that the general skeptical argument is sound.[109]

5.2 Bostrom's Sims

Could we not reconstruct Putnam's argument, but for Sims instead, to show that whenever I utter the phrase 'I am a Sim' it would be false?[110] A skeptical argument like the one challenged by Putnam could in fact be formulated with Sims instead of BIVs. This argument would look like this:

[$P1^{Sim}$]. If I know that P then I know that I am not a Sim.

[$P2^{Sim}$]. I do not know that I am not a Sim.

[C^{Sim}]. I do not know that P.

Could we make this argument just as impotent as the defused skeptical argument concerning BIVs using the same method as Putnam did? It seems that all we really have to do to accomplish this is to accept semantic externalism. According to this position a Sims truth conditions for statements concerning the external world would not be disqoutational. So for a Sim, 'I am sitting on a chair' is not true iff I am sitting on a chair but rather if sense impressions make it seem as if I am sitting on a chair. So, in Sim-English, the word 'chair' does not refer to actual chairs but to computer simulations of chairs. This is so because

[109] Gallois. p. 280f.
[110] With 'Sim' meaning a simulated person in a simulated world.

computer simulated chairs is what typically causes chair-like sense impressions in Sims. For a normal human being, on the other hand, the word 'chair' refers to actual chairs, because that is what typically causes chair-like sense impressions in normal humans. So for a human 'I am sitting on a chair' does have disqoutational truth conditions. This means that for a Sim, the word 'Sim' refers to what typically causes sense-impressions of Sims, not to actual Sims. Now we have made a similar distinction between Sims and humans as we earlier had between BIVs and humans. Following this we could conclude that all my utterances of 'I am a Sim' are false. But this is not to tell the whole story.

5.2.1 English or Sim-English?

Remember, a BIVs existence has very special circumstances. In the universe where it exists, no sentient beings exist that aren't brains in vats. Furthermore, the universe got to be that way completely by accident.[111] A Sim, as described by Bostrom, is in a very different situation. Its existence is not by accident, but caused by some sentient being creating it and the environment it inhabits. If we were Sims, in the way described by Bostrom, then we and our world would be created by our descendants to resemble their ancestors and their ancestors world.[112] So in the case of the Sims there is a causal connection to a higher level of reality that someone in the BIVs predicament would lack. But this connection does not seem to make that big a difference for the truth conditions of the statements of the Sims. The fact that chairs in their world were sculpted to look like chairs in the real world doesn't necessarily mean that the Sims concept 'chair' refers to real chairs rather than to simulated chairs. The typical cause of the Sims chair-like sensations are simulated chairs, not real chairs.

The key question here seems to be: How does a Sim acquire language and concepts? Bostrom doesn't himself answer this question, so we will have to examine the different possibilities. I think it is reasonable to say that the most plausible account of the origin of language among Sims is one of the following two:

[i]. The simulators give the Sims a predetermined language similar to their own.

[ii]. The language of the Sims develops on its own.

It's important to note that [i] does not deny that the language of the Sims evolves once they acquire it. Of course the Sims will refine the use of their words and introduce new concepts to

[111] Putnam (1981). p. 6.
[112] Bostrom (2003) p. 1.

each other as they explore their world. [i] just states that Sims are given a linguistic and semantic *starting point*, i.e. a standard vocabulary with standard meanings associated with the words (perhaps along the lines of the wordlist and the grammar device your word processor uses, but significantly more complex), by whoever is running the simulation. Account [ii] states that language develops within the simulation without any interference from any higher level of reality. This would require that the simulation that the Sims are living in runs at least from the point in evolution when humans started to communicate. This may seem implausible at first, since our descendants will probably be interested in certain periods of time in their history and therefore wouldn't want to simulate the whole world from the first human every time they e.g. want to investigate the assassination of JFK. This problem could easily be sidestepped by imagining a save and load system that our descendants may use to jump to different times in the simulation. They simply save the state of the simulation to a file at a certain point and load that file when they want to return there, much like we do when we save a document on our computer or progress in a video game. Another issue is that the simulators will be interested in understanding what the Sims are talking about. If language develops entirely within the simulation it may develop into a language that the simulators don't understand. This means that they would have to invest effort and time into translating the Sim-languages into their own. Because of this it would seem rational of the simulators to give the Sims a language that they wouldn't have to translate. But this means that the simulators wouldn't just be observing the simulation but also making changes in it. If they want to study it for scientific purposes it would seem reasonable that simulators do not interfere, since that in some sense would be to tamper with the evidence. It's plausible to assume that language in the real world developed on its own, this would mean that if the simulators give the Sims a predetermined language the simulation would not be like the real world in this sense.

If [ii] is correct the Sims can properly be said to be speaking Sim-English, and their words cannot be said to refer to anything that doesn't exist within the boundaries of the world they live in. Hence a Sims word 'Sim' will refer to what usually causes its sense impressions of Sims, i.e. simulated people on a lower level of reality. On this account a Sims utterance of 'I am a Sim' will always be false, and the skeptical argument bites the dust, Putnam-style. If [i] is correct, matters are more complicated. In this case the words will have a predetermined meaning and reference when the Sims are uttering or thinking in terms of them. So, on this account a Sims word 'Sim' would in fact, at least in some sense, refer to actual Sims if the meaning of the word Sim is predetermined by the simulators. This is because on this account

the word ‘Sim’ would have a causal connection to actual Sims. As a consequence, when a Sim utters the statement ‘I am a Sim’ the statement would be true, and this would mean that the skeptical argument doesn’t lose its strength. But this depends on the strength of the semantic externalism one assumes. A stronger thesis might argue that the concepts of a Sim could never refer to anything outside its own world, even if it had a predetermined language.

The plausibility of [i] and [ii] will depend on what kind of simulation one thinks that one is in. If I think there is a great chance of me being in a historical simulation that is being observed for scientific purposes, then [ii] seems more plausible. If I believe that I am in a simulation that has been created for entertainment purposes, like a sophisticated version of *The Sims*, or if I think I am in a staged sociological or psychological experiment, then [i] seems more plausible. But this is all highly speculative. We, the Sims, are in no position to settle what kind of simulation we might be living in.

There is of course a third possible account of language among the Sims, namely that the universe develops deterministically. If this is correct, then a simulation running from the big bang and forwards will always turn out the same way. This implies that even though the language of the Sims develops entirely within the simulation, the language that the Sims acquire will be isomorphic and contain the same words as the language of real humans without the simulators interfering in the simulation. But the fact that the words in the Sim-languages are the same and that they are used in the same way as in real languages does not entail that they refer to the same things. Given the truth of semantic externalism the Sims word ‘Sim’ will not refer to actual Sims but to what typically causes Sim-like sense impressions in Sims. This leads us to the conclusion that, on the third account of Sim-language, a Sims utterances of ‘I am a Sim’ will always be false, and the skeptical argument will be defused.

6. Concluding Discussion

Now then, we've been through the three questions posed at the start of the paper. All that's left to do now is to tie these questions together with Bostrom's simulation argument.

6.1 Consciousness in Computers

We've seen that it's necessary, in order for the argument to work, that a computer simulation of the human mind could yield consciousness of the kind that you and I employ. If not, the possibility that we are currently living in a simulation would be nonexistent. Bostrom assumes a specific kind of *multiple realizability* that allows complex computational processes to produce conscious experiences. This seems to imply some sort of *functionalism*, the view that mental processes and states are, at bottom, computational processes and states. But functionalism, as we've seen, conflicts with the idea that *qualia* are essential to consciousness. Following both the arguments from absent or inverted qualia and the knowledge argument we can reasonably conclude that qualia cannot be functionally reduced. Additionally, such intentional and representational states as beliefs and desires also seem impossible to reduce to physical states or processes. Functionalism also seems to run into with trouble concerning its old buddy multiple realization. This is because in order for two physical systems be in the same mental state, like the one associated with pain, they both have to go through the exact same physical process, which seems counterintuitive. The options available to avoid these problems are either to deny that qualia, mental representation and intentionality exist, or to grant them existence over and above the physical systems that they seemingly belong to. Since the first option is, in most eyes, unacceptable, it's more reasonable to assume that mental states could emerge from complex physical systems; that genuinely distinct mental states *supervene* on the physical.[113]

However, there is a great threat to this non-reductive physicalism, namely *the exclusion argument*. This argument arises from the fact that every physical event that occurs has a sufficient physical cause, i.e. the cause of me raising my arm when I have something to say at a seminar can be described entirely in physical terms without any gaps in the causal chain. The physical domain is, so to speak, causally closed. This notion doesn't leave any room for mental states to be the cause of any physical effect. Since each physical event has a sufficient physical cause, mental states can be excluded from the causal chain. This, the problem of

[113] 'Supervene' is a term used by Bostrom in the short account of his position in the philosophy of mind that he gives in the original paper, so this is very likely to be the sort of position that best fits the bill.

downward causation, draws us towards *epiphenomenalism* if we want to withhold our non-reductive stand in regard to mental states. Epiphenomenalism is, to put it mildly, not a lovable position. If we can consistently avoid it, we should. And it seems as quantum physics provides us with a way of showing that the causal closure of the physical domain might not hold. According to a common conception of quantum physics, microphysics is not causally closed. The quantum system develops according to a wave-function with superposed states, and *something*, that isn't microphysics, causes the wave to collapse into one of the superposed states. This *something* might very well be mental states. Although this reading is far from unanimously accepted, it's consistent, and in order to avoid denying qualia and intentionality, and also escaping the threat of epiphenomenalism, it seems plausible to assume.

We've now gotten to a point where we have a plausible account of the mind that is compatible with multiple realizability, it allows complex physical systems, such as brains, to have genuine mental states that have causal efficacy. But is it compatible with the simulation argument? Does it permit computers to be conscious? Our current position can be used to form a reply to John Searle's notorious Chinese room argument, the reply is known as the *systems reply*. Searle argues that the person inside the Chinese room doesn't understand Chinese, and we would agree with him. However, we could argue that there is understanding in the room, but the understanding is not that of the man. He is just a part of a greater system, and the system could understand. Searle says this leaves us with an unacceptable consequence, namely that consciousness is everywhere, e.g. in his stomach. I don't think that our version of the systems reply gives us any reason to assert such a statement. The question is whether or not Searle's stomach has the sufficient level of complexity that is required for consciousness to emerge. I'd say it probably hasn't. However, a system consisting of a book of rules, or even of memorized rules, governing every possible question in Chinese, a human brain, and inputs/ outputs might be sufficiently complex to do this. Similarly, a physical state in a computer, that has immense computational power, could be complex enough to yield consciousness. We are currently not in a position to say otherwise. But a quarrel remains. If we do build a sufficiently complex computer that in fact is conscious, how would we find out if it was? Since we don't know how to program it to be conscious or have any way of measuring consciousness we could never tell. So, even though consciousness in computers cannot be ruled out, any conscious experience in a computer would have to be accidentally caused and it will never be observed.

There is actually, as I write this, a project entitled *Blue Brain* in progress at EPFL in Lausanne that aims to make a functioning recreation of a biologically correct human brain, down to the molecular level in a computer. The project is a long way from reaching its goal but it is none the less on the right way. As of now the project has resulted in a biologically accurate model of a neocortical column on the cellular level. However, getting back to our discussion, in the FAQ section of the webpage the question "will consciousness emerge?" is asked. The answer is simply this: "We really don't know."[114]

6.1.1 Levels of Reality

In the original paper Bostrom argues that the simulation argument implies that there could be several "levels of reality".[115] The "real" world would be at the basement level, the simulated world would be another, and simulations within the simulation could yield yet another level of reality. If we are currently living in a simulation this means that there are at least two levels of reality: Ours and our simulators. Given the fact that computers, or computer programs could be conscious, on what level of reality would our conscious experiences emerge? This is a tricky question, and Bostrom does not address it in his paper. There are, as far as I can see, two possible answers to this question. I will try to give a thorough account of both. Our options are:

[a]. The simulation yields another, distinct, level of reality. This reality can be said to have its own physics consisting of simulated materials, like e.g. simulated brains. The conscious experiences of the Sims supervene on the simulated physical states in its simulated physical brain. So, on this account the conscious experiences emerge on the simulated level of reality.

[b]. The simulation is run on a computer on the original level of reality. The conscious experiences of the Sims supervene on the complex physical states in the computer that simulates them. In this case the conscious experiences emerge on the original level of reality.

Given how we experience our world and its content, the first option, [a], seems more intuitive. According to our natural conception of the world we do in fact have a physical brain, and the physical states of this brain are what cause our mental states, whether these physical states of our physical brain are simulated or not. This is also more intuitive since this would place our

[114] For more information on the Blue Brain project visit http://bluebrain.epfl.ch.
[115] Bostrom (2003). p. 9.

minds on the same level of reality as the world that we experience. On this account the physics we experience would be on the same level of reality as our beliefs about, and our sensations of the physics. To say that a physical object, and the quale it instantiates in a mind, are on different levels of reality seems counterintuitive. However, considering the fact that we need the appeal to quantum physics to escape epiphenomenalism, [a] might lead us to the conclusion that our mental states have no causal efficacy. To quote Bostrom: "Simulating the entire universe down to the quantum level is obviously infeasible."[116] And I would assume that a simulation of the human brain down to the quantum level is also pretty much infeasible. Even though such a simulation would need less computational power than to simulate the whole universe at the quantum level, it's still a significantly more complex simulation than the one proposed by Bostrom. So, following [a] epiphenomenalism is likely to be true.[117]

Following [b], on the other hand, our conscious experiences supervene on "real" physical states in the computer running the simulation in which we live. But this means that our experiences would not be on the same level as the things we experience. This doesn't seem particularly plausible. How can a mental state, on one level of reality, be about or represent something on a different level of reality? The best way to get around this problem would be to deny that the simulation gives rise to a distinct level of physical reality. Instead we could claim that the simulation does not yield physics, only conscious experiences of physics. With Searle's distinction between simulation and duplication it sounds implausible that a computer simulation would really constitute a physical world. This would mean that something like Berkelian *idealism* is true about the simulated world, where Berkeley's God is substituted for the structure of the simulation. Furthermore, this leaves our solution to the problem of downward causation intact since there is no need to simulate the quantum level. On the basement level of reality, quantum physics is already there. So, on account [b] epiphenomenalism is avoided due to the fact that the conscious experiences of the Sims are located at the basement level of reality.

So, the choice comes down to either [a]: Physical reality and epiphenomenalism, or [b]: Idealism and mental causation. Both options are unpalatable. If [a] is true then it's the end of the world, as Fodor said.[118] If [b] is true then the objects that we see around us do not consist

[116] Ibid. p. 4.
[117] Unless we can somehow argue that macrophysics isn't causally closed. Even though this seems implausible such arguments do exist. See for instance Sturgeon, Scott, "Physicalism and Overdetermination", *Mind, Vol. 107, No. 426* (1998).
[118] See quote on page 17f.

of anything physical. I would personally prefer idealism over epiphenomenalism any day of the week, but we really don't have any reason to assume that either one is true, even though neither is inconsistent. There is, however, a way of denying both idealism and epiphenomenalism. This is simply to deny that we are living in a simulation. If we do this then we'll find ourselves to be living in the real world, and we can enjoy both physical objects as well as causal efficacy of the mental.

6.2 Evaluating the Credence of the Simulation Hypothesis

Bostroms claims that if the third disjunct of the simulation argument, that the fraction of all people with our kind of experiences that are living in a simulation is very close to one, is true, then we should give an equally high credence to SIM (that we are currently living in a computer simulation). We should do so because we do not seem to have any particular reasons to believe that we are either real or simulated. Our empirical evidence is, more or less, indifferent to these two distinct possibilities.

The best objection raised against this is that on an externalist/ internalist conception a simulated human-like experience might be vastly different from a real human-like experience. But this objection doesn't cause any worries for Bostrom's argument, as long as we haven't got any reason to believe that we are simulated rather than real. Well, do we? Following the lines of David Hume's discussion about the search for the self we can conclude that we really do not know what kind of entity the self actually is. If I look into myself I will not find any entity that is *me*, neither real nor simulated. If I have impressions of being a human, living in the real world, this alone won't trump the possibility that I am a Sim, living in a simulated world. And, in consequence, if I have impressions of reading a paper on Sims in simulations, this does not give me any particular reason to think that I am reading a real paper rather than a simulated paper, even though these experiences may differ vastly. Since my own experience is all I have access to I will never be able to compare my paper-reading experience with those of others. This means that if I am a Sim, my experience is that of reading a simulated paper, and I will never know what it is like to read a real paper. How, then, I am supposed to be able to tell, without knowing what it's like to have real rather than simulated experiences, whether or not I am a Sim? The answer seems simple: I can't. So, my empirical evidence does not seem to influence the credence in SIM. Whether I am simulated or not, I will never find out, unless of course my simulators tell me this in a fashion such as described by Bostrom, or if

God reveals the true conditions of my existence to my immortal soul after I die. Those who wait will see.

All in all I'd say Bostrom's reply holds, and so given the truth of (3), following the fact that our knowledge of the real reality of our, and the worlds, existence is severely limited, we will have to ascribe a very high credence to SIM. But perhaps our previous discussion of the philosophy of mind could give us reason to doubt the indifference principle. We seem to have reason to believe that epiphenomenalism is false, since our mental lives seemingly has impact on our physical lives. It seems evident that our wanting causes our reaching and our itching causes our scratching. But is this knowable? No, unfortunately it isn't. There is nothing inconsistent in epiphenomenalism, and it cannot simply be ruled out because it's counterintuitive. On the other side we have idealism. This might not be as terrible as epiphenomenalism, but it's still not a comfortable view of the world. It seems obvious that objects in our world consist of physical stuff, that they are more than just "collections of ideas", as Berkeley called them. But, just like epiphenomenalism, it's a consistent view of the world, and it cannot be ruled out because of the fact that we'd like it to be false. As with SIM, our empirical evidence isn't strongly correlated with either epiphenomenalism or idealism, so these two will not be likely to have any effect on our credence in SIM. But it should also be said that our wanting the falsity of epiphenomenalism and idealism will influence our wanting the falsity of SIM, but wanting and believing are not the same thing.

6.3 Could it Be True?

Could it be true that I currently am a Sim? The answer to this question depends on how a Sim would acquire language. If the language of the Sims is given to them by their simulators, then it could indeed be true. But if the Sims language develops on its own, then there can be no causal link between the word 'Sim', in Sim-English, and the actual Sims. How the Sims acquire language depends on what kind of simulation they inhabit. If it's a simulation that is studied as an historical account, for scientific purposes, then there is a good chance that Sim-English develops on its own. If not, the simulators will be influencing their object of study which would seem to violate the ethics of scientific research. But this seems to presuppose that the world develops deterministically. If it doesn't, a simulated world is very likely to develop into a world that isn't anything like the original world. Unless determinism is true each historical simulation will have to be staged, and if they are staged the language will probably be given to the Sims.

If we're living in a simulation created for other purposes, like psychological or sociological studies, then there is a good chance that our utterances of 'I am a Sim' are true. This is due to the fact that in these types of simulations historical accuracy would be less important for the simulators than it would be to understand what the Sims are actually talking about. There is reason to think that the most common type of simulation in the future will probably be another, namely computer games. For instance, the computer game *The Sims* has sold more than 50 million copies, and the sequel, *The Sims 2*, has sold more than 100 million.[119] These games are actually simulations of people, living in a simulated world. And with *The Sims 3* launching soon the number of virtual people living inside computers is likely to go up. Of course these Sims are not conscious, but who knows, maybe they will be in the future, more advanced, versions of the game. If they will be, the words that they use (in *Simlish*, as the language of the Sims is called) are likely given to them by the creators of the game, so their utterances of 'I am a Sim' would probably be true.

So assessing the probability that the statement 'I am a Sim' is either true or false comes down to what type of simulated human-like observer will be most common. Following the principle of indifference proposed by Bostrom, given the truth of (3), our credence in the truth of 'I am a Sim' should equal the fraction of all human like-observers who are currently living in a simulation and who's word 'Sim' has a predetermined meaning given by their simulators. Accordingly, our credence in the falsity of 'I am a Sim' should equal the fraction of all human-like observers, living either in a simulation or in the real world, who's word 'Sim' has no such predetermined meaning. However, even though the statement might be false relative to some Sims, this still will not rule out the metaphysical possibility of them being Sims, and so Putnam's argument does not influence our credence in SIM.

6.4 Conclusion

Bostrom's simulation argument presupposes that consciousness could emerge from complex computer simulations of the human mind. Whether this assumption holds or not is a difficult question to answer. If we grant that mental states supervene on complex physical states, then there seems to be no reason to believe that these physical states have to be those of brains rather than of computers. We are currently not, and we're likely to never be, in a position to deny consciousness in computers. But on the other hand we are not in a position to assert it either. Our current standpoint in this question inevitably depends on intuition. Some find it

[119] Whitehead, Dan, "The History of The Sims", *Eurogamer* (2008). http://www.eurogamer.net/article.php?article_id=94927. 2008-05-08.

intuitively true that computers will eventually get to have conscious experiences; others find it just as counterintuitive. Consciousness is, unfortunately, not observable. But, if it were the case that consciousness does emerge from a complex simulation of the brain, then for the simulated mind either idealism or epiphenomenalism would have to be true if it's living inside a simulation. And following this, the counterintuitive qualities of these positions also make the simulation argument seem highly counterintuitive.

Given that it is true that mental states could supervene on complex physical states in computers and that the human civilization eventually simulates minds in large numbers, then following the indifference principle, we are right to believe that there really is a good possibility that we're currently living in a simulation. However, if we are, then either epiphenomenalism or idealism would have to be true. So if we find ourselves to have reason to believe that both epiphenomenalism and idealism are false then we will have reason to believe that we are living in the real world. But both these theories seem impossible to refute on the basis of empirical investigation. Since consciousness is unobservable epiphenomenalism will not be definitely refuted, and since all empirical investigation is consistent with idealism neither will it. Even though these positions are uncomfortable, they should not influence our credence in the theory that we are currently living in a simulation. They are more reasonably viewed as consequences of SIM rather than reasons to deny it.

But even if it's metaphysically possible that we're currently living in a simulation, and we have reasons to give this possibility a high credence, it's not at all certain that my statement 'I am a Sim' is true relative to me. For you see, if the simulation I am living is run without any influence from its simulators, other than the creation of the simulation, then my word 'Sim' does not refer to the predicament that I am in, so the statement 'I am a Sim' is false. But if I am living in a simulated world that has predetermined languages, then my word 'Sim' would have a causal connection to actual Sims, and I would probably be right when I utter the words: 'I am a Sim'.

Bibliography

Aranyosi, István A. *The Doomsday Simulation Argument* (2004). http://www.personal.ceu.hu/students/03/Istvan_Aranyosi/Doomsday%20Simulation/The%20Doomsday%20Simulation%20Argument%20by%20I.A.%20Aranyosi.pdf, 2008-03-14.

Bickle, John. ”Multiple Realizability”, *Stanford Encyclopedia of Philosophy* (2006). http://plato.stanford.edu/entries/multiple-realizability/, 2008-04-02.

Bostrom, Nick. *Are You Living In a Computer Simulation?* (2003). http://www.simulation-argument.com/simulation.pdf, 2007-11-11. (Also in *Philosophical Quarterly, Vol. 53, No. 211* (2003). Blackwell Publishing.)

Bostrom, Nick. “The Simulation Argument: Reply to Weatherson” (2005), *Philosophical Quarterly, Vol. 55, No. 218.* Blackwell Publishing, Malden, MA. (Also available at http://www.simulation-argument.com/weathersonreply.pdf.)

Brueckner, Anthony. “Brains in a Vat”, *The Journal of Philosophy, Vol. 83, No. 3*. (1986). The Journal of Philosophy Inc, New York, NY. (Also available at JSTOR: http://www.jstor.org/stable/2026572.)

Brueckner, Anthony. ”Brains in a Vat” *Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/brain-vat/, 2007-10-28.

Chalmers, David J. *Consciousness and its Place in Nature* (2003). http://consc.net/papers/nature.pdf, 2008-04-02. (Also in *The Blackwell Guide to Philosophy of Mind* (2003), edited by S. Stich and F. Warfield. Blackwell Publishing.)

Cole, David. ”The Chinese Room Argument”, *The Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/chinese-room/, 2008-03-26

David, Marian. ”Kim’s Functionalism”, *Philosophical Perspectives, 11, Mind, Causation and World* (1997). Blackwell Publishing, Malden, MA. (Also available at JSTOR: http://www.jstor.org/stable/2216127.)

Fodor, Jerry. “Making Mind Matter More”, *A Theory of Content and Other Essays* (1990). MIT Press, Cambridge, MA.

Gallois, André N. ”Putnam, Brains in Vats and Arguments for Scepticism”, *Mind, New Series, Vol. 101, No. 402.* (1992) Oxford University Press, New York, NY. (Also available at JSTOR: http://www.jstor.org/stable/2254335.)

Graham, George. ”Behaviorism”, *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/behaviorism/, 2008-04-10

Hájek, Alan. ”Interpretations of Probability” *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/probability-interpret/, 2008-04-21.

Hume, David. “Of personal Identity”, *A Treatise on Human Nature* (1739-40). http://www.class.uidaho.edu/mickelsen/ToC/ hume%20treatise%20ToC.htm, 2008-04-20.

Jackson, Frank. “What Mary Didn’t Know”, *The Journal of Philosophy, Vol. 83, No. 5.* (1986). The Journal of Philosophy Inc, New York, NY. (Also available at JSTOR: http://www.jstor.org/stable/2026143.)

Kant, Immanuel, *Critique of Pure Reason* (1787). Macmillan, London (1929). (Also available online, e.g. at http://arts.cuhk.edu.hk/Philosophy/Kant/cpr/.)

Kim, Jaegwon. *Philosophy of Mind* (2006). Westview Press, Cambridge MA.

Lau, Joe. ”Externalism About Mental Content” *Stanford Encyclopedia of Philosophy* (2003) http://plato.stanford.edu/entries/content-externalism/, 2008-04-25.

Levin, Janet. ”Functionalism”, *Stanford Encyclopedia of Philosophy* (2004). http://plato.stanford.edu/entries/functionalism/, 2008-04-10.

Putnam, Hilary. *Reason, Truth and History* (1981). Cambridge University Press, New York, NY.

Putnam, Hilary. *Representation and Reality* (1988). MIT Press, Cambridge, MA.

Schmidt Galaaen, Öisten. *The Disturbing Matter of Downward Causation* (2006). University of Oslo (Ph.D. Dissertation) https://webspace.utexas.edu/deverj/personal/ test/disturbingmatter.pdf, 2008-04-02.

Searle, John. *Minds, Brains, and Programs* (1980) http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-

00/bbs.searle2.html, 2008-04-18. (Also in *Behavioral and Brain Science* (1980), Cambridge University Press.)

Searle, John. *The Problem of Consciousness* (1994). http://cogsci.soton.ac.uk/~harnad/Papers/Py104/searle.prob.html (Also in *Philosophy in Mind: The Place of Philosophy in the Study of Mind* (1994), edited by M. Michael, J. O'Leary-Hawthorne and J. P. Hawthorne. Springer.)

Stoljar, Daniel. ”Physicalism” *Stanford Encyclopedia of Philosophy* (2001). http://plato.stanford.edu/entries/physicalism/, 2008-04-10.

Tye, Michael. “Qualia”, *Stanford Encyclopedia of Philosophy* (2007). http://plato.stanford.edu/entries/qualia/, 2008-04-12.

Weatherson, Brian. *Are You a Sim?* (2003) http://www.simulation-argument.com/weatherson.pdf, 2007-12-14. (Also in *Philosophical Quarterly, Vol. 53* (2003). Blackwell Publishing.)

Whitehead, Dan. ”The History of The Sims”, *Eurogamer* (2008). http://www.eurogamer.net/article.php?article_id=94927. 2008-05-08.