

Minds, Substrate, Measure and Value Part 3: The Problem of Arbitrariness of Interpretation

By Paul Almond, 11 May 2008

Website: www.paul-almond.com
Email: info@paul-almond.com

© Copyright Paul Almond, 2008. All Rights Reserved.

Minds, Substrate, Measure and Value

Part 3: The Problem of Arbitrariness of Interpretation

By Paul Almond, 11 May 2008

Abstract

This is the third in a series of articles exploring the relationship between minds and physical systems (substrates) on which they are based. The first article was *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence* (available at <http://www.paul-almond.com/Substrate1.pdf> [1], <http://www.paul-almond.com/Substrate1.htm> [1] or <http://www.machineslikeus.com/cms/minds-substrate-measure-and-value-part-1-substrate-dependence.html> [2]). The second article was *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence* (available at <http://www.paul-almond.com/Substrate2.pdf> [3], <http://www.paul-almond.com/Substrate2.htm> [3] or <http://www.machineslikeus.com/cms/extra-information-about-substrate-dependence.html> [4]).

The previous articles argued that probability in some thought experiments in which you are unsure about your status is based on the substrates on which algorithms are running and that the nature of the substrate is statistically relevant to the *measure* with which an algorithm runs and affects the measure of any minds associated with such algorithms.

This article provides a deeper explanation for this by discussing the real problem, which is *arbitrariness of interpretation*. Previously called “multiple realizability” by John Searle [6,7], this is the problem caused by the need to apply an interpretation to a physical system to say that it is running an algorithm and the possibility of applying any interpretation to obtain any algorithm, leading to an apparent observer subjectivity in the algorithms that a physical system is running. The idea of interpretation is formalized with a hypothetical machine, an *algorithm detector*, which uses a *detection program* to make interpretations, and other, similar, devices. It will be shown that the problem of arbitrariness of interpretation is really one of arbitrariness about selection of an appropriate detection program to use in an algorithm detector.

Searle argues that multiple realizability makes the strong AI hypothesis incoherent. This conclusion is unnecessary, although the strong AI hypothesis needs some clarification. Any observer subjectivity can be removed by admitting all interpretations as corresponding to algorithms that are running. This does not mean that all algorithms are equal. Some algorithms, and algorithms of some types, will occur with greater measure than others, due to a larger proportion of all interpretations “finding” them. The set of all interpretations is infinite, so absolute numbers of interpretations cannot be counted. Instead, the measure of an algorithm, or a type of algorithm, relative to that of another

algorithm, or type of algorithm, is defined as being the number of all interpretations that produce that algorithm, or type of algorithm, as a proportion of the number of interpretations that produce the other algorithm or type of algorithm, as the length of the detection program tends to (but does not reach) infinity.

This provides the real explanation for the “weak substrate dependence” or statistical substrate dependence which I argued to be necessary in the previous two articles [1,2,3,4]. It was suggested that “redundancy” or “inefficient use of matter” tends to increase the measure with which a particular algorithm may be running in a physical system. The reason for this is that redundancy is likely to allow more ways in which different interpretations can produce algorithms. As an example, if we imagine increasing the thickness of the wires in a computer then this allows more (speaking informally) ways for detection programs in an algorithm detector to make interpretations and find algorithms in that system, which will have an effect on the probability of being of being in various situations in thought experiments like those in the previous articles.

Admitting all interpretations like this removes observer subjectivity, making Searle’s “multiple realizability” argument against the strong AI hypothesis invalid.

Previous Article by Nick Bostrom

Since writing the previous two articles in this series [1,2,3,4], I have become aware that Nick Bostrom published an article in 2006 [8] considering issues of splitting and combination of minds and suggesting that consciousness can somehow exist with some sort of degree depending on the nature of the substrate. I would like to ensure that readers are aware of this.

Introduction

“Could I write a mind? Could I sit down and describe a scene, a thought and an emotion so well that it took life on the page?”
- *Divergence (Fiction)*, Tony Ballantyne, 2007 [9].

This article follows on from *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence* (which is at <http://www.paul-almond.com/Substrate1.pdf> or <http://www.paul-almond.com/Substrate1.htm> [1]) and *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence* (which is at <http://www.paul-almond.com/Substrate2.pdf> or <http://www.paul-almond.com/Substrate2.htm> [2]). These articles discussed the concept of substrate with regard to minds. A copy of *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence* has also been posted on the www.machineslikeus.com website at <http://www.machineslikeus.com/cms/minds-substrate-measure-and-value-part-1-substrate-dependence.html> [3] and a copy of *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence* on the same website at <http://www.machineslikeus.com/cms/extra-information-about-substrate-dependence.html> [4]. Comments about both of these articles can be posted here.

“Substrate” in artificial intelligence (AI) and philosophy of mind is the underlying system, often the physical system, that causes a mind to exist. Substrates might be human brains or computers (though some people like John Searle [5,6,7,10] dispute this). *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence* [1,2] argued that complete substrate independence in AI is an incoherent concept, using a thought experiment about combining computers, and *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence* [3,4] gave extra information to support this argument. The argument used in the first article was based on the assumption that the position being argued against was that different computers running identical algorithms correspond to different situations in which you could be and that the number of possible situations depends on the number of such computers. An alternative view would be that it is numbers of unique algorithms – not computers – that determines probability and my argument about combining computers would not be seen as a problem by someone who holds this view. To deal with this I said that the thought experiment about combining computers shows a special case of a more general problem, which is the problem of *arbitrariness of interpretation*. The problem is that the strong AI hypothesis states that minds are associated with processing of symbols by an algorithm running in some physical system, but interpretation is required to determine what algorithms are running in a physical system. Any symbols and computational states we want can be extracted from a physical system if we are prepared to interpret it in a sufficiently contrived way. This may seem only relevant in debates about the nature of mind, but it is actually relevant in thought experiments like those discussed in the previous two articles, and without a resolution of this issue probability is incoherent.

Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence [3,4] discussed the problem of arbitrariness of interpretation briefly. This article will give a more detailed discussion. It will be shown that there is vagueness in the generally understood idea of the strong AI hypothesis and that, without qualification, it is incoherent, meaning that we have no coherent method of calculating probabilities in thought experiments like those discussed earlier. The sort of problem being discussed here has been described by John Searle [6,7]. I will give a more formal description of the problem, which should make it clearer what the problem is, and that it really does exist, using a conceptual machine called an *algorithm detector*. Unlike Searle, I do not intend to try to destroy the strong AI hypothesis, but rather to show how it needs *clarifying*. The clarified strong AI hypothesis which results will not be a watered-down version: it will actually be an *extreme* version of the strong AI hypothesis.

This article will discuss the strong AI hypothesis more generally than the previous articles, but I will make sure that the relevance of physical substrate, an important idea in this series, is not overlooked.

The Strong AI Hypothesis

The term “Strong AI” comes from arguments made by John Searle [10] which distinguished between two philosophical positions about the mind. The “strong AI” view was regarded as being the philosophical position that a computational system can think

and that it has a mind. The “weak AI” view was regarded as the philosophical position that a computational system might think and act like it has a mind, but that this does not necessarily imply that it *really* has a mind.

Some confusion might be caused by the term “strong AI” now because many AI researchers use it differently to how it was first used. They regard strong AI as merely the view that a suitably programmed computer can behave intelligently and behave as if it has a mind, regardless of the philosophical issue of whether or not it has a mind [11]. The term “strong AI hypothesis” is therefore used now to describe the claim that a properly programmed computer would have a mind.

In the earlier articles I discussed “strong AI” but I hope it is apparent that I was referring to “the strong AI hypothesis” – the term that I should have used.

The strong AI hypothesis, then, is basically the idea that computers can have minds, but there is an important implication of this. If an appropriately programmed computer can have a mind then this follows:

A mind exists when the appropriate algorithm is being run on a physical system.

We should not get distracted here by irrelevant issues. This does not necessarily mean that the algorithm must be written by humans, or must be run “tidily” with elements that we would recognize easily as instructions. For example, an artificial neural network in this view could have a mind if it is doing the right sort of processing because the interaction of all of its artificial neurons gets the correct algorithm run. Similarly, a human brain could be regarded as running the correct algorithm, even though human brains clearly are not programmed like conventional computers, because the interaction of all the parts in the human brain could, in principle, be formally described by an algorithm and the fact that you and I experience consciousness would indicate that whatever information processing occurs in such an algorithm is equivalent to a mind. This point is lost on many people who think that an artificial neural network is philosophically different to a conventional computer.

There is a problem with this view, however. I will call it *the problem of arbitrariness of interpretation*. It has previously been raised by John Searle who calls it *multiple realizability*. The problem is that there is no obvious, formal method of saying what algorithms a physical system is running.

The Problem of Arbitrariness of Interpretation

Searle’s best-known argument against the strong AI hypothesis is probably his *Chinese room* argument, widely regarded as invalid by advocates of the strong AI hypothesis, but his observation of the problem of arbitrariness in interpretation is potentially deeper. This is what Searle says:

“...computation is not an intrinsic process in nature like digestion or photosynthesis, but exists only relative to some agent who gives a computational interpretation to the physics. The upshot is that computation is not intrinsic to nature but is relative to the observer or user.” [6]

What this argument says is that symbols and meaning do not exist in a physical system by themselves, but need an observer to allocate meaning to the physics. For example, 1s and 0s do not exist by themselves inside an electronic computer. Instead, we decide that particular voltages correspond to 1 and other voltages correspond to 0, allowing us to say that an electronic computer has a particular computational state. We interpret the electronic computer’s physical state as having a certain meaning.

The problem with this is that the obvious interpretation is not the only one. It would be possible to interpret any physical system in any number of ways to get any computation we want from it. Searle calls this *multiple realizability* or *universal realizability* and says this:

“The same principle that implies realizability would seem to imply universal realizability. If computation is defined in terms of the assignment of syntax, then everything would be a digital computer, because any object whatever could have syntactical ascriptions made to it. You could describe anything in terms of 0’s and 1’s.” [6?]

We must interpret what a computer is physically doing to get 1s and 0s and an algorithm, but Searle is saying that you can get any algorithm out of anything if you make an appropriate interpretation, a point he makes clear in the following statement:

“For any program and any sufficiently complex object, there is some description of the object under which it is implementing the program.” [7]

Searle gives an example:

“...the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movement that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain.” [7]

(Note: *Wordstar* is a word processing program for personal computers.)

You would have to do some of these interpretations in a contrived way and some people would say that the computation really comes from the way you interpret the matter, rather than the matter itself, but – and this is an important point – *there is nothing in the strong AI hypothesis which tells us which interpretations of a physical system are reasonable and which are contrived.*

I disagree with Searle’s conclusion that the strong AI hypothesis needs discarding. It will be obvious now that I think that the strong AI hypothesis needs clarifying. I do, however,

accept the basic point made by Searle – that this apparent subjectivity of interpretation is a problem.

We can get an idea of this problem of interpretation from part of Greg Egan’s novel, *Permutation City* [12], in which arbitrariness of interpretation was actually a major plot device. In *Permutation City*, two characters, who are “copies” made by mind uploading [13,14], hide in a computer system by having themselves placed in a complex, encrypted way, among the normal processing of that system. An observer may see that system, for example, running a virtual reality simulation of a fountain, but a bit of the fountain simulation is being stolen to do a bit of the job of running these two stowaways and an environment for them. This is similar to steganography, the practice of hiding messages in other data, such as in computer image files.

Suppose we wanted to check for intelligent agents hiding, encrypted, in a system like this. We would need to run some sort of analysis of the computer – some sort of decryption – to show that they were there. If the decryption is not complex enough it may not find them, but if the decryption is too complex then it may find agents just by arbitrarily interpreting the computer hardware in the right way.

As another example, I will give this encrypted message:

AY5NSL Q23 FQK Q33WDN 7UIKDLQPD IKELDP HJK IEKSL POLSO

I can assure you that there is a way of decrypting that to get “Hello world.” Alternatively, you can decrypt it to get “The quick brown fox.” Of course, decrypting it in this way may be contrived, and whatever decryption process we use may need to be supplied with more information than we are actually getting out of the message, but any decryption process needs to be supplied with some information. Ultimately, there is no way of saying that one particular interpretation is valid while others are invalid. You might claim that one method of decrypting the message recovers what was encrypted in the first place, but surely the history of the message does not matter: only what we have now should be an issue. You might say that the simplest interpretation that produces a mind is the correct one, but it should be easy to imagine situations with multiple interpretations.

The previous articles [1,2,3,4] described thought experiments in which whether or not a system is running a particular algorithm can affect probabilities. There is no escape from the problem of arbitrariness of interpretation, no matter how we approach probability. When we base probability on counting computers, arbitrariness of interpretation causes a serious problem in deciding which computers are running algorithms that count, but we might try to ignore this by using some “common-sense” approach. Such a common-sense approach would not be very effective, though: the thought experiments about probability in the previous articles are really special cases of the problem of arbitrariness of interpretation. We are more obviously confronted by the general problem when we get probabilities by counting unique algorithms rather than computers, because we must be able to say how many unique algorithms are in a particular system and arbitrariness of

interpretation has the potential for generating unlimited numbers of unique algorithms to mess up our counting.

Why the Problem of Arbitrariness of Interpretation Cannot Be Ignored

The Idea of Reference Class

The problem of arbitrariness of interpretation cannot be ignored because, if the strong AI hypothesis or anything like it is correct, the decision about what algorithms are being run by what physical systems determines our reference class of possible situations in thought experiments. There are descriptions of what “reference class” means in terms of Bayesian probability, but to put it simply (and maybe not as rigorously) the reference class is the set of possibilities from which some statistically based selection is thought to be made. As an example, if I think that there is a 10% chance that the next person who talks to me will be carrying a briefcase then this implies that I have in mind some “reference class”, some imaginary collection of all the people who are possible candidates for being the next person I meet next, and that 10% of the people in this reference class carry briefcases. I might arrive at my reference class in a number of ways. For example, if I am on a train I might decide that my reference class is everyone on the train, and I might arrive at my 10% probability if I have reason to believe that 10% of the people on the train are carrying briefcases.

The Problem

The first two articles of this series [1,2,3,4], when discussing weak substrate dependence, discussed thought experiments in which you are unsure about the nature of the physical substrate on which your mind is based. Situations like these arise in the novels *Permutation City*, by Greg Egan [12], and *Recursion*, by Tony Ballantyne [15]: in these books characters are confronted with the possibility that they may not be based on organic brains, but may be computer simulations of their brains running in a virtual reality. The reference class in these thought experiments depends on what algorithms are running where. For example, if you are asked if you think you could be existing in your own brain, or on some computer, then whether or not that computer is running an appropriate algorithm that could be considered adequate for providing your experiences is relevant. If the computer is not running such an algorithm then it is impossible that you could be running on it and the probabilities for any other situations in which you could be are correspondingly higher. If you know that the computer is running such an algorithm then you have to face the possibility that your experiences could be due to the running of that algorithm, and the probabilities of being in any other situations must decrease. The previous articles ignored this and concentrated on the issues of combining computers and when we say that a computer running in an algorithm is one computer or two computers. That is also, in a way, a reference class issue, but the reference class problem presented by arbitrariness of interpretation is more basic: without being able to say whether or not a given algorithm is running on a computer we cannot even begin to determine our reference class of possible situations in thought experiments like this and any probability

is incoherent. The reference class problem resulting from arbitrariness of interpretation is worse than this, however: it now becomes a serious issue outside what may seem to be the contrived circumstances of the thought experiments. This is because, in the absence of any way of firmly saying what algorithms are running on what physical systems every situation becomes like those in the apparently contrived thought experiments. There is no way of saying that any given physical system is not running an algorithm that corresponds to your mind, and which should be considered part of your reference class of possible situations.

If you think that particular physical systems are running particular algorithms, arbitrariness of interpretation will result in incoherency no matter how you choose your reference class. The previous articles [1,2,3,4] mainly discussed the idea that the reference class should be made of each version of the algorithm that is running on a physical system somewhere. Some people, however, disagree with this and say that we should only be interested in each *unique* algorithm. This distinction will not help with the problem of arbitrariness of interpretation because both approaches, and any conceivable approach, will depend on what algorithms are running in physical systems – and arbitrariness of interpretation makes it incoherent to ask about this.

It may seem that common sense should sort this out and that it is just obvious that a given algorithm is being run by some physical systems and not others, but common sense and human intuition are insufficient: probability is supposed to be something which we can assign using formal methods. There needs to be a formal method of deciding whether or not a given algorithm is running in a given physical system, or at least some idea of how one would work. As we will now see, this causes a big problem.

The Algorithm Detector

Basic Idea of the Algorithm Detector

It will be hard for many people to see that the problem of interpretation exists. So quickly and so easily do we apply computational interpretations to physical systems, we are not aware, unless we think about it, that we even needed to apply an interpretation. For example, when a circuit designer sees the circuit diagram for a calculator it will seem obvious to him/her that he/she is “seeing” something that implements the algorithms that we associate with calculators. Many advocates of the strong AI hypothesis, if they could see the structure of a human brain in sufficient detail, and could make sense of it, would say that they are “seeing” something that implements (albeit in a messy way) an algorithm corresponding to a mind. I want to show that reaching such a conclusion is by no means trivial and that, unless we do something about it, it implies some subjectivity.

To show that the problem of interpretation exists, rather than discussing how human brains “see” algorithms in nature, I will simplify things by introducing a conceptual machine for detecting algorithms. I will call this machine an *algorithm detector*. An algorithm detector is a stand-in for a human brain, with none of the abilities of a human brain except the ability to “see” algorithms in physical systems.

If there is an obvious way to interpret physical systems as having algorithms then we should be able to formalize that interpretation method as a computer program. Nothing extraordinary is being claimed here. I am merely saying that if we imagine any reliable human knowledge about determining what algorithms a physical system contains, we should be able to formalize that knowledge as a computer program. The algorithm detector is a machine that has such a computer program, together with sensory devices that allow it to examine a computer, a human brain, or any other physical system and determine what algorithms that system is running.

Description of an Algorithm Detector

Here is a more detailed description of how an algorithm detector would work:

The algorithm detector analyzes physical systems and determines what algorithms they are running.

A computer in the algorithm detector runs a *detection program*. The detection program controls the movement of *physical probes* to get data about the system that it is analysing. The detection program contains a formally expressible method, or a number of such methods, for determining the presence of algorithms running in a physical system.

Physical probes move around under control of the detection program. Each has a tip that can measure physical values and send the data back to the detection program. For example, a physical probe may measure the local magnetic or electrical field strength. Physical probes are not subject to any conceivable damage from the environment, nor any conceivable interference with their movement. They can be moved, ghost-like, through matter and have no noticeable effect on the physical system on which they are used. For example, a physical probe could move through the casing of a computer to measure physical values inside, or it could be placed inside a human brain – all without the computer being affected or the human subject being aware of the probing.

The number of physical probes available to an algorithm detector should not be an issue, but it should not constrain any philosophical discussion. We may imagine an algorithm detector with a limited number of physical probes available for simultaneous use, or we may assume that the number of available probes exceeds that which is needed in any situation which we are considering. It may be simplest just to imagine the algorithm detector as having an unlimited number of physical probes.

An output device gives the results of the algorithm detector's analysis. We will assume that this is a printer.

To use the algorithm detector it is placed near the physical system that we want to analyze and turned on. The detection program starts to move the physical probes around inside the system, taking readings of physical values in it. The detection program analyzes the data that it obtains from its probes and intermediate results of this analysis may cause it to make further movements of its probes. When it knows enough about what

is inside the system the detection program prints a list of algorithms that the physical system is known to be running. This may be done by printing the program listings of all the algorithms, or by assigning each algorithm a short code (which could only be done if the machine is looking for a small subset of all possible algorithms, or if the results are prioritized in some way).

Notes on the Algorithm Detector

The above description only specifies what an algorithm detector does, and not how well, or how generally, it does it.

An algorithm detector may be designed to detect only certain algorithms that interest us or to detect all algorithms. An algorithm detector that detects only a single algorithm may be programmed with the specific details of the algorithm for which it is looking, but this does not mean that an algorithm detector must be programmed with details of every algorithm that it is going to find: it could have some general method for finding any algorithm. If an algorithm detector is not designed to find a particular algorithm, but has some general method for finding any algorithm, then I will call it a *general algorithm detector*. A general algorithm detector is not necessarily guaranteed success. Being able to find any algorithm is not the same as being able to deal with any way of *encoding* algorithms in the physical substrate.

Algorithms may be encoded or “hidden” in a physical system in different ways. An algorithm detector may be designed to find those algorithms encoded in only a particular way, or to find algorithms encoded in any way.

I define an *omniscient algorithm detector* as one capable of detecting any algorithm encoded in a physical system in any way whatsoever. An omniscient algorithm detector would need to contain every “knowably sound” method for detecting any algorithm in a physical system. This has some similarity, in terms of scope, with the conceptual machine proposed in Penrose’s “proof” [16] (flawed in my opinion [17]) against computational AI. An omniscient algorithm detector may not necessarily be something that we could actually make, but most of the discussion in this article will involve algorithm detectors that are not necessarily omniscient. Most of the consideration will be about quite realizable algorithm detectors that only contain the methods that anyone has got round to putting into the detection programs.

An algorithm detector might be able to find a large number of different algorithms, but it may be inconvenient to use: if it finds many algorithms then it could produce a huge printout. We might imagine a more user-friendly algorithm detector which can accept an input consisting of the program code of one or more algorithms and which then attempts to find those algorithms only. This might require the algorithm detector already to have the specific capability of finding those algorithms, so that it will only work when certain algorithms are provided as inputs, or it could be a general algorithm detector, able to search for any algorithm that we supply as an input.

The Problem with the Algorithm Detector

If the strong AI hypothesis, as most people understand it, is coherent and well-defined, we should be able to formalize it. Any understanding we have of what constitutes an algorithm running on a physical substrate should be capable of being formalized as an algorithm and used to run an algorithm detector. The strong AI hypothesis in itself is inadequate for this and does not tell us how to make algorithm detectors. It is therefore incomplete.

We will consider an example:

Suppose we want to find algorithms that play chess. We hire a computer programmer to write a detection program to detect algorithms that play chess and we put this in an algorithm detector or similar device. We agree to pay the programmer a bonus for every chess algorithm that his program finds.

We set up our machine next to a chess computer and it moves its physical probes around inside it before giving us a printout indicating that it has found a chess algorithm. This should not surprise us, because we expect to find chess algorithms in chess computers.

We then position the algorithm detector near another machine and it analyzes this machine before reporting that it has found another chess algorithm. This puzzles us, because we do not see any obvious signs, from the machine's behaviour, that it is doing anything related to chess. On looking inside the machine, however, we see that a wire that should send the output to a monitor is broken, so the machine *was* running a chess algorithm all along inside: we just could not see it, but the algorithm detector's physical probes and detection program had determined its presence from the internal structure of the machine.

Next, we place the algorithm detector near a third machine and it tells us that this machine is running a chess algorithm. As with the previous machine, we cannot find any external evidence that it is running such a thing. When we look inside, unlike with the previous machine, we do not find a simple failure of some output device. When we investigate further, however, we find that the machine is running a chess algorithm. The way in which the hardware is used to encode the algorithm is somewhat complicated. For example, an individual 1 or 0 is not encoded as a single voltage level: instead there are a number of different permutations of voltage levels in a number of different wires that correspond to a "1". Nevertheless, after we have been looking at the insides of the machine for a while, it is obvious to us that it is running a chess algorithm.

We now place the algorithm detector near a fourth machine and it detects a chess algorithm here too. On looking inside the machine, however, we merely see a lot of complicated hardware. We see no evidence of any chess algorithm. The programmer demands his finder's bonus. We are sceptical, however, that there is any chess algorithm in this machine. We challenge the programmer on this and he assures us that the algorithm is there, but that it is very hard to see it. He tells us that he has more experience

of finding algorithms than we do and that he has formalized this experience in the detection program that he wrote for us. We wonder if he rigged his program to falsely detect chess algorithms to defraud us of money, so we inspect the detection program's source code. The program is long and so complex that we cannot understand it. We tell the programmer that the great length of his program is suspicious. What if it actually contains the very chess program for which it is supposed to be searching within itself, so that it can just do some irrelevant and complicated computation and then, in some circumstances, print the listing of a chess program to make it appear that it found it? The programmer replies that the program is merely very long because it can detect chess programs that are hidden in very complicated ways and only a long program can have the sophistication needed to find such programs.

This is the problem. How can we ever know what a valid algorithm detection program is and what is not? Without a way of being able to formally describe what constitutes a valid detection program the strong AI hypothesis is incompletely defined, because it asserts that machines running certain algorithms have minds without stating what it means to say that a machine is running an algorithm. As the discussion of reference class and probability pointed out, this is not a trivial consideration. Without a resolution of this the strong AI hypothesis, as understood by most people, with its lack of a formal interpretation method, gives incoherent results when we try to obtain the probabilities that we are in various situations.

Incidentally, another, similar, point might be made – that an algorithm detector, if it existed, would be a physical system, that the problem of interpretation also applies to the algorithm detector's software itself and that it is impossible to say, without interpretation, what detection program an algorithm detector is running, launching us into an infinite regress of interpretation. I will not be making such an argument because, although it may initially appear tempting, it would be weak.

Inputs and Outputs

An algorithm detector need not be limited to providing us with the program listing. We should be able to imagine an algorithm detector that also reports to us the sequence of inputs and outputs associated with that program. If the algorithm itself is hard to find then these will be just as hard to find. For example, in Greg Egan's novel *Permutation City* [12] and Tony Ballantyne's novel *Recursion* [15], simulations of human minds run hidden in virtual reality simulations. An algorithm detector could be imagined that found programs corresponding to human minds like this and told us what inputs and outputs were associated with them. As another example, an algorithm detector might find a program that is doing arithmetic and tell us what numbers it is adding. I will call such an algorithm detector an *input and output reporting algorithm detector*.

We could go further and imagine a special case of an input and output reporting algorithm detector that not only finds inputs and outputs, but actually allows us to *interact* with the system that it has found. I will call this an *interactive algorithm detector*. Such an algorithm detector would have physical probes that are not limited to

measuring local physical variables, but can change them as well. The detection program would tell us what outputs are coming out of the algorithm and would prompt us to provide inputs, for example using a keyboard. It would then direct the algorithm detector to use its physical probes to alter the physical system to ensure that the algorithm receives those inputs. Depending on how the algorithm is encoded into the physical system, providing a single input of a 1 or 0 from the “outside world” might involve changing many physical variables.

As well as algorithm detectors equipped to tell us about algorithms and their inputs and outputs we should also be able to imagine less powerful devices that lack the capability to tell us what algorithms are being executed, but can merely find them and tell us what input and output values are associated with these algorithms. I will therefore define these types of detector:

An input and output reporter is a device that tells us what input and output values are associated with algorithms in a system. An input and output reporter is just an input and output reporting algorithm detector without the capability of actually finding out with what algorithm it is dealing.

An interactive input and output reporter is a device that tells us what input and output values are associated with algorithms in a system and allows us to interact with the system by changing output values. Such a device is therefore just an interactive algorithm detector without the capability of actually finding out with what algorithm it is dealing.

Both of these devices have some characteristics in common. The same system of physical probes would be used, able to move through matter without restriction under control of a detection program, and able to read physical values. The detection program also applies an interpretation to the data that has been received and tells us what the device has found. In the case of an interactive input and output reporter, the detection program would also prompt us for input values which are then supplied to the algorithm that the system has found by using the physical probes to change physical variables.

Limitations on Interaction

There will be situations in which interaction with the algorithm cannot occur in real-time or in which it cannot happen at all. This could be because there may be some interpretations for which computational states do not correspond to extractions of information from the physical system at different times and in the same order. For example, an algorithm might be detected that goes through computational states 1, 2 and 3, but it may be that Computational State 3 is obtained from the state in the physical system that occurs first, Computational State 2 is obtained from the state in the physical system that occurs next and Computational State 1 is obtained from the physical state that occurs last. That is to say, the physical system is being viewed backwards to make the interpretation. Some interpretations could be independent of time in the physical system, the information about all the computational states being obtained from the physical system at the same time. An example of this would be a hardcopy printout of a computer

program. In principle, the right sort of interpretation applied to this could obtain a description of an algorithm moving through computational states. Other interpretations may be so far removed from our intuitive expectations that we could not even categorise them as easily. For example, an interpretation could involve extracting information from the physical system at many different places and times to make a single computational state, so that from our point of view the physical basis of each computational state is delocalized in space *and* time.

This will need considering later. For now, it merely needs pointing out so I do not appear to have missed it. Although many extreme interpretations will be like this (I will define these as *atemporal interpretations*.), there will many that will not – interpretations that work as we would intuitively expect, with the physical state of the system corresponding to some computational state and a later physical state of the system, in time, corresponding to the next computational state, and I will define these as *temporal interpretations*. We could interact with an algorithm found using a temporal interpretation, but not one found using an atemporal interpretation.

This, incidentally, has possible implications for our view of time that will be discussed later.

How I Will Deal With This

An obvious objection to the above is to say that I am saying that no formal interpretation method exists simply because I cannot see one and that the strong AI hypothesis is therefore invalid. I am not trying to destroy strong AI and join the Searle or Penrose camps – or announce some position of my own against strong AI: I think that both Searle and Penrose are incorrect. Rather, I think that our view of the mind has to rely on computation, but that the obvious lack of any formal interpretation method in strong AI means that something needs to be done. I will be trying to show that none of the *obvious* things that can be done work and that we are left with the *strong suggestion* that a drastic approach for getting a formal interpretation method is needed. I will be describing this approach later and therefore suggesting what is intended to be a repaired version of strong AI. What I suggest will not reject the relationship between minds and computation. Rather, computation will be shown to be important to any understanding of minds and other things. The way in which computation relates to minds, will, however, be different to what is argued against by Searle. Rather than minds being viewed as computational in themselves, minds will be shown to be correctly considered as associated with computationally based descriptions (descriptions that are expressed formally) and computationally based interpretations (interpretations that are expressed formally) of physical reality.

Objections

Objection 1: Of course you will be able to find arbitrary algorithms when you are allowed to *interpret* reality, but there is no reason to make any interpretation at all. You should simply take algorithms that you can observe directly, without any

interpretation, as existing, and ignore any algorithms that can only be “found” by interpretation.

Answer

This objection shows a misunderstanding of what is involved in “perceiving” an algorithm. We *never* directly perceive algorithms but instead perceive physical variables in the world around us and interpret these to decide that some algorithm is there. When we read a printed program listing we decide that particular patterns of ink on a page can be interpreted as symbols which in turn can be placed in a particular order and correspond to a sequence of instructions in a language, which can be interpreted as an algorithm according to the rules of a language. When we say that a computer is running a program we have decided that particular patterns of voltage, or other physical variables, correspond to 1s and 0s, according to some rules of interpretation, and further rules of interpretation dictate how these 1s and 0s are arranged to give the computation states through which the computer is supposed to be going.

It becomes obvious that all claims of the presence of algorithms rely on interpretation when we imagine an algorithm detector with no detection program inside it. It would be incapable of telling us what algorithms are there. It is no different when humans are doing the observing. A truly “direct” observation of the presence of an algorithm would require the observation to be made by human sense organs, without the need for any information processing at all. In reality, the brain is needed to process the sensory information and decide that a computer is being observed running an algorithm and, to anyone who subscribes to the strong AI hypothesis, this must be equivalent to an interpretation algorithm.

Objection 2: You are complicating things needlessly. Humans intuitively know when a computer is running an algorithm and when it is not.

Answer

For the strong AI hypothesis to be coherent, the concept of what constitutes an algorithm should not depend on intuition, but should be formally defined.

Objection 3: Some interpretations are just stupid.

Answer

This is just the previous objection stated in a different way. Unless you can formally state how we tell sensible interpretations from stupid ones there is no formal idea of what constitutes the running of an algorithm.

Objection 4: You can make any interpretation you want to find any algorithm you want, but some of these interpretations would need to be made by very long algorithms. For the more extreme interpretations, the detection program might

need to be longer than the algorithm being detected and might effectively be pretending to “find” an algorithm that is actually contained within itself. Short detection programs can be accepted, but long ones are contrived and invalid.

Answer

This objection may seem plausible when we consider it just for really extreme cases because in one way it is correct: the objection describes the “mechanics” of how a detection program may be contrived to find any algorithm that we want.

The problem with this objection, however, is that that there is no specific point at which it comes into force and the detection program becomes “contrived”. Suppose we are sure that detection programs with a maximum length of L bits are valid. What if we increased L by one bit? It is hard to see any obvious philosophical reason why an addition of one bit in length should make much difference. We could keep increasing L in one bit increments, so that the detection program could make progressively more contrived interpretations. For what value of L should we say the interpretations are permissible? The issue is not trivial because it will affect probabilities in the sorts of thought experiments that we have been considering [1,2,3,4]. The reference class issue means that choosing some arbitrary value of L would not do, because this would make probability values, which should be objective, dependent on this arbitrary choice and we would have no coherent concept of probability.

I do not think there is any obviously correct value of L and this is a problem.

One approach could be to suggest that while there may be no obvious cut-off point, the interpretations become increasingly contrived as L increases. Doing this, however, means accepting that all interpretations are valid – just that some have some “more validity”, in some sense than others. This would strongly imply some kind of measure of minds, as I stated must be the case in *Minds, Substrate, Measure and Value – Part 1: Substrate Dependence* [1,2]. This is the approach that I will be suggesting for dealing with the problem of arbitrariness of interpretation.

Objection 5: Programs and minds are real only if they interact with reality. The more extreme interpretations made by an algorithm detector’s detection program could not possibly interact with reality. As an example, you may be able to make a contrived interpretation of the matter in a wall as running the *WordStar* word processing program, as described by Searle in his example [7]. This *WordStar* program could never be used to word process any real documents. A real word processing program is accessible through the keyboard and can send information to the outside world. The program said to exist in the contrived interpretation can never do that. This tells us which programs are real and which are not.

Answer

We should ask what is meant in the objection by “interact with reality”. If an appropriate algorithm detector, or similar device, found the *WordStar* computer program in the atoms of the wall in Searle’s example then who is to say that the algorithm is *not* interacting with reality? The detector could report the inputs and outputs to us and it could be argued that the inputs and outputs do exist, and that we are not intelligent enough to see that they are there. Against this, it could be argued that the system should interact with “macroscopic variables” in reality that are apparent to us, but it could also be argued that there could be other aspects of reality that are not apparent to us – that the algorithm’s behaviour changes variables in reality that, just like the algorithm itself, are hidden to us.

Why should we say that the effects of running an algorithm have to be noticeable in the “outside world”? Suppose we applied that standard to your own brain. What if you were so completely paralyzed that your conscious mind could make no noticeable outputs whatsoever? This does not need to be hypothetical: humans can end up in this situation as a result of neurological conditions. An observer applying the standard that “interaction with reality” is required should pronounce you non-existent at this point, yet I think most people should be able to imagine being in such a state and feeling very much real! A simpler example, and one that may appear at first to be deceptively easier to dismiss, is that of a human placed in a sound-proof metal box through a door which is then welded shut. Someone determined to say that a system must make meaningful outputs to be regarded as existing could say that there are no outputs to indicate the presence of any intelligence in the box. We could further use hypothetical mind uploading as an example. If you subscribe to the strong AI hypothesis you must think that an adequately constructed computer program simulating a human mind must be conscious, so you should be able to imagine the situation of being such a program. What if you were in such a situation, interacting with the outside world, and your inputs and outputs were disconnected from the outside world and connected to a program providing a virtual reality? From the point of view of an outside observer, you might seem to be non-existent, without a careful examination of the internal workings of the computer equipment running you, but would it be reasonable to think your own mental experience would cease just because of this? You could claim that these situations are not the same, because you know that there is really a paralysed person there, or that there is really someone in the metal box, or there is really a simulation of someone running in a virtual reality, but after the algorithm detector gives us its results we could equally well say we know that there is an algorithm hidden inside some physical system – even if it is found by some extreme interpretation, because there is no obvious cut-off between reasonable interpretations and extreme ones.

This idea that inputs and outputs may not be necessary to say that an algorithm exists could be argued against by stating that, even if they are not obviously provided, inputs and outputs can always be found for any information processing system. As an example, in the case of a computer running a simulation of a brain and the virtual reality with which it is interacting, some variables in that program (or memory addresses, or registers, or some combination of variables, etc) must correspond to the inputs and outputs that the

brain would have if it were connected to the outside world. As another example, if an algorithm, with no obvious inputs or outputs in the “outside world”, is adding numbers then some variables or similar things in the program must correspond to the numbers that are being added. In cases like this we should be able to imagine, if we have the physical capability to reach into systems, physically affecting it to change these variables, memory locations, registers, etc and similarly to read the values stored in them. In this way, we could regard these as inputs and outputs that are not yet connected to the outside world and, by our process of reaching into the system we could connect them to the outside world. This should be easy to imagine in the case of the electronic computer. We could imagine having some diagnostic machine with a keyboard and screen, with wires running from it into the computer. Some of these wires would be used to measure the voltages at some places and some of them would be used to allow us to control the voltages at some places, allowing us respectively to read values and change values in the computer. By doing this, we could argue that we have effectively made “hidden” outputs observable in the outside world and gained the capability of affecting “hidden” inputs.

Doing this could allow us to detect an algorithm in a slightly different way to that previously discussed: not by looking in detail at the internal processes of the algorithm itself, but by finding appropriate inputs and outputs in the system and showing that the relationship between them is consistent with the execution of some algorithm. Although different, this approach would be just as valid. In this context, the objection would be about how “apparent” the inputs and outputs of the system are. The idea of the objection would be that, even if a human were paralysed, in principle, by connecting wires to the appropriate parts of the brain corresponding to inputs and outputs we should be able to interact with the brain and show that an algorithm corresponding to a mind is there. Likewise, if a human is locked in a metal box then by reaching into the box we can find appropriate inputs and outputs that tell us a mind is there. With the example of the simulated brain in the virtual reality, by running wires into the computer we can find ourselves interacting with the mind. The idea would be that in all these cases we can find inputs and outputs that are apparent and can be readily measured, whereas in extreme situations this will not be the case.

The objection is unsuccessful, however. Whether we are obtaining the algorithm directly by inspecting the system, or obtaining inputs and outputs that tell us it is there, interpretation is required. For example, if we decide that a paralysed human has a mind, just by “finding” appropriate inputs and outputs we have to apply an interpretation in which particular firing operations in neurons correspond to outputs and particular changes in neurons correspond to inputs. When I described the algorithm detector I described a second type of machine called an *input and output reporting algorithm detector* which is able to tell us what input and output values are associated with an algorithm that it has found. I also described the input and output reporter and the interactive input and output reporter. These machines would all do for inputs and outputs what the algorithm detector does for algorithms. They would find inputs and outputs associated with an algorithm that may be hidden in a system and tell us what they are. If we decide that we really need to be able to observe the inputs and outputs of some algorithm then the problem merely changes from what interpretation to apply – what detection program to use – in the

algorithm detector – to what interpretation to apply – again, what detection program to use – in the device that will be telling us what the inputs and outputs are. All of the same issues now become relevant. It is meaningless to talk about “directly” observing an input and output when some interpretation will always be required. If we decide that only short detection programs are valid then, just as with detecting algorithms, we have problem of there being no obvious cut-off point between acceptable and unacceptable detection programs.

One point could be made about the capability of humans to interact with systems. It could be argued that by a system being apparent in the real world, we should be able to observe its inputs and outputs, but that (at least in principle) we should be able to interact with the system ourselves by means of these inputs and outputs. This is not the case for all extreme interpretations. Those interpretations that I called *atemporal interpretations* are ones in which the order of computational states does not correspond to physical states of the system occurring in the same order in time (or maybe any order in time). This objection could have merits with respect to some extreme interpretations – those which I called “atemporal interpretations” – although I think it ultimately unfounded even in these cases, as I will explain later. The problem is that it does nothing to invalidate any of the unlimited supply of *temporal* interpretations. Even if we decide that interpretations have to be made by “playing by the rules” and that each successive computational state must be correspond to a later state of the physical system than the previous one, we could still make temporal interpretations as extreme as desired to obtain any algorithm – or any collection of input and output values as evidence of the running of an algorithm – that we want. The problem of arbitrariness of interpretation is not resolved.

Objection 6: For us to say that something exists, it should be apparent in the real world.

Answer

This is really the previous objection, above, worded slightly differently, and the answer previously given deals with it.

Objection 7: It is not true to say that there is no obvious cut-off point between valid and invalid interpretations. You said that there is no sense in which any observations can be made directly, but there is one sense in which an interpretation can be direct: by not involving any processing outside the human brain. Some interpretations can be made by an unassisted human, but other, more extreme, interpretations are too complex for an unassisted human brain to make, and humans can only find algorithms with such interpretations by using a computer to run some kind of interpretation algorithm. While we might argue about the semantics of the word “direct” there is a difference between interpretations that we can make unassisted – interpretations in which we can look at the physical system and just see the algorithm – and extreme, contrived interpretations in which the algorithm only appears unless when we use a computer to run some complex program.

Answer

For probability to be meaningful in the sorts of situations that have been considered there must be an unambiguous, formal way of deciding what a valid interpretation is. The objection does not provide this, instead suggesting a method based on the arbitrary standard of human biology. We may be able to immediately “see” that an algorithm is running in some situations, but what if human brains were much more capable? We may then be able to examine a physical system and “see” that some interpretation can be made corresponding to finding an algorithm in a way that is unknown to us with our more limited brains. We do not even need to imagine brains vastly different to our own: variation between existing human brains would mean that some human brains can make interpretations that others cannot. In fact, we do not need to consider different brains really: one is enough. The objection does not deal with the problem that varying amounts of thought may be needed to make an interpretation. Someone may examine a physical system and immediately “see” that it is running an algorithm. He/she may think about the system for a few minutes before “seeing” that another interpretation of the system “reveals” another algorithm and may think for six months before “seeing” a further algorithm. Where do you draw the line? If you can solve that problem, you then have the issues of how intelligent the person is, how much sleep he/she had the night before, how motivated he/she is to “see” algorithms, whether or not he/she has used any drugs that compromise thinking abilities, etc.

All these are just symptoms of the main problem, which is that interpretations can be made by different algorithms, and there is nothing profound about those algorithms that happen to be accessible to an unaided human brain.

Objection 8: You are saying that there is no way of saying which interpretations are correct, but you cannot know that. Some unknown physics could make particular interpretations valid. Something, going on in matter, could explain why some interpretations are correct and not others and this could coincide with our intuitive ideas of validity of interpretations.

Answer

The strong AI hypothesis, as generally understood, does not claim that some “unknown physics” is responsible for the process of physical systems acquiring meaning and giving rise to algorithms and computations. Instead, the occurrence of certain physical processes is supposed to be associated with computation and *that is it*. Advocates of such abstraction cannot reasonably seek refuge in unknown physics.

The very idea of specific physics being involved would suggest that the type of physical system was relevant, as the underlying physics may be different in different physical systems. This would suggest some sort of substrate dependence and that some view other than that in the strong AI hypothesis, as most people understand it, would be needed.

There is also the issue of why this physics should coincide with our intuitive ideas about the validity of interpretations.

Objection 9: It is impossible to define how the algorithm detector would work because there is no clear concept of what would constitute a “physical variable” to be measured. A physical variable could only be some variable in fundamental physics if we are sure there is a fundamental physics and if we know what it is. If we allow a physical variable to be something in “higher level” physics then we have the problem that this physical variable could itself be the result of interpretation being applied to lower level physical variables. For example, if physical probes measure the local gas pressure, this gas pressure is the result of an averaging over many different molecules in the gas. An interpretation is therefore being applied to the lower level variables associated with the molecules in the gas to obtain the pressure value. This associates the physical probes themselves with a particular interpretation. If this interpretation is valid, why not others? We could extract any physical values we wanted just by specifying a suitable physical probe design that made a sufficiently contrived measurement and it would be impossible to say what a “correct” physical probe is. This makes the algorithm detector itself incoherent, so we need not worry about what would be valid programming for it.

Answer

It could be argued that this problem could be solved by requiring physical probes to measure values that are part of the most fundamental physics currently known. The idea would be that any problems due to interpretation actually being built into the design of the probe are presumably minimized by doing this. We need not worry about whether or not construction of such a machine is practical: the algorithm detector and related machines are just conceptual devices to help us think about philosophical problems. Against this it could be argued that we may not know what that bottom level of physics is, or there may not even be one, and even our best attempt to do this could involve us in some interpretation, just as if we were using physical probes that read abstracted, macroscopic variables. We could reply that the sorts of interpretations used to obtain such variables would be the particular ones associated with physics, which do not seem to have much to do with finding algorithms, but we could be asked to prove this and also to explain what is special about these interpretations that makes them valid. Continuing further with such an argument would get us into the subject of the next objection, so I will leave that here. Instead, I would point out that, if we decide that the specification of the probes requires an arbitrary interpretation, this does not cause my argument problems. It may cause problems in building a useful algorithm detector, but that was never my objective. The objective was to show that if we had an algorithm detector, there would be arbitrariness in the interpretations that we had it make, that arbitrariness being in selection of its software. If it can be reasonably argued that some arbitrariness must be built into the design of the machine itself – if there is arbitrariness not only in the machine’s software, but in the choice of physical probe to be used – then that merely adds to the main point: that there must be arbitrariness in the interpretation performed by such

a device and that, as the device represents the process of interpretation itself, the concept of particular algorithms being run by hardware rests on arbitrary interpretation.

Objection 10: What you say about algorithms, you could as easily say about physical objects. For example, the concepts of “cat” or “chair” could be expressed as interpretive algorithms that allow us to decide if these objects are there. The problem is that we might imagine making a very complex interpretation that allows us to find a cat or chair, when none is apparent in the real world. Is this supposed to mean that such objects exist in some sense? As this is absurd, and as the same sort of logic leads to this, we should discount this.

Answer

I will be discussing this issue in a later article. For now, I point out that it does not remove the main problem raised by this discussion. There is nothing in this objection that somehow finds a way of declaring some interpretations valid and others valid. The observation that there may be a similar issue when dealing with physical reality in general may be uncomfortable, but it will not make the problem go away. What I have said so far will have given an idea of what, to some, will be the fairly appalling philosophical step that I am prepared to take to resolve this.

Readers of Greg Egan’s books [12] might like it though: I will discuss that later.

Searle’s objection does not go far enough

I find myself in a strange position, when I am going to disagree with Searle, of saying that one of his objections to the strong AI hypothesis does not go far enough. Nevertheless, this is the case with his “multiple realizability” objection. Here, again is one of the quotes of Searle that I gave:

“...the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movement that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Wordstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain.” [7]

Searle seems to view the idea of “interpreting” a physical system to obtain computational states out of it a bit like making words with alphabet soup: you can put things together in different arrangements and orders to make anything you want. As I have described in this article, however, the concept of interpretation is more general than this. An interpretation is not limited to being a way of “sticking pieces of reality together”: an interpretation is any formally describable way of obtaining the algorithm from reality and is any algorithm which can be applied to reality (or more accurately used to control a machine like an algorithm detector, which is applying the interpretation to reality). This means that there is not even the requirement that the wall should be “big enough”. A small piece of wall can also be said to be implementing any algorithm if we like, if we are prepared to use a sufficiently complex interpretation. To take this to an extreme, we could even apply any

interpretation we wanted to a single electron in an atom in the wall to obtain any algorithm we wanted. This may seem counter-intuitive, and some readers will be thinking that it is obvious in some cases that all the information that we think we are extracting from the wall is just coming from the interpretation itself. As I have pointed out earlier, the problem is that there is no obvious cut-off point at which interpretations suddenly become “cheating” like this.

Actually, we might go *further* than saying that any interpretation we want could be made from a single electron. Do we even need the electron? That is something for later articles.

I may seem here to have been trying to strengthen Searle’s case against the strong AI hypothesis, but that is not a problem for me. I have no problem with the basis of Searle’s “multiple realizability” objection and I am using it myself in this article. It is better for me if it is as strong as possible and I certainly want the version of it that I use here to be the strongest one that can be justified. The difference between my position and Searle’s will be in how this objection is *answered*.

Searle’s Answer to the Problem of Arbitrariness of Interpretation

I will discuss Searle’s answer to this problem now, but I want to emphasize that I disagree it.

What I call “arbitrariness of interpretation” Searle calls “multiple realizability”. We may both use slightly different arguments to try to show that the problem exists, but we appear to agree that the problem exists. Searle’s answer to it is definite: he states that the strong AI hypothesis must be incoherent, but not in a way that means it needs more clarification. He thinks it is *completely* incoherent to associate a mind with the running of an algorithm. Instead, Searle thinks that such things as “thought”, “minds”, “understanding”, “intention” and “consciousness” must be associated with specific physical processes and that other processes do not cause these things just by appearing to be processing information in the same way. As far as Searle is concerned, only some processes have the “causal power” needed for minds.

Searle does not say *which* physical processes have the right causal power: he admits that he does not know. He does say that (obviously) the human brain has the relevant causal power and that there is no reason to presume that a computer would have it, just because it was performing what looked like the correct computation. This does not mean that Searle thinks that intelligent *behaviour* by a computer is unlikely. Rather, he would interpret such behaviour differently. He would say that processing of symbols counts for nothing and a computer which appeared to be conscious would probably just be a zombie-like simulation of consciousness – something with merely the appearance of a mind.

What this position implies is particularly obvious when we consider mind uploading. The strong AI hypothesis would indicate that if we could scan a human brain with sufficient

accuracy to make a good enough computer simulation of the physical processes occurring in that brain, then the computer simulation would itself be conscious. Searle would say that there is no reason to presume this, because what makes us conscious is the particular physical processes occurring in brains and, while a computer simulation may capture the behaviour, it does not capture the actual physical processes themselves. This would apply, as far as Searle is concerned, if the mind uploaded “copy” were pleading for survival, telling jokes and engaging in philosophy. It would apply even if its behaviour could not be distinguished from the original biological person. As far as Searle is concerned, behaviour, algorithms and processing of symbols count for nothing while the physical nature of the system counts for everything.

My Answer to the Problem of Arbitrariness of Interpretation

I think Searle is wrong. I propose a solution that does not discard the strong AI hypothesis, but instead clarifies it. What I am proposing is this:

We should admit all interpretations equally.

In other words, if a correctly functioning algorithm detector can be made to report the existence of an algorithm then the algorithm should be considered to be there – no matter what sort of detection program was needed or how extreme it was. This applies whether we are detecting the algorithm itself or using one of the other devices that detects inputs and outputs and inferring the existence of the algorithm from these: there is no profound difference.

This may appear philosophically extreme: it *should* appear philosophically extreme, at least with regard to conforming to our intuitive expectations. If true, it means that a chair or a wall, can be said to be running any algorithm imaginable, because any conceivable algorithm could be “found” by interpreting observations made of it in the right way. This “many-interpretations” approach completely deals with the problem of arbitrariness of interpretation, or “multiple realizability” because the arbitrariness is in the selection of the interpretation that is going to be correct. If all interpretations are valid then there is no selection, and no arbitrariness.

The Extra-Strong AI Hypothesis

This is not an attack on the strong AI hypothesis: clarification of it is not a weakening of it. In fact, I would say that what I am proposing here is an *extreme* form of the strong AI hypothesis – one in which the idea that minds are associated with symbol processing is taken so literally that any interpretation of a system which gives the correct symbol processing is recognized as corresponding to a real mind.

It may seem strange to some readers to regard this as “extra-strong” AI, but consider Searle’s view: Searle says that no computational interpretation should be viewed as

corresponding to a mind. *That* is most certainly an attack on the strong AI hypothesis. I am going to the other extreme and admitting not only the particular computational interpretation of a system that strong AI advocates would insist that I should admit as a mind, but all the other interpretations that describe things like minds as well. I am about as far away from Searle's view as I could be.

It may appear, from this, that I am explicitly assuming that minds must be computational. This is not actually the case: things are more involved than that. Some readers will wonder how that can be: after all I am associating minds with the correct algorithms. I will explain later in this article. For now I point out that I said that minds should be viewed as *associated* with the correct algorithms. This is not the same as giving the algorithms themselves some sort of metaphysical power, and saying that "an algorithm is running" is really an informal way of saying something about reference classes and minds: I will discuss this later in this article.

Many-Interpretations and Measure

Why the Physical Substrate Matters

Considering different interpretations is equivalent to considering different *detection programs* that could be used in an algorithm detector machine, or similar device, as described previously. I have said that for any physical system we should admit all interpretations – all detection programs – as corresponding to algorithms that are based on the system. This may *seem* to make the physical nature of a system irrelevant. After all, if a physical system is running all algorithms, then why should it be of any importance at all if, for example, it contains electronic circuits which under one interpretation could be said to be running the *WordStar* program? This may seem to push us back into incoherence. The thought experiments in the earlier articles [1,2,3,4] suggested that if multiple systems are running something like your mind then the substrate is statistically important, but does this not suggest that the substrate and anything else is irrelevant? It may seem that this implies that your brain can be said to be causing your mind no more or less than a chair could be said to be causing it. According to such a criticism, it would be just as likely that your experiences, including all your perceptions of the external world, are being caused by a chair or table than your by your brain, which would make it dubious that you were even in your own brain! This would seem to over-rule everything said in the earlier articles about substrate and measure. *I want to be clear that I am not advocating such an idea.*

How should we resolve this issue? *Relative measure*, a concept from the previous articles, provides the answer. Just because any system can be said to have interpretations in which it runs all algorithms, and just because we treat each of these interpretations equally, it does not follow that all algorithms that can be produced by such interpretations are equal. In any given physical system, some algorithms – the ones we would normally associate with "common-sense" interpretations – will run with particularly high measure, relative to that of others, and the measure with which a physical system runs a particular

algorithm, relative to the measure with which it runs some other algorithm, will depend on the physical nature of the system and the two algorithms.

There are two reasons for this:

- Some interpretations will produce the same algorithm, so the measure of some algorithms is higher, relative to that of others, because more interpretations in the set of all possible interpretations produce them.
- Some interpretations will produce algorithms that are very similar, if not identical, so that the measure of some *classification*, or general type, of algorithms will be higher, relative to that of others.

To see why the physical nature of the system should matter we can use the following argument:

Consider some physical system, P . We are applying an algorithm detector to P .

Suppose we have some detection program I_1 to put in the algorithm detector. I_1 has length L_1 bits and when applied to P applies a particular interpretation of what P “means” and produces algorithm A .

Now, I suggest that the following statement is true:

If algorithm I_1 has L_1 bits and, in an algorithm detector, produces an algorithm A by interpreting P , if we are considering all detection programs with lengths up to some maximum lengths of L bits, then the number of detection programs which, when used in the algorithm detector, also produce A is largest when we can find an algorithm I_1 such that L_1 is smallest.

or to put this more simply:

If an algorithm detector can produce some algorithm A by applying a short detection program to some physical system then we should expect many other detection programs also to produce A . If, however, we need a very long detection program to produce some algorithm A when it is applied to the physical system, we should expect interpretations that produce that algorithm to be rare.

A similar argument could be made about algorithms of some classification or general type.

This would explain why “sensible” interpretations should have high measure, relative to that of extreme ones. Considering your brain as an example, the sort of interpretation that produces an algorithmic description of your mind will be a relatively “sensible” interpretation: it would not need a very long program in an algorithm detector or similar device, because the physical system itself is arranged in the right way to facilitate this

sort of interpretation. Such an interpretation will therefore have high measure relative to some of the absurd interpretations that we might imagine.

Infinite Sets and Counting

I have just been talking about counting interpretations, but that is an informal way of viewing things. We cannot really count interpretations or algorithms. There is no absolute number of detection programs that produce A because such a number can only be given for some maximum detection program length of N bits and N is finite in the above reasoning, but unlimited in reality. That is to say, the set of all possible detection programs is infinite, but here we are considering some finite set of detection programs

This is resolved, by using the same sort of technique used in my article *Occam's Razor Part 4: An Overview of How Occam's Razor Works* [18] of letting N tend to infinity. We can make N progressively larger, determining the number of all interpretations that produce an algorithm A₁ and another algorithm A₂, and we can obtain the number of all interpretations that produce A₁ as a *proportion* of those that produce A₂ as N tends to infinity. This can be taken as the measure of A₁, relative to that of A₂.

This is the closest we can get to “counting” instances of an algorithm produced by different interpretations. Although we might informally think about “how many” instances of an algorithm are produced by all the different interpretations, the fact that the set of all interpretations is infinite means that it is only technically correct to consider relative measure. This does not really change what was said above about how the physical substrate affects numbers of algorithms. To be technically correct we just need to talk about measure without reference to absolute numbers of interpretations or algorithms. An algorithm, or some general type of algorithm, is obtained from some physical system, by interpretation, with a high measure, relative to that of some other algorithm or type of algorithm, when the shortest detection program that can be put in an algorithm detector to obtain the algorithm, or an algorithm of that general type, is as short as possible.

This sort of reasoning encompasses the “common-sense” view of particular algorithms being run by physical systems. When we examine a physical system we might easily “see” that it is “obviously” running some algorithm, but that really means that some uncontrived interpretation can produce that algorithm from the physical system. This uncontrived interpretation means that the algorithm runs with high measure, relative to the measure with which other algorithms, associated with more contrived interpretations, could be said to be running.

What the “Other” Interpretations Really Mean

Yes, your brain, or a table or chair might be said to be running the *WordStar* computer program, but the extremely contrived interpretation needed to obtain such an algorithm from these systems means that the measure will be extremely small relative to that with which other algorithms resulting from more “sensible” interpretations are being run by

various systems. On the other hand, the physical nature of your brain makes it simple to apply an interpretation that produces an algorithm corresponding to your mind, so the measure of that kind of “sensible” interpretation will be higher relative to that of other, more contrived interpretations. As was discussed in the previous articles [1,2,3,4], relative measure is an indication of probability, so it is much more likely that your experiences are due to the algorithm resulting from the obvious, reasonable interpretation of your brain than it is that your experiences are due to some extreme interpretation of the matter in, for example, a wall.

Many-Interpretations and Weak Substrate Dependence

In the previous two articles in this series [1,2,3,4] an argument was made, with thought experiments, attempting to show that minds exist with different relative measure and that the measure with which a mind exists depends on the physical nature of its substrate – the physical nature of the system running the relevant algorithm. The thought experiments involved situations in which there are multiple physical systems running algorithms that could correspond to your situation, so that there is uncertainty about your situation, and considered the statistical consequences of combining these systems. It was stated that the measure with which a computer runs an algorithm is greater when the redundancy, or inefficiency in use of matter to run the algorithm, is greater. At the time it was stated that this seemed to be true from the thought experiments, but that a deeper understanding was needed to understand how the physical nature of a computer would affect measure. It may seem that we digressed from this with the discussion of arbitrariness of interpretation in this article, but this is actually relevant to the issues discussed earlier. We now have the deeper understanding that I said was needed and are in a position to reconsider the thought experiments in the earlier articles in the context of many-interpretations.

In the thought experiments I took the common-sense view that there is a “preferred” algorithm – the one which seems to be run in a reasonable interpretation of the physical system. For example, in the case of an electronic computer this would be the algorithm that appears to be associated with the arrangement, and states, of the electrical components in the system. Now, in this article I have proposed that instead we should accept multiple interpretations, but we will ignore most of these here. We will instead focus on those interpretations which match “common sense” – which give the “preferred” algorithm that most people would reasonably say is running. As has been stated, this algorithm is run with particularly high relative measure anyway, because of the low amount of information needed to construct such interpretations. If we have two physical systems, and for each of them the same algorithm is the “preferred” one run with very high relative measure, then we can compare the relative measures for this preferred algorithm in each case. Suppose both systems are electronic computers of the same general design, but one has much thicker wires. If we imagine an algorithm detector or similar device moving its physical probes around in this system, then there are many more ways in which it can obtain the sorts of values needed to make an interpretation leading to the “preferred” algorithm than there are in the other system with the thinner wires: there is more matter in which the physical probes can be moved around.

This deals with a number of possible objections to the earlier articles. Some objections would use the idea that the previous proposal that inefficient use of matter in computing causes algorithms to run with higher measure is not a formally expressed idea. Such objections could be that different kinds of splitting could be conceptualized or that “inefficient use of matter” is a vague concept. Objections like this are easily dealt with now by saying that the argument about “redundancy” and “inefficient use of matter” was just an informal version of the argument in this article. To be technically correct, we should not be looking at “redundancy” or “inefficient use of matter”, but at the *measure* with which the algorithm is running. Even selecting a single algorithm as the “preferred” algorithm would be a simplification. It would just be one of an infinite set of algorithms which are all running with some measure. Many-interpretations is the real explanation of weak substrate dependence.

Dealing with Groups of Algorithms

Uncertainty About Yourself

The sort of reasoning here would apply to individual algorithms, but also to algorithms of some similar type. In the thought experiments in the previous two articles [1,2,3,4], it was assumed that you know that various computers are running specific algorithms that correspond to your mind. A problem with this scenario is that you are unlikely to know what algorithm corresponds to your mind. Some readers may have noted that with the previous articles this was not dealt with and, in the thought experiments in those two articles, you know that specific systems are running versions of the algorithm that corresponds to you in virtual realities that may be the same or different.

Dealing with Uncertainty

How do we deal with situations like this if there is uncertainty about what algorithm corresponds to your situation? As an informal simplification, the idea in the first article [1,2], that an algorithm tends to be run with greater measure when inefficient use of hardware is involved, still applies, the real explanation of this being that in such situations the algorithm can be produced by more interpretations (using the word “more” in an informal way, given what was said about no absolute numbers being countable), so we can discuss things in terms of the concepts in the first article, to some degree. If the ideas about measure from the first article are valid for thought experiments involving different, identical versions of you, they will hardly go away when we have different versions of you. Whatever we can say about the measure with which a single version of you is run in some computer will still apply whether it is the same or slightly different as other versions of you in other computers.

Let us consider the kind of situation in the first article [1,2] where there is uncertainty about what substrate you are on. The difference now is that you do not know exactly what algorithm would correspond to your mental experiences. You are confronted with a number of different algorithms, running on different computers, all of which are possible candidates for your mental experiences. Each of these slightly different versions of you is

running in a virtual reality and the virtual realities are not necessarily the same. They could be very different; however they all involve the simulated version of you being in a situation indistinguishable (based on what you know) from your current situation now (even though what happens to you in ten minutes may be a different matter). You know about the physical details of the computers running each such version of you, and about how the software is implemented in each version. In short, you know all of the details about how all a number of non-identical versions of you are implemented in non-identical virtual realities on non-identical hardware.

How do we deal with this? You cannot rule out any of these versions, because they are all close enough to what you know about your current experiences – your own mental state – to be possible candidates for it. Let us further assume that you do not know enough to declare any of these situations a “better fit” for your current situation than any other. In this kind of situation it should be clear that any distinction between different versions of you is irrelevant in terms of how we assign probabilities. It is pointless to talk about differences in versions, with regard to probabilities, if we cannot distinguish between them. What has been said about substrate, measure and interpretation applies to each individual algorithm that could be found, by interpretation, in any of these systems, and because we cannot distinguish between them we can treat them all as if they are identical versions of you, on different substrates and in possibly different virtual realities, with different futures. It is not necessary to have a collection of identical versions of you to assign probabilities.

Most of the measure will come from “obvious” interpretations in the various computing systems on which you could be running, so we should not need to deal with the contrived versions, found by longer detection programs, to assign probabilities. We should be able to deal just with those interpretations that find algorithms with measure that is high, relative to that of algorithms found by other interpretations. These interpretations will have high measure because of the nature of the physical substrate. For example, if the physical substrate is a human brain, then the “obvious” interpretations will be those that pull things like a specific human mind out of the physical system because less information will be needed to do this. Many such reasonable interpretations will be possible, for example by obtaining physical variables from slightly different points in a neuron. There is not just one reasonable interpretation. There will be many-interpretations that can be made by relatively short detection programs and which are *almost* the same. In practice, however, we would not need to worry about this when assigning probabilities to different situations as in the thought experiments that have been considered. By looking at a human brain, we could determine what the most “obvious” algorithm resulting from interpretation is. We would not need to work out all the individual interpretations which result in such an algorithm, or minor variations of them. Instead, we could say that a relatively high measure group of interpretations that are compatible with our experiences is implied by this single “obvious” interpretation. We could then use our knowledge about the physical structure of the brain to get an idea what the “measure” of algorithms in such a group is, relative to measures obtained in a similar way in other physical systems.

Two Separate Issues

There are two separate issues relevant to the many-interpretations view and it is easy to mix them up:

1. the issue of which interpretations should be considered valid for saying that a given algorithm is running, or that some computation is proceeding through given computational states in some physical system.
2. the issue of what algorithms, computational states or sequences of computational states should be regarded as corresponding to minds.

These are separate issues. To state whether or not a given physical system produces a mind we must be able to deal with both issues. This series of articles will mainly deal with the first issue. I have not yet decided whether or not any later articles in this series will deal with the second issue at all.

Both of these can be considered as issues of possible sets which determine the set of possible situations – the reference class – for an observer. This series of articles has considered thought experiments in which you are unsure about which substrate you are on and when considering this kind of uncertainty about your situation, and the chances of being in different situations, we are really considering things in terms of reference class. In probability calculations the reference class is the set of possibilities from which something is chosen and in this sort of consideration the reference class is the set of all possible situations in which you could be existing. A consideration of the above two issues, in reference class terms, is be as follows:

The first issue, the main subject of this article, is the issue of what formally expressed interpretations, or descriptions of your situation based on external reality, should be considered as forming the set of possible candidates for your situation. I will call this the *interpretation set*.

The second issue is the issue of what sort of information structure, algorithm or computational state, produced or found by an interpretation, should be considered a member of the set of candidates for your mental state. I will call this set the *abstraction set*.

Each member of the abstraction set has its own interpretation set of formally describable interpretations of the physical world that relate the physical world to that mind/mental state. This means that the situation of any thinking being is described by a pair of items chosen from the two sets: the abstract description of the mental experience comes from the abstraction set and the interpretation that relates these experiences to the physical world comes from the interpretation set corresponding to that particular abstract description. Each member of the reference class of possible situations in which a thinking being could exist is therefore an abstract, formal description of a mental state from the abstraction set and a formal description of how that mental state relates to the physical

world (that is to say, a formally expressed interpretation of the physical world) from the interpretation set for that mental state.

It should be noted that mental states do not all occur with the same frequency in the reference class of possible situations. A mental state with more (using that word informally) interpretations in its corresponding interpretation set will feature in more situations in the reference class of possible situations, because every possible combination of a mental state and an interpretation that relates the physical world to that mental state is a member of the reference class of possible situations.

Here is an example of how the abstraction and interpretation sets would determine the reference class of possible situations. Rather than consider infinite, or even very large sets, let us suppose that we have small sets:

There is an abstract, formally described mental state, M_1 . M_1 is related to physical reality by formally described interpretations I_1 , I_2 and I_3 .

There is a mental state M_2 , related to physical reality by formally described interpretations I_4 and I_5 .

There is a mental state M_3 , related to physical reality by formally described interpretation I_6 .

The reference class of possible situations involves every possible mental state and interpretation pair as follows:

$M_1I_1, M_1I_2, M_1I_3, M_2I_4, M_2I_5, M_3I_6$.

In this example, different mental states have different numbers of interpretations. Some readers might ask how anything can be counted, given the true infinite sets, but this is dealt with by considering descriptions with lengths that tend to, but do not reach, infinity.

This may be unwieldy to handle statistically, or even to think about, so I can make one suggestion about how it could be simplified for convenience in situations in which we are much more interested in the different possible interpretations than in the different possible mental states: we could just assume that one plausible mental state (that is to say, a mental state from the abstraction set) is the mental state (whether it is or not) and ignore the rest of the abstraction set, considering just the interpretation set corresponding to that mental state.

In the example just given, if you selected M_1 as your mental state then your reference class of possible situations would be:

M_1I_1, M_1I_2, M_1I_3

This is likely to be an effective simplification technique, because in most cases it will probably be hard to determine any significant difference between the interpretation sets for different mental states in your abstraction set.

Whether these issues are viewed in terms of the descriptions that I initially gave or in reference class terms is partly a matter of convenience, but the second approach has an advantage. A problem of the first approach is that it is philosophically abstract. Knowing what position to take on these issues would mean knowing what a mind is, and what is needed to make one exist, but it is hard to see what empirical questions would be answered by possession of such information and an argument might therefore be made that no questions would be answered. Alternatively, considering these issues in reference class terms means that the probabilities of you being in various situations that you obtain in thought experiments like those discussed earlier [1,2,3,4] are dependent on the approach that you take: considering things in terms of reference class actually has ramifications for your possible situation. This makes the reference class view preferable. A similar reference class view can be taken of the strong AI hypothesis itself, as will be discussed later.

The second issue – that of the abstraction set – is about what formal descriptions of your mind are valid. To some readers, this may seem nonsensical: they may point out that there is only one valid description of your mind – the one that happens to be correct. We do not know, however, what this description is. In fact, your knowledge about your own current state of mind is limited, long before we get as far as computational states or reference classes. What does happiness feel like? Do you know *exactly* how happy you are now? If you are about to do something can you say *exactly* how strong that motivation is? If you had to make a *formal* description of your current state of mind, assuming that you could attempt such a thing (and I am not pretending that I could), limitations would be imposed by your lack of knowledge about own mental state. At best, all you could do is state which kinds of formal descriptions are valid and which kinds are invalid. These valid descriptions may vary in terms of the detail of how they describe your emotions and motivations, your perception of something that they found beautiful, your opinions of this article, and so on, and limitations in your own self-knowledge would mean that you have a lot of valid descriptions. Every possible, formally describable situation that you could be in involves one of these formal descriptions of your mental state and a formal description of how that mental state relates to reality (the interpretation). The “correct”, formal description of your situation would be one of these, but you would not know which it is after you have already removed all those which contain invalid mental states or interpretations. Assuming that all valid formal descriptions of your situation are equally likely (we could introduce complications by getting different probabilities involved for different formal descriptions here, but I will not do that), the “correct” formal description of your situation is randomly chosen from one of these, so the set of possible formal descriptions of your situation is a reference class. A reference class of possible, valid formal descriptions of your situation results then from limitations in self-knowledge as well as uncertainty about the correct interpretation.

Some of this self-knowledge may come from limitations of human biology. We might imagine, for example, a machine which can know exactly how happy it is by running some software subroutine which tells it. What if a machine could make similar enquiries about every aspect of its functioning? Could it “know” a formal description of its own computational state and avoid the need for a reference class at all? I think many readers will intuitively suspect that some kinds of limitations about self-knowledge would apply in any situation like this and, indeed, there is reason to think that this is the case and that no thinking entity would be able to avoid the need for a reference class of possible situations involving various formal descriptions of its mental state to deal with limitations in its own self-knowledge. This issue of limitations on self-knowledge has been discussed by Hofstadter [19].

Is this a computational view of the mind?

Ways in which it is and ways in which it is not

The many-interpretations view may appear obviously computational, but things are not that simple. The answer to this question is “yes” and “no” depending on what the term “a computational view of mind” means.

First, let us consider what *is* computational about the view:

Any mental state is computational to the extent that an algorithm is used to describe it. An observer who is unsure about his/her/its mental state will have a set of possible mental states, each such mental state being expressed as an algorithm.

Any interpretation that means a mind exists, based on physical reality, is computational to the extent that an algorithm is used to describe that interpretation. Any observer who is unsure about his/her/its situation in reality (for example, on what substrate he/she/it really exists) will have a set of possible interpretations, each such interpretation being expressed as an algorithm, implying that it is a computational interpretation.

There is another, obvious way in which the many-interpretations view could be regarded as computational: it allows computers to have minds.

This may appear very computational, so let us consider what is *not* computational about it:

Although the interpretations and the mental experiences/mental states in the reference class of possible situations are described computationally, this is not the same as stating that the mind results from computation. The many-interpretations position is merely assuming that things in the physical world must relate to the state of any mind that exists, and that this relationship and the mental state must have a formal description. Algorithms are used for that formal description, because of the apparent generality of algorithms, and the many-interpretations view assumes that any interpreting algorithm that can describe

the relationship between a mental state and variables in the physical world is associated with a mind.

Does this mean that the mind is being assumed to be “computational” or “an algorithm”? Not really: the same sort of computation could be said to apply to anything else currently recognised by science. For example, let us consider fire engines instead of minds. Any existence of the type of object that we call “fire engine” is based on what exists in a physical system. All of our basic knowledge of the physical system could be described in terms of various physical variables. There would be some relationship between these physical variables and the various features that a fire engine might have, and this physical relationship could be described by an algorithm. There would be an algorithm corresponding to any fire engine, describing how it relates to the underlying physical variables that cause its existence, but would it be reasonable to say that this is equivalent to claiming that fire engines are computational? If we accept these semantics we should recognize that such an idea of “computational” is so general as to allow *anything* to be computational if it happens to have features that relate to reality in a formally describable way. If we decide that this does not make fire engines computational then we similarly have no reason for saying that minds are computational, just because a mental state is formally – algorithmically – describable and has a formally – algorithmically – describable relationship with physical variables.

Although the many-interpretations view allows computers to have minds, this is not a profound feature of it. According to the many-interpretations view, minds would be associated, with varying degrees of measure, with all physical systems, and described computationally in each case. If a computer is doing the right things needed to have a mind it merely means that the physical arrangement of the matter in the computer makes some interpretations have particularly high measure: those that describe how the mental states of that sort of mind are related to physical variables – how the mental states can be produced by *interpretation* of physical variables.

What this all means is that the many-interpretations view has some elements of a computational view of the mind in it, which are just products of a formal view of nature in general and the idea that algorithms are particularly well suited for expressing formal descriptions. Relating a mental state to physical reality by means of an algorithm may *seem* computational but the computation is in the way that the relationship between the mental state and physical reality is expressed rather than the mental state itself.

Some readers might point out that this leaves us with no view of what a mind is. If a mind is associated with the underlying physicality, but is not that, and if it can be described computationally, but is not computation, what is it supposed to be? The answer, and it will not please some people, is that it is a mind, in the same way that a fire engine is a fire engine. We might say that a fire engine’s existence is contingent on some underlying matter, but a fire engine is generally regarded as being a fire engine in its own right.

What this all comes down to is that whether or not this is a computational view of the mind depends on what you mean by “computational view”. What is important, however,

is that there is no *explicit* assumption that “the mind is computational”. The many-interpretations view merely takes the position that the relationship between mental states and the underlying physicality should be capable of formal description.

We can, of course, still talk about how algorithms are “running” in a physical system. In fact, I do so in some parts of this article. We should be aware, however that this is a simplification. Saying that some algorithm is running according to some interpretation is really saying that the interpretation describes a relationship between the physical system and some entity that is formally described by that algorithm.

What this means for the strong AI hypothesis

The strong AI hypothesis is that minds are associated with processing of symbols by an algorithm running in some physical system. Searle says that this idea is rendered incoherent by what he calls “multiple realizability” and I call “arbitrariness of interpretation”, as any symbols and any processing can be obtained with the appropriate interpretation. The many-interpretations position deals with the problem of arbitrariness of interpretation by associating any valid interpretation with a mind and giving consideration to the *measures* of different kinds of minds, which will depend on the physical system. I think the many-interpretations position deals with Searle’s multiple realizability objection against computers having minds, but a question that should be asked is whether the many-interpretations view vindicates the strong AI hypothesis or shows that it was wrong all along.

Someone trying to defend Searle’s position from the many-interpretations view could argue that the many-interpretations view does not support the strong AI hypothesis because the strong AI hypothesis assumes that a particular “computational” interpretation of what a physical system is doing is the “correct” one, while the many-interpretations view denies that any interpretation is the correct one. Furthermore, the many-interpretations view does not really claim that minds are caused by processing of symbols, but rather accepts computation in description of mental experiences and the interpretation with which a mind is associated. Also, the many-interpretations view involves the concept of measure, whereas the strong AI hypothesis does not involve any such consideration.

This objection cannot make Searle’s position correct. Searle clearly does not expect computers to have minds. He may have qualified this by saying that he is not saying that computers do not have minds, but he is clearly sceptical. The many-interpretations view accepts that a computer could have a mind, and in that important respect it is much closer to what the strong AI advocates are saying than what Searle is saying.

Should the strong AI hypothesis be thrown out? I do not think so, but it needs clarifying. If it seems obvious to us that a system is doing the right sort of processing then we might apply the strong AI hypothesis to say that the system has a mind. If, on the other hand, we applied the many-interpretations view to the same system, then (if our initial view that the system was doing the right sort of processing was reasonable) we would be likely to find

that the physical substrate was such that interpretations leading to that sort of mind had a particularly high measure. The reason for this is that any reasonable person claiming that a system is doing a particular kind of computation must have made a relatively simple interpretation of the system, suggesting that a high measure interpretation relating the system to a mind is available. If an advocate of the strong AI hypothesis explained why some system had a mind and an advocate of the many-interpretations position explained why the same system had a mind with particularly high measure, then the two explanations would not be very much different with regard to what they said about the system. The semantics of one explanation would be about computation in the system, while the other explanation would be associating the same sort of computation with a high measure interpretation of the system.

It is obvious that the strong AI hypothesis needs some clarification to agree with the many-interpretations view. The most obvious need for such clarification is that the strong AI hypothesis might tell us that some system has a mind, on account of performing some computation, but the many-interpretations view would say that *any* system has that mind, even though it may be with very low measure. The strong AI hypothesis and the many-interpretations view are therefore not talking about the same things.

From the many-interpretations viewpoint, the strong AI hypothesis is a reasonable, informal statement of what a physical system needs to be doing to have minds with relatively high measure, based on observation of what that system appears to be doing.

The Strong AI Hypothesis in Reference Class Terms

The Reference Class Justification of the Strong AI Hypothesis

I have discussed the two different issues of valid formal descriptions of a mind and valid formal descriptions of interpretations and how these could be viewed in reference class terms. A similar approach can be taken to the strong AI hypothesis itself. The strong AI hypothesis states that minds are associated with the processing of symbols. This is not about the abstraction set. It may seem that accepting the idea of the abstraction set means accepting the strong AI hypothesis, but the abstraction set is just the set of formal descriptions of your mental state that could be valid. Some people like Searle say that specific physical processes are involved – however, a consideration of the many-interpretations view shows this to be a flawed position. People making these sorts of argument treat the idea of “substrate independence” as if it is an assumption that needs some sort of support. I would say, however, that substrate independence (at least weak substrate independence) is arrived at naturally, simply by avoiding prejudice. A justification for this works by turning things around. Instead of looking from outside a mind and asking what would be sufficient for a mind to exist, you look out from your own mind, and ask what would be sufficient to constitute an *explanation* of your mind. The argument is as follows:

1. Any real knowledge that we have should be capable of being formally stated, at least in principle. This should seem a reasonable starting point. If we cannot have this we should ask why we are even trying to do philosophy, and even people like Searle and Penrose have not claimed that the workings of minds are beyond formal description. Theists might.
2. Your mental state – the subjective experiences that you are having now – should be formally describable, at least in principle. This follows from (1).
3. The correct formal description of your mental state is one description from a set of possible descriptions of your mental state. The size of this set depends on how much you know about your mental state. If you have complete knowledge about your mental state then there will only be one member of this set. If you have incomplete knowledge of your mental state then this set will contain many different possible, descriptions. This set is the *abstraction set* to which I referred earlier in this article.
4. Your mental state exists because of something in the “outside world”. By “outside world” all I mean is that part of reality that is not your mental state. Whatever happens in the outside world, there must be some relationship between events in the outside world and your mental state. This relationship is capable of formal description. That is to say, there must be some way of formally describing how events or things in reality “map onto” your mental state – how various characteristics of reality correlate with various characteristics of your mental state. If you know what the nature of reality is, but do not know how events in physical reality correlate with your mental state then an unprejudiced view is one in which, for any given mental state, you consider every such possible formal description of a way in which reality could relate to the formal description of that mental state as being a member of the set of possible, formal descriptions of how reality relates to the description of your mental state. If you have no idea what the nature of physical reality is then you should take a similar sort of position, except that your reference class is now every possible, formal description of how reality relates to the description of your mental state in every possible reality that can be formally described. In either case, each of these formal descriptions of how reality relates to the formal description of your mental state is nothing more or less than an interpretation of the kind we have been considering in this article! The set of such formal descriptions is therefore the *interpretation set* to which I referred earlier. Simply by not making assumptions about how minds relate to physical reality, we end up with every possible interpretation in the interpretation set.
5. If a particular interpretation – a particular formal description of how your mental state relates to the outside world – could conceivably be considered by you to be part of your interpretation set for some possible mental state then, unless you are a solipsist, you should also consider it reasonably *possible* that any similar

interpretations of reality are associated with other minds. This is not dependent on it being *likely* that you could find yourself in such situations: the mere possibility of including them in your reference class means that you have no strong philosophical argument against them. Nor does it mean that you need to be sure that minds are associated with such interpretations: that would be a matter for further philosophical discussion. It just means that all you can know about a particular formal description of how a mental state relates to physical reality is your decision about whether or not you would admit it into your interpretation set for that mental state, and therefore into your reference class of possible situations. If you do admit it then you have every reason to view such situations as involving minds. If you do not admit it then you should have some good philosophical reason for this prejudice in selection of your interpretation set.

6. We can go further than the previous step. It is possible to make an unlimited number of formal descriptions of how your mind could relate to physical reality and, without a strong philosophical argument stating otherwise, you should admit these into your interpretation set for any corresponding, possible mental state. From (5) you should view it as reasonably possible that other such formal descriptions imply the existence of minds. Further, you cannot possess any other evidence that will tell you more about the existence or otherwise of minds associated with such formal descriptions: all you can ever have is knowledge of physical reality and knowledge of what you regard as constituting your reference classes of possible situations.

At this stage, we have ended up with the *many-interpretations* view with regard to the interpretation set for any mental state that could be involved in your own situation and a strong suggestion of it with regard to the existence of other minds. What is important about how we ended up with it is that we did not assume anything like “substrate independence”: instead this followed from the approach of not having any prejudice about how the interpretation set should be constructed. We did not assume a “computational” paradigm. No assumption was made that minds are caused by “the correct computation”. Instead, the only assumptions in the first four steps, about our minds, were that minds can be formally described, that minds are caused to exist by something in reality, that the relationship between this “something” and the mind being caused to exist can be formally described, and that in the absence of any information about which such formal description of the relationship between the physical world and our mental experiences is the correct one we should accept any possible such relationship as forming part of the reference class. Critics could point out that the “interpretation” takes the form of an algorithm, but that is not equivalent to an assumption of minds following from computation: it is merely due to algorithms being a general way of describing things.

Step 5 could be contentious to some people. Once you have accepted that certain types of interpretation could be viewed as reasonable for you, it suggests that you should accept it as reasonable for minds to in such situations. For example, if you think that it is possible for your mental experiences to be due to your mind being simulated on a computer, then

you should accept that it is a reasonable idea that minds are associated with the correct processing on computers. This does not mean that you have to believe, literally, that you are running on a computer: most people probably do not have such scepticism about reality. All it means is that, if it is a conceivable situation then that sort of possibility could be admitted into your reference class.

Although such a viewpoint does not *explicitly* assume the computational view of minds, the computational view follows as a kind of informal simplification of the many-interpretations view. In the many-interpretations view your mind exists because the mental experiences that you are having now are described by a particular interpretation of something in reality. That particular interpretation is computational, in the sense that an algorithm can be used to express it, but this should not be confused with a computational view of minds. A computational paradigm is simply used to describe the relationship between features of your mental state and features of reality. This will have the surface appearance of a computational paradigm, however: physical systems that seem to be doing “the right computations” will be capable of being related to mental states by relatively simple interpretations.

Critics could suggest that some assumption here is being made that events in reality somehow magically cause a mind by “interpretation”. Even without the many-interpretations view (and the sort of argument above seems to lead to it anyway) any description of how minds occur as a result of things in physical reality must involve some description of how things in physical reality relate to the occurrence of particular thoughts in a mind. All we are doing in the above argument is saying that, for this idea to have any meaning, such a description would have to be capable of being formally expressed and that, unless we have some reason to eliminate certain types of descriptions it should be viewed as reasonable that they are associated with minds.

This argument obviously relates particularly to two controversial areas of philosophy: Nick Bostrom’s simulation argument and John Searle’s argument against strong AI.

The Reference Class Argument and the Simulation Hypothesis

Bostrom’s simulation argument [20,21] is worth discussing here, apart from its merits as an argument in its own right, because such discussion will show how the reference class view of strong AI turns things around. The simulation argument explores the possibility that right now you are in a computer simulation being run by future humans, or descendants of humans, or aliens, etc, although it does not seek to show that this is likely: it actually argues something a bit more subtle. This idea of uncertainty about your situation has been considered a lot in this series of articles, in thought experiments. I will not go into Bostrom’s simulation argument in detail here, nor will I get involved in the issue of whether or not it is valid, but instead I will just use it as a convenient philosophical vehicle. What should concern us here is that the argument is based on the logical *possibility* of being in this sort of situation and, according to Bostrom, the simulation argument relies on substrate independence as one of its assumptions. This would mean that, according to Bostrom, someone who accepted substrate independence

should include simulation possibilities in their reference class of possible situations. On the other hand, if you reject substrate independence then the entire simulation argument would appear flawed at the first step. Substrate independence is generally viewed as relying on a computational view of minds, so the simulation argument might seem to rely on such a view.

I do not think that the simulation argument does rely on any assumption of substrate independence, or at least if it does the assumption is so trivial as to be nothing more than the absence of other, more specific assumptions that would require more defence. Let us imagine the simulation scenario applied. It would mean that right now our minds exist because something is happening in a “computer” somewhere in a physical reality that might be similar to what we think of as physical reality or might be utterly alien to us. It might seem that this would have a dramatic effect on our ontological status, making us “computational beings” as opposed to “non-computational beings” but is this really true? On one hand we are here just because of what happens in our brains. On the other hand we are here because of what happens in something called a “computer”. In either case, something in physical reality is related to your mind and that relationship has a formal description. There is nothing profoundly different happening in the simulation scenario. If the simulation scenario is false then you are here because of what the physical parts of your brain are doing. If it is true then you are here because of what the physical parts of a computer are doing. Neither scenario rests on philosophical principles fundamentally different than those of the other scenario.

The only thing that might make the simulation scenario appear “computational” is the computational *intent* of whoever or whatever makes the computer in the simulation. This means that we cannot just declare substrate independence to be an unreasonable assumption and get rid of the entire argument: it is actually the other way round. If you do not know what the particular formal description of your situation is, then you need a reason to start editing your reference class of possible, formal descriptions of your situation – and you cannot just remove the “computational” ones, because there is nothing intrinsically special about them. This means that, depending on how you view the semantics, there is no assumption of substrate independence in the simulation argument, that any such assumption is the default position, that other more specific assumptions would be needed to avoid making it or that it is so trivial as to be not worth stating.

This discussion about the simulation argument may appear similar to the discussion of the reference class view of the strong AI hypothesis in general and it is: consideration of the simulation argument just puts it in a particular context. While the explicit assumption of substrate independence, and the computational view of minds, are not needed by the simulation argument, there is still relationship between them and the simulation argument. By following the sort of argument we have been considering here, it should be apparent that, unless there is reason to the contrary, the simulation scenario should be accepted as a logical possibility, merely because it cannot be completely eliminated based on what we know, and that such a possibility does not need any explicit idea of “substrate independence” or any “computational” view of minds. However, once you accept the simulation argument as a logical possibility, the possibility of existing on other substrates

that are radically different to what you might seem to exist on is admitted into your reference class of possible situations (that is to say, into the interpretation set for any given mental state, which is used to construct this reference class). This would be a strong reason for accepting substrate independence and the possibility that computers could be associated with minds.

This reverses the relationship between the simulation argument and the strong AI hypothesis. Instead of relying on the strong AI hypothesis, simulation scenarios can be shown to be something that we might reasonably accept as logically possible without even considering such a hypothesis, yet once we view such scenarios as logically possible, we have no justification for stating that substrate independence is invalid or claiming that the strong AI hypothesis (or at least the qualified/clarified version of it considered in this article) is invalid. *The simulation argument, far from being dependent on substrate independence, actually supports it.*

A further issue for the simulation argument, which I will not go into here, is that the many-interpretations view associates measure with minds and this could skew the statistics being considered in the simulation argument: when we are trying to determine where most observer moments are experienced, an instant in a computer simulation of a person, or an instant in a real brain should not be regarded as one observer moment each: the physical nature of these systems can affect things. This will be discussed in more detail in a later article.

The Reference Class Argument and Searle's Position

The argument supporting the reference class view of the strong AI hypothesis opposes Searle's arguments against strong AI [5,6,7,10]. In fact, arguing against Searle is the most obvious use it. Searle argues that strong AI is based on a computational view of minds and that there are no grounds for such a view. He argues that, instead, specific biological processes must account for consciousness. While he does not rule out the possibility that specific processes in a computer could be associated with the existence of a mind, he implies that it is not a possibility that should be taken very seriously. Much of Searle's reasoning attempts to show the incoherence and incorrectness of a "computational" view of the mind. The problem is that, according to the view taken in this article, no such position need be explicitly assumed. Accepting the idea of a computer thinking as a reasonable possibility need not be based on any "computational" view of minds, but simply on acceptance of the idea that, in some situations, we might include a similar kind of description of how our own mental states relate to the physical world in our reference class.

An advocate of Searle's position might protest that consciousness is caused by specific physical processes rather than computation, but whatever the specific biology is that Searle might think is needed to cause minds, the relationship between the matter that is taking part in that specific biology would still be something that we can formally describe, in principle, unless we resort to some supernatural idea that we are based on "something" that is "beyond science" or "beyond logic", whatever such terms are

supposed to mean – and Searle says that he is not claiming the supernatural. Let us imagine, for a moment, that Searle is right, and that some specific biological process (and possibly some other specific processes) are needed to produce minds. The biological process behind your mind could be formally described. This means that if we consider your mental state and the physical world then there is some formal description of how various characteristics of your mental state relate to various characteristics of the physical world. If no description at all were possible then the reason for your mental state would be a mystery and if the description could not be formal then we would be admitting something beyond the scope of what is, after all, merely unambiguous language: we would be in the realm of the supernatural. Whatever that process is, the description of how things in the physical world relate to a given mental state *must* be one of the interpretations in an interpretation set for that mental state – and must therefore contribute to your reference class of possible experiences. An advocate of Searle’s position cannot plausibly deny the existence of the abstraction set and the interpretation sets because as soon as the possibility, in principal, of formally describing a mental state is admitted and the possibility, in principal, of formally describing just one way in which the physical world relates to your mental state is admitted – the “special” (and unknown) way in which minds are supposed to work according to Searle is admitted, the existence of members of the abstraction set and interpretation sets have been admitted, apparently making them inescapable and establishing a reference class of formally describable situations.

The objection that minds are not computers does not even have relevance at this point. There is nothing in acceptance of the abstraction and interpretation sets that implies an explicitly computational paradigm of minds, even though the members of the interpretation set are likely to be expressed as algorithms. This may seem a contradiction, but expressing a possible relationship between a possible mental state and physical reality as an algorithm does not say anything about what mental experiences are (if that question can even be answered) or what is required to have them. It simply amounts to formally describing how characteristics of the physical world relate to mental states. Algorithms are used not due to any assumptions of computation in minds, but merely because they are suitable for making *general* formal descriptions.

This weakens Searle’s case weakened because it removes the relevance of the objection that Searle makes against computational views of minds by not demanding them. Nor could an advocate of Searle’s position validly say that it also negates the strong AI hypothesis: the correct “symbol processing” by a physical system would still be a quite valid concept in terms of determining that a mind was associated with it with some degree of measure: the concept of symbol processing would just need clarifying so that it can be understood as an interpretation of the physical system. At worst, strong AI would be defined somewhat vaguely. Ultimately – and this is the main issue with regard to whether strong AI or Searle comes off worse – there would be every reason to associate the correct sort of behaviour in a computer with a mind.

An advocate of Searle's position could try the following sort of reply:

“So what? I accept that there is a reference class of formally described situations, each consisting of a formally described mental state and a formally described – indeed algorithmically described – relationship between the physical world and the mental state. This does not mean that all such conceivable interpretations – all ways of describing some relationship between the physical world and a mental state like mine – actually relate to minds. We know that we are conscious as humans and it is sensible to assume that only situations involving *that* sort of process – or that sort of interpretation if you want to stick an algorithmic description on it – are in the reference class.”

In other words, someone might reply to what I have said by accepting it and then simply restricting the interpretation set for any given mental state, and therefore the reference class of situations, to rule out those unwelcome “computers”.

Such an approach would be giving up a lot of ground: Searle's arguments about the incoherence of computational views of the mind would no longer be relevant and it would even have been accepted that, if Searle's “special” sort of process exists, even *that* would involve a computationally described relationship between physical reality and a mental state. All that would be left is what should by now be appearing to be a rather arbitrary downsizing of the interpretation set for any given mental state and, therefore, of the reference class of possible situations.

This diminishment of the reference class encounters problems beyond its mere arbitrariness. An advocate of Searle's position may choose to admit only those interpretations that we would associate with the sorts of processes on which we are based, but can you be so sure that you know on what processes you are based? Unless you can actually show simulation scenarios to be impossible, you would have to admit those into the reference class. Furthermore, even if each of us is here in the same “sort of” way with regard to what atoms are doing, we have no guarantee that any sort of similarity is maintained at deeper levels of reality. Trying to limit interpretation sets to exclude everything except the “sensible” biological interpretations amounts to excluding possible interpretations on which you could be based! Nor could you validly dismiss this by saying that simulation scenarios are based on an assumption of substrate independence: I showed, above, that the simulation argument rests on no such assumption and that things are actually the other way round.

I am not saying that simulation scenarios are likely to apply. The point being made is a bit deeper. The logical possibility of simulation scenarios simply brings into focus a more general issue about your lack of knowledge about the exact formal description of the relationship between your mental state and physical reality, making attempts to limit the reference class of such descriptions questionable.

That is all I will say about Searle's position for now. I will briefly comment on it again, in more general terms, later in this article. Really, though, there is too much to be dealt with in an article about other matters and it probably needs an article devoted to it.

Deeper Issues: Constantine's Question

Books, Minds and Motion

In Tony Ballantyne's novel *Divergence* [9], a character, Constantine, wonders what sort of substrate is needed to support a mind and whether or not a mind could exist just by its thoughts being described in a book. I used a quote from *Divergence* at the start of the article, and here is an expanded version of it:

“Could I write a mind? Could I sit down and describe a scene, a thought and an emotion so well that it took life on the page? No, the page is not a suitable medium to allow movement, and this language is too ambiguous and overblown to capture the simplicity of the underlying mechanism of thought. Constantine had once been told that a mind was a sentence that could read itself. A book might have thoughts written within it, but something external had to be applied to the book in order to read the words. But what if words could be written in some medium that allowed the words to take on a life of their own and refer back to themselves? What if the instructions telling the book how to read itself were also written in the book itself?” [9].

This sort of speculation touches on the essence of this series of articles. It may seem ridiculous to some people to talk about a mind existing by being described in a book, but why should it necessarily be so absurd? Our minds exist by virtue of being described in our brains. If we really think that the strong AI hypothesis is true then our minds exist because *our brains tell the right story to make them exist*. Similarly, if we think that computers can have minds by doing the correct symbol processing (though this is an informal way of expressing things) then, really, we are not requiring computers to do anything more than similarly tell the right story, in enough detail. This is a sort of answer to Constantine's question. A story in a book *can* cause a mind to exist, if we use a very flexible definition of the word “book” which regards brains as very wet books and computers as very technological books. Of course, Constantine is asking somewhat more than this: he is wondering if a literal book – ink on paper – containing an adequate description of a mind's thinking could act as a substrate for a mind.

As well as describing the algorithm detector, I also described interactive devices related to the algorithm detector. These sorts of devices are not only capable of finding patterns in physical variables and interpreting them as corresponding to particular algorithms, but also of altering physical variables so as to alter the inputs of those algorithms, allowing interaction with them. I stated that not all algorithms that could be found could be interacted with in this way by humans. For a human to interact with an algorithm he/she would need to influence its computational states one after the other, in time, but the algorithm's computational states may not correspond to the physical system at separate moments of time which follow on one after the other, or it may be that the interpretation that we are considering really involves assuming particular inputs and outputs. I raised the question of whether or not we should demand the possibility of interactivity to say that an algorithm is running.

Let us now relate this to things printed on paper, instead of running on computers or in brains.

Suppose we have a hardcopy printout of the sequence of inputs and outputs through which some computer program goes. The right sort of algorithm detector or similar device, applied to this printout, could read the pattern of ink and use it to extract evidence of the algorithm and the computational states through which it goes. Even though no algorithm appears to be “running” as most people would understand it, the computational states can still be extracted from the paper by an interpretation that would not be very contrived. The detector could use some interpretation that extracts the first computational state from the paper at some time, then the next at a later time, then the next, and so on. To detect longer runs of the algorithm the interpreter simply analyzes a longer printout for longer. Indeed, the only thing special about this sort of interpretation of the printout is its relatively high measure. Qualitatively, a printout of the algorithm can be said to be running it.

There will be other extreme, contrived interpretations that pull any algorithm we can imagine out of the hardcopy printout. The one that we see in the pattern of ink on the page just happens to be the result of a particularly high measure interpretation because the way in which the ink has been put on the page statistically favours it.

Interaction

Suppose we require all interpretations to allow interaction in some way – that only those interpretations made by *interactive* algorithm detectors or similar devices are admitted. We would have no problem making such interpretations from things like brains or conventional computers. Because the computational states are arranged neatly in time, one after the other, we should be able to imagine an interactive algorithm detector or input and output reporter reading the relevant information from the matter to obtain a computational state and then obtaining the inputs that we want to provide and using them to affect the matter in the system and alter the current computational state by providing it with inputs. This can be done because the ordering of computational states in time allows the detection device to set up the inputs for a given computational state before it happens. People do this with us all the time when they talk to us.

This could be done for an algorithm described in a book. Even if the book does not list any inputs and outputs, but only lists an algorithm, then an appropriate algorithm detector could still read the pattern of ink and apply some interpretation to report that it has found an algorithm running with a particular sequence of inputs and outputs. It could allow interaction with this algorithm by just running the program that it has found itself and not making any changes to the physical system to provide it with your inputs. In other words, the arbitrariness of interpretation problem does not just allow an algorithm detector to choose arbitrarily what algorithms it has found: it also enables it to choose arbitrarily what inputs and outputs are occurring in that algorithm. Alternatively, if you think that the algorithm detector does need to interact with the system properly then that can be arranged too. Imagine a printout of a computer program to take numbers as inputs and do

arithmetic on a piece of paper. An algorithm detector could read that information from the paper and also make contrived interpretations of the states of the matter in the paper at future instants in time to correspond to computational states of the system running this program, occurring one after the other in time – and it could allow you to interact with them.

What this means is that even if you demand interactivity with a system it can be provided, even when we are looking at a computer program, or a description of a mind printed on paper, provided that you are prepared to make sufficiently extreme interpretations which could reduce measure a lot. Even if you have a blank piece of paper you can make any interpretations you want in the same way. The absence of any useful information, such as the ink of computer program listings, to be extracted merely means that your algorithm detector has to contain more information to do the job and is finding lower measure interpretations. Right now, there is some interpretation that could be made of my chair that could be used to allow me to have a conversation with a computer simulation of Albert Einstein – if I use a smart enough interpretation device. Now, you might argue that the interpretation is using information almost completely obtained from inside the interpretation machine itself and you might argue that the information I provide in the conversation is merely affecting subsequent states of the chair to provide inputs when they are interpreted in contrived ways, or even that my words do not interact with the chair at all, but I could say that contrived interpretations are not fundamentally different from obvious ones or I could say that all of the system’s possible responses to inputs are contained in the information originally extracted from it by some contrived interpretation. The only thing that could really be discussed here is *measure*.

Is “change” or “motion” really needed?

This way of looking at things may seem strange, but we pretty much do this already if we think that brains are “running” us. Is a printout really sufficient? According to the many-interpretations view a hardcopy printout of an algorithm is not only sufficient: it is not even needed, its existence affecting no more than the measure with which a particular algorithm is run. What about what Constantine (the fictional character) says in that quote though? If Constantine is right, the simple existence of an algorithm would not be enough: you would need some sort of “motion” to make it run. I think Constantine could have easily said “process” or “change” and meant the same thing. This idea may seem attractive. After all, it seems that our thinking is due to “processes” in our brains, but the idea runs into problems. Even if we accept that some sort of “process” is involved, so that somehow a thinking system needs to be “animated”, there is still the problem of interpretation. If the strong AI hypothesis is correct we still have the problem of interpreting the process or animation in some system as corresponding to physical systems and we have the problem that, if there are any processes at all going in, it will be possible to make an interpretation to find any algorithm that we want. The problem of arbitrariness of interpretation still pushes us into a many-interpretations situation, where what has been said about different interpretations and measure still applies.

The Reference Class

How should we deal with this? Once we have realised that the problem of multiple interpretations will not go away – once we accept a many-interpretations view – we have our answer. We should really view this as a reference class issue. In the thought experiments discussed previously [1,2,3,4], things were examined from the point of view of an observer who does not know what substrate he/she is on, or in what situation he/she exists. It has been suggested that in such a situation you should view your situation as being one of the situations in the reference class of all the situations that you know about. When we introduce the idea of “interpretation” to this what we really mean is that each interpretation is associated with a particular formal description of a situation and the reference class of possible situations is all possible combinations of a formally described mental state and an interpretation which associates that mental state with physical reality and that this reference class includes all formal descriptions of situations which can be obtained from physical reality. The probabilities of being in some kinds of situations are higher because a bigger proportion of the set of all interpretations correspond to them: from a reference class point of view, formally described situations like this occupy more of the reference class of all formally described situations.

We can form a coherent view like this, avoiding any arbitrary selection of particular interpretations (which Searle is so insistent we need to do), by simply making our reference class “all possible, formally described situations”. Suppose now, that we start to demand that things are in “motion” or that some sort of “process” or “change in time” is needed for an interpretation to produce a mind. What we are doing now is eliminating possibilities from our reference class. Furthermore, we have no formal basis for deciding which of the interpretations satisfy this criterion of “motion”. It may seem obvious that a brain or a computer satisfies it due to changes in the configuration over time, presumably equivalent to changing mental states, and that a description of a mind’s thoughts which simply exists does not satisfy it, but in between these we could imagine grey area interpretations where it is far from clear if the criterion of “motion” has been fulfilled. I suggest that we have no reason to enforce such a requirement for motion or change and no logical justification for it. We have a coherent view with the complete reference class containing situations including all possible interpretations and there is nothing to suggest that part of this reference class should be invalid. Furthermore, if we try to declare part of this reference class invalid, due to lack of motion or some similar condition, we will find ourselves in the same sort of mess that requires us to adopt the many-interpretations view in the first place. All we know is that our situation comes from the set of formally describable situations: anything else represents an artificial narrowing of our reference class. If we reduce the reference class based on “motion” or lack of it, why can we not similarly reduce it based on requiring that computers that support minds must have green cases, or that systems that support minds must be based on biology only? *The only sensible reference class to take is that which includes all formal descriptions of your situation.* This would mean that you should accept all possible, formal descriptions of your mental state into the abstraction set and all possible, formal descriptions of how this mental state relates to the underlying physical world into the interpretation set for that mental state, and therefore into the reference class of possible situations.

What “all possible, formal descriptions” means needs considering. It would at least need to include all the obvious, high measure, possible interpretations which could be made of part of the known physical world. A more complete view would include all possible interpretations that could be made of the known physical world, irrespective of measure. However, you are unlikely to have complete knowledge of the physical world. The only information that you have is that which is available in your mental state and, in principle, this could be related to many different physical worlds by many different interpretations. The most complete view would therefore seem to be that for any single mental state in the abstraction set, the interpretation set should include all possible, formally describable relationships between that mental state and *all possible, formally describable worlds* and that, therefore, your reference class of possible situations should take account of all possible, formally describable worlds.

I need to make a point here to prevent possible confusion. This article has argued that any interpretation, even an extreme one, of reality, which is associated with a mind means that that mind really does exist. Here, I have said that construction of your reference class of possible situations should really take into account of all possible worlds. Now, unless you subscribe to modal realism and think that all these possible worlds exist, it does not mean that all the minds associated with all these interpretations really do exist: if only your world is real then the many-interpretations view would just say that that part of the reference class corresponding to minds associated with interpretations of the real world corresponds to minds that actually exist: you would just not be completely sure which of the possible worlds is the real one.

Of course, if you subscribed to modal realism, the view that all possible worlds exist, then you would think that all the interpretations of other worlds correspond to real minds as well. In a later article, the many-interpretations view will be developed into a cosmological form which will actually suggest that a type of modal realism applies.

Motion is not needed

This leaves us in what may seem like a counter-intuitive situation, because it suggests that part of Constantine’s fictional speculation in *Divergence* is wrong – that there is no requirement for “motion” (I will observe, however, that Constantine appears to be questioning his own reasoning about the need for motion.). We can now answer Constantine’s question. Suppose a book did describe the workings of a mind in enough detail. Some interpretations based on that book would produce mental states corresponding to the experiences described in it, even without the mental states being neatly pulled out of the book in chronological order. In fact, some interpretations could produce an algorithmic description of the experiences of the mind in the book by interpreting the book as it is in a single instant of time. Other interpretations of the matter in the book may produce the book by obtaining the mental states in reverse chronological order, so that the latest experiences of the mind described in the book are extracted from it first and the earlier experiences are extracted later: the order does not matter as the entire book is available at any instant anyway. Other interpretations might be more

complex, not falling tidily into any of these categories. We might intuitively think we should discard these strange interpretations, but we have no basis for artificially reducing our reference class like this. In fact, accepting all interpretations means accepting much stranger interpretations than those involving finding minds in books. In principle, any interpretation can be made from any physical system anyway. A brain, or a computer or a book does have any magical effect which causes particular kinds of interpretations to be possible. Rather, the physical structure of a brain or a book increases the measure of particular kinds of interpretation. A book describing Einstein's thinking in detail might be said to be causing a version of Einstein's mind to exist, but some interpretations of the atoms in your desk will also cause the same mind to exist – except they will be such extreme interpretations that they will have relatively low measure. What the book does is provide a physical system which allows the Einstein interpretation to be made, from the pattern of ink molecules on the pages, by relatively short interpretation algorithms and therefore with relatively high measure, when considered alongside interpretations made from your desk.

This may seem weird. If correct it means that it is possible, *in principle*, to discover that you are a character in a book (thought that does not mean that it is at all likely). It may be your natural impulse to say that that is wrong because there is no motion, but why should it be? From one point of view, all we are doing is saying that your situation must have a formal description and that every possible, formal description should be considered as likely as any other. If we require “change” or “motion” we are artificially introducing prejudice into the reference class, unless we can explain why we need to do it. Some readers may point out that we experience “motion” or “change” and of course we do. Our experience of thinking seems to be a sequence of computational states or at least, if you do not want me to make assumptions about computation, states of *something*. In that respect, we experience change, but that does not mean that the change must be occurring in the physical system on which the interpretation is made. It is possible to take a physical system which has no change and no motion, and produce an interpretation of an algorithm going through a set of computational states which appear, from the inside, to be changing.

When we start to think like this, it is natural that we might take things beyond just considerations of computation and consider reality in general. Some modern philosophers are suggesting similar sorts of things about the nature of time. In later articles I will try to make a more general, cosmological view of things and look at these kinds of issues. In particular, I will put this idea of what we might consider our own reference class to be in a cosmological context.

If you still have problems with the idea that it is possible in principle to find you are a character in a book then I will point out that, in a sense, you have already found this to be the case. Once, you did not know about brains, or about the idea that your conscious experience is due to physical changes in your brain. Eventually, however, you became aware that your thoughts, your experiences, what you are – your entire mental situation – occur just because matter in your brain is moving in certain ways (unless you do not accept this, maybe instead believing in some kind of non-physical “soul”). You found

that you are here because a particular arrangement of matter happens to be “telling the right story”. Even if you believe in a view like Searle’s, that particular physical processes are needed, that merely means that you believe that the story told by matter needs to be very specific for you to be here. Finding out that you exist because the right pattern of ink is existing would be little different in principle to finding out that you exist because the right pattern of neuron firing is occurring.

Books and Measure

It may be possible in principle, but how likely is it? Let us imagine a thought experiment, similar to those in the first article, where there is uncertainty about the substrate on which you exist. Unlike those thought experiments, this one could actually be done without futuristic technology. You could try it as a real experiment with a friend. Do not expect any spectacular results, though. Here is the thought experiment:

You wake up in your bed and you remember that the night before you told your friend to write a short story describing you waking in bed the next morning. Assuming that the many-interpretations view is correct there must be some formal, abstract descriptions that correspond to your state of mind – your experiences – and these can be produced by making the appropriate interpretations of physical reality. We do have the problem of what characteristics a formal description of your experience should have to be regarded as a valid description of your experience – to be regarded as a member of the reference class of possible descriptions of your current experience – but that is not the problem we are dealing with here: we are dealing with the issue of interpretations that produce such descriptions.

For any mental state, the interpretation set for your situation would include these two types of formal description:

- formal descriptions based on the most “obvious” kinds of interpretations of matter in your brain that produce a description of your subjective experience.
- formal descriptions based on the most “obvious” kinds of interpretations of matter in the book written by your friend that produce a description of your subjective experience.

There would be many more possible kinds of interpretations than these, but we are just focusing on the brain and the book and assuming that they are the most likely places for you to be, based on the fact that the physical arrangement of matter in both cases particularly facilitates interpretations that lead to you.

This means that you should accept a probability that you are really in the book written by your friend. As I have suggested before, this is not all that weird. If your friend had built a clone and configured its brain appropriately, as in the film *The Sixth Day*, you would have to deal with this. Likewise, if your friend had built a computer simulation of you then, unless you have good arguments against it, it would be a real possibility. In all these cases, your friend can create special configurations of matter that contribute to the

reference class of possible situations that you could be in. Books are hardly different. We should not be prejudiced against the book on account of lack of “change” or “motion”. As I have argued already, for any mental state we should just accept an interpretation set containing all possible interpretations and they can be made from the book too.

We should however, consider measure. Interpretations that produce descriptions of your experiences can be produced from your brain by relatively short detection programs, so the measure for these will be relatively high. On the other hand, a book does not describe a human mind with the same level of detail as a brain. Descriptions of what characters in books are thinking are described more vaguely, in more general terms, than a pattern of neuron firing in a brain describes a thought. Interpretations that produce sufficiently detailed descriptions of your experience from the book will therefore need to be made by longer detection programs to “fill in” the extra information and this will significantly reduce measure. This raises the issue of how detailed the description produced by an interpretation needs to be to be part of the reference class of descriptions of *your* situation, but it seems reasonable to assume that it must be very detailed to be a candidate for your experience and this is an entirely separate question from the question of what physical configurations of matter can produce it – the main subject of this article.

All this suggests that, unless your friend is a particularly skilled writer, able to describe your experiences with the same level of detail as your neurons themselves (and I know of no human writer who can do such a thing), you should not worry too much that you have woken up in his book: interpretations that produce your experiences from the book will have such a small measure, in comparison with those made from your brain, that for practical purposes you can dismiss the idea that you are in the book.

How much detail is needed?

This deals, at least partly, with a big question about simulating minds, and one that should concern people interested in ideas such as mind uploading. One of the questions with mind uploading is that of how much detail would be required for it to be successful. Would a scan of a brain to be used in mind uploading need to encompass every atom or just every neuron, or could it even work with less accuracy than this, recording less detailed information that could be used for a passable copy? In a previous article I discussed the possibility of doing mind uploading without any brain scan at all, obtaining the information instead by observation of behaviour, a process that I called *indirect mind uploading* [22]. That article has been commented on in the past and one problem that some people have with the idea (and I expressed this concern in the article itself) is this issue of how much detail is needed to say that it has worked.

This issue of detail is the same one that we face when considering minds recorded in books. This article got close to this issue previously when I talked about someone writing a book that described your thinking in “enough detail”. How much detail is enough? The issue of detail may seem to be an intractable problem, but the many-interpretations view provides a kind of answer.

Let us suppose that something resembling a copy of your mind is being made. Whether it is in another human brain after some kind of brain scan or reading and cloning (as in the film *The Sixth Day*), a digital computer with some kind of mind uploading (as in the novels *Recursion* [15] and *Permutation City* [12]), a book after a skilled writer has described you, or even someone happening to have a mental breakdown on another planet which happens, by a freak chance, to make his/her/its mind resemble yours as it is now, need not concern us: all that interests us is that some physical system exists somewhere that could plausibly said to be running a copy of your mind as it is right now. One of the questions that commonly arises here is that of the level of detail that the copy would need to be “valid”. In the many-interpretations view this question can be dealt with by declaring it simplistic because there is no absolute “level of detail” required for validity. If we imagine a particular abstraction of your mind – some formal description of your experiences at some level of detail – then *all* of the above situations allow interpretations which produce that abstraction or formal description. Even the situation in which the physical substrate is a book cannot be said not to be detailed enough because, like all of the physical substrates described above, any algorithm or computational state can be produced from it by the appropriate interpretation.

As has already been discussed, different ways of mind uploading you are not the same. In some you will be represented with much lower detail than in others and any interpretation that produces you will need much more information to describe it, so such interpretations have much lower measure. The question of whether or not a “low resolution” or “crude” copy of you is a valid simulation of your mind may be meaningless, but the question of whether or not it produces you with very much measure is reasonable. If you did take the idea of mind uploading seriously, and you did want to try to use it continue yourself, this issue of measure may be cause for concern.

This does not deal with things entirely. As discussed earlier, there are two important issues. In reference class terms these are the issue of the interpretation set and the abstraction set. When we talk about how much measure a particular abstract description of your mind is provided with by some physical substrate we are looking at things from an abstraction set point of view. This assumes that we have some abstract description of your mind and we can consider with how much measure it exists due to different physical systems. We have, however, the problem of uncertainty about the abstract description of your mind. As discussed earlier, you have limited self-knowledge and do not know with certainty what the abstract description of your mind really is. This means that you cannot know with certainty how detailed your mental experiences are, although you may be able to make an estimate. When we say that different physical systems would “run” you with different measure depending on the level of the detail in the physical description we cannot be sure *what* needs to result from interpretations of these systems to qualify as you. There are two issues of how much detail is needed: the issue of how much detail is needed in the physical system (which is easily shown to be simplistic in the many-interpretations view) and the issue of how much detail needs to be in a formal description of your mental state for it to be part of the abstraction set and therefore affect the reference class of possible situations. We have not dealt with the second issue, and it is not the main subject of this article. With regard to the first issue, the many-interpretations

view has provided an answer: there is no minimum level of detail needed in some physical representation of a mind to qualify as “enough”: all that level of detail affects is measure.

Not being able to deal immediately with the issue of how much detail is needed in descriptions in the abstraction set, however, is not as serious as it may seem. The issue of how much detail is needed in the physical system is more important when considering measure and copies. If we are comparing two physical systems, System A, which appears to have a crude representation of your mind, and System B, which appears to have a much more detailed representation, then we may not be sure how detailed the abstract description of your mind needs to be that is supposed to result from interpretation of each of these systems, but it does not matter very much: whatever the formal description of you is, and however detailed a formal description of you needs to be in the abstraction set, it is reasonable to think that it will be produced by interpretation with much higher measure from System B than from System A, as the physical nature of System B (its “greater detail”) reduces the amount of information needed to make interpretations that produce the “correct” formal description of your mind, however detailed that formal description is. We still have the issue of what constitutes the abstraction set, so some of Constantine’s speculation is still with us, but in the many-interpretations view, *the issue of what constitutes a valid formal, abstract description of a mental state is completely removed from the physical nature of a substrate.*

Mind Uploading: Continuity and Measure

None of this resolves another controversial issue about mind uploading – that of whether or not an uploaded version of you is a continuation of you, as some people think, making mind uploading a viable life extension method. The discussion just given has not dealt with this at all, instead just showing that questions of which substrates would be valid for mind uploading, whether or not “change” or “motion” is needed for a mind and whether or not the uploaded version of you is detailed enough to have a mind are meaningless and that the level of detail is only relevant with regard to measure.

There is one way in which the issue of measure could matter when considering mind uploading: some mind uploading processes may cause a version of you to exist with much less measure than that with which you exist now. Unless some attempt is made to conserve measure, and unless whoever designs the technology has confronted the sorts of issues discussed in these articles, a huge collapse of measure in mind uploading is almost inevitable. This is not due to any obvious “superiority” of organic devices like brains – far from it. It is because humans tend to make computers as efficient as possible. Faster computers, with greater memory, means the smallest components possible, meaning that that the tendency is to build computers which run algorithms with the smallest amount of redundancy – and the lowest measure possible. The idea that if you are going to be mind uploaded you should maybe ask for an inefficient computer may seem strange.

If you do not think that mind uploading provides a way of continuing yourself then this will probably not matter much. If, however, you think that mind uploading is a viable life

extension method then the possibility of a dramatic decrease in measure during the process should possibly be an issue for you: I am not saying that it is an issue, merely that it creates the *possibility* that, in some statistical sense, *you die*. A similar kind of point was made by Bostrom in his article about substrate [8]. It is also possible that measure is not something that we should view as important – though if humans, or our descendants, come to the conclusion that measure conservation is not important we end up with a civilization acting in ways utterly alien to our standards. Some of the arguments that might be used to defend mind uploading from this issue of measure collapse would be similar to arguments used to support quantum immortality, a controversial idea.

The “Haunting” of Matter

I do not want to persuade anyone under false pretences to accept the ideas in this article – to accept them casually, without full awareness of the implications. This article implies that every piece of reality can be interpreted in an infinity of different ways to produce different algorithms, some of which correspond to minds. Therefore, any object in reality could be said to be “haunted” by an infinity of algorithms (speaking informally), some of which describe minds. This is useful to you because it allows you to exist. Presumably, you exist because of a relatively high measure, sensible interpretation of your physical substrate, so in a sense you are just “haunting” your brain along with any minds produced by other interpretations. You just have higher measure. Some readers may find this concept disturbing and many readers will find this, in itself, reason for dismissing the whole argument. I would suggest that it is best to consider the argument in terms of how good its logic is, and how well it deals with the relationship between physical systems and algorithms. If uncomfortable ideas are needed to attain coherency then that is how things are.

What I Am *Not* Saying

I want to be clear about what this article is not saying. All kinds of mediums, psychics, New Agers and similar people could attempt to relate these ideas to all kinds of claims of contact with spirits, etc. This would be wrong. Even if we accept the existence of other minds associated with contrived computational interpretations of matter, they would not be “observable” by us in anything like the conventional sense of the word and we could never communicate with them. The only way we could know of their existence is by inferring it in arguments like this and, if we want to do so, by applying the relevant extreme interpretations to the physical system.

Why could one of these these “other minds” never meaningfully communicate with us? If it had conventionally observable inputs or outputs then it would be observable in this way. An appropriate input and output reporter (a similar sort of device to the algorithm detector) could find these inputs and outputs using what should be, by definition, an uncontrived interpretation with a short detection program: if a long detection program were needed then finding the inputs and outputs would be too complex for us to “observe” them conventionally. Furthermore, an algorithm detector with a suitable detection program should be able to find these inputs and outputs, examine the data

associated with them, and determine the existence of the algorithm, and if observing the inputs and outputs is that easy then a short, uncontrived detection program should suffice. The mere fact that an algorithm is only findable by making contrived explanations effectively removes it from everyday human experience. It can never interact with the macroscopic variables that we can measure in reality. We can only ever know that such an algorithm is running at all on a philosophical basis – by being prepared to accept extreme interpretations in the interests of philosophical consistency.

Reluctantly accepting the existence of something, on a theoretical basis, to avoid inconsistency is different to saying that the thing plays any part in human affairs or is talking to you. *Anyone trying to use this article to explain “paranormal events” will be misrepresenting me.*

General Implications for Searle’s Argument

This article has taken a point made by Searle seriously – that what he calls “multiple realizability” [6,7] appears to make it arbitrary to say what algorithm a physical system is running, and therefore, if minds are based on algorithms, what mind or minds exist in a physical system. Searle says we should therefore discard computational views of consciousness.

This article, however, has suggested an alternative – that we do not need to select an algorithm, but can simply admit all of them, while recognizing that the amount of information needed to make an interpretation means that some algorithms will have high measure relative to that of others. *This issue of relative measure is completely ignored in Searle’s claim of observer subjectivity.* It is not necessary to discard a computational view of consciousness, although the strong AI hypothesis, when viewed in this way, could be regarded as being qualified. Searle’s argument was not intended to force us to qualify the strong AI hypothesis, but to force us to throw it out, yet by simply accepting many algorithms it becomes irrelevant. *Searle’s multiple realizability argument therefore fails.* The many-interpretations position causes issues with other arguments by Searle too, as will be explored later in this series.

Strong and Weak Many-Interpretations

What I have said here cuts across so much of philosophy that it is going to be contested. It might be contested, however, in varying degrees. It may be useful to consider the many-interpretations view in two ways: the *strong many-interpretations view* and the *weak many-interpretations view*.

- The **strong many-interpretations view** would be the position that I have advocated in this article. Any interpretation of physical reality that is associated with a mind is regarded as being associated with a mind that really exists.
- The **weak many-interpretations view** would not make claims about what minds do or do not exist. It would simply use the many-interpretations view to construct the reference class of possible situations in which you could exist.

Many-Interpretations and Many-Worlds

Earlier in this article I mentioned Everett's many-worlds interpretation of quantum mechanics (MWI), also known as the relative state formulation of quantum mechanics [23]. This is an attempt to explain "weird" results in experiments like the double-slit experiment with single photons. MWI lacks any wavefunction collapse mechanism. Instead, it claims that the wavefunctions are all that exists and that when there are multiple possible outcomes for some event all outcomes occur, but that the process of decoherence soon prevents the wavefunctions from affecting each other, resulting in what can be regarded as separate worlds.

In MWI, particles do not exist at a basic level. Instead, there are only the wavefunctions and decoherence or splitting leads to the appearance of particles. MWI is controversial and I am not assuming that it is true in this article. If, however, it is true then it would seem to affect the interpretation set for any mental state, and therefore the reference class of possible situations.

If MWI is false then any interpretation in the interpretation set for a given mental state should presumably be of the arrangement of particles, or of whatever the particles are made out of. However, if MWI is true then the most accurate way of constructing the interpretation set for a given mental state, and therefore the reference class of possible situations, would be to use interpretations of the quantum wavefunctions themselves. This may seem not to make much difference, beyond going to a slightly deeper level, but there is one respect in which the difference is important: as decoherence and splitting of worlds occurs, the wavefunctions in single worlds are effectively "sliced" progressively more thinly making the situation equivalent to a computer in which "less matter" is used as in the thought experiments in the earlier articles [1,2,3,4]. This means that, all else being equal, splitting, by "thinning out" the quantum wavefunctions, would therefore reduce the measure of interpretations associated with your mind in any single world relative to the measure of interpretations associated with your mind in a world before the splitting occurred. This effect will, however be balanced by the proliferation of worlds, provided that you continue to exist in all these worlds.

This could have relevance to a number of philosophical issues, including the simulation argument [20,21], quantum immortality and quantum suicide and the Doomsday hypothesis.

As an example of how this could affect some philosophical issues, I recall seeing a web posting in which someone, using reasoning similar to that in the Doomsday hypothesis, asked how, if worlds proliferate in the many-worlds interpretation, we should not all expect to find ourselves in the last fraction of a second of our lives. The idea was that as worlds proliferate, observers proliferate and the number of observer moments in which you are living in the last fraction of a second of your life would be vast compared to the number of observer moments in which you exist at any appreciable time before it. Therefore, observer moments in which you are about to die imminently should be much more common than any other type of observer moment. I will not go into the issue, here,

of whether or not this kind of reasoning, this statistical handling of observer moments, is valid. *That* is controversial. If we accepted it as valid we might answer by saying that the measure of the worlds resulting from splitting is reduced as the total measure is the same, but that does not really explain why these observer moments should count for less. The many-interpretations view, however, would provide a detailed answer and in the many-interpretations view the absence of this effect would be entirely expected. The proliferation of worlds at later times would not result in an equivalent proliferation of observer moments because observer moments would result from interpretation and the more thinly sliced wavefunctions of such worlds would reduce the scope for such interpretation.

Of course, if you think MWI is nonsense you do not need to worry about this.

Putting Probabilities on Strange Ideas

Earlier in this article I said:

“The most complete view would therefore seem to be that for any single mental state in the abstraction set, the interpretation set should include all possible, formally describable relationships between that mental state and all possible, formally describable worlds and that, therefore, your reference class of possible situations should take account of all possible, formally describable worlds.”

This tells us how to construct a reference class of possible situations in which we could exist and by doing this we can determine the probability that we are in any given type of situation. If, for example, situations of some given type occupy about 60% of the reference class then we should assume that there is a 60% probability that we are in that kind of situation. Someone might argue that this makes no sense as the reference class is infinite, but that is easily dealt with by making the sort of point about descriptions of length that tends to infinity that I have made a number of times now: I will not bother making it again here.

This approach to constructing a reference class of possible situations does not distinguish between “conventional” and “strange” situations and, in principle, allows us to put probabilities of being true on various ideas which many people might think are beyond our knowledge. One such example is the simulation scenario, suggested by the simulation argument [20,21] which has been mentioned previously in this article and which some people might think is the sort of issue we cannot know anything about beyond speculation. I disagree – and I am not saying that something like the simulation scenario is correct, but merely that it is not beyond what we can know about. We should not need literally to count individual possible worlds to perform such a calculation. With enough computing power, and the ability to formally express the problem, we may be able to determine probabilities by making some approximation of the reference class and exploring that. This means that in the future we can fully expect to know, for example, the probability that the sort of simulation scenario considered in the simulation argument

applies – in a formal way, without just having to guess. The probabilities could be further refined by using more computing power.

A consequence of this is that you should expect any civilization appreciably more advanced than ours – if any exist anywhere – not to be spending much time making informal guesses about all kinds of strange ideas about their possible place in reality: they would know the probabilities.

Comparison with Greg Egan’s *Permutation City*

Readers of Greg Egan’s science fiction novel *Permutation City* [12] will find some of the ideas being discussed here familiar. A character in this novel proposes a cosmology based on the idea that all interpretations of matter are valid and, in the novel, this idea turns out to be correct. The suggestion may be made that this is “just *Permutation City* turned into a philosophy article”. Given the popularity of Egan’s book, some important differences between this article and the idea in Egan’s novel are pointed out in the appendix.

Further Objections

Objection 11: There is no observable evidence for what you propose.

Answer

This article has suggested a logical argument for the position it takes about minds, interpretation and measure and that argument is itself evidence.

Objection 12: It violates Occam’s razor.

Answer

It may seem to violate Occam’s razor to suggest that all of these other interpretations and (where relevant) minds should be accepted as “existing” but there is some controversy about what the modern version of Occam’s razor should be taken as meaning. Some people think we should minimize the amount of “stuff” we think exists. Others (I am one of them) think that we should minimize the amount of information content in our underlying description (or theory) of reality which explains where all this stuff comes from. This disagreement shows itself most obviously in different attitudes to Everett’s many-worlds interpretation of quantum mechanics [23]. Some people say that the huge number of “other worlds” in the many-worlds interpretation is a violation of Occam’s razor. Others say that the many-worlds interpretation, as an underlying model of reality, is not hugely complicated and all the “other worlds” that it predicts are implied by the information content of this underlying model and do not need to be separately assumed. This is relevant here. If you think we need to minimize “stuff” then all the other interpretations, minds and algorithms proposed in this article are certainly a lot of “stuff”. If, however, you merely think that the underlying model of reality should be simple then I would point out that this article proposes a fairly simple idea – allowing all algorithmic

interpretations of reality to correspond to algorithms that are running and, where relevant, minds that exist – and that this infinity of algorithms and minds is implied by this and does not need assuming in its own right.

Whatever view of Occam's razor you have, I also point out that this article provides a logical argument justifying its proposal. If that logical argument turns out to be wrong then that is one thing, but if the argument is valid Occam's razor would hardly overthrow it. Occam's razor cannot validly tell you that things do not exist if they are shown to exist by a correct logical argument. You should accept the simplest view of reality which is logically coherent, and this article has suggested that selectively allowing some interpretations and not others is logically *incoherent*. However we understand Occam's razor, it cannot justify an incoherent position.

Objection 13: You deal with the impossibility of counting interpretations in an infinite set by determining the fraction of all interpretations that are of a given type as the maximum length of the interpreting algorithm tends to infinity. This is invalid: you need an actual number of interpretations of things to make sense and you cannot get it.

Answer

This sort of approach is widely regarded as valid in mathematics. Infinities do not prevent us from dealing with things and getting *real* numbers as results. The basic derivation of differential calculus uses a similar kind of approach, as I stated in the previous article [3,4]:

“Situations like this do not necessarily prevent extraction of useful information, however. An example is the derivation of the basic method of differential calculus. We define the gradient at a point on a curve as the y step divided by the x step over some small section of the curve and then we let the size of this section tend to zero - as if it were becoming infinitesimally small. The objection could be made that we cannot compare infinitesimally small values, but we never do compare them. The size of the section tends to, but does not reach, zero, and we can make a useful comparison of the y and x steps. This is not a comparison for a zero size section of curve, but nor is it a comparison for any particular absolute size: it is something more abstract. We do not need absolute values to make a comparison. It is the same with the issue of "infinite minds". Even if we find ourselves having to infer infinite amounts of minds, it is relative numbers that we are interested in for the purposes of assigning probability and value.”

Objection 14: You only deal with the issue of interpretation, but do not resolve the issue of what algorithms correspond to minds. Suppose a particular interpretation produces the *WordStar* algorithm: who is to say that does not correspond to a human mind? That would be absurd, yet seems completely allowed by what you say. You have not solved the problem at all.

Answer

Yes, it would be absurd, but that is not what I am saying.

There is some truth to the objection, but it does not recognize that there are two different problems: deciding what algorithms are running in a physical system and deciding what algorithms should be regarded as causing minds or, in some thought experiments like those in the first two articles, a particular mind in a particular situation.

This article only offers a solution to the first problem (What algorithms are running?) but this is needed for any solution to the second problem (What algorithms are equivalent to minds?) to be of any use.

The objection's claim that I am proposing arbitrariness in the issue of what algorithms correspond to minds is wrong: it confuses the argument I have made about arbitrariness in interpretation – in deciding what algorithms are running in some physical system – with an argument I have not made about arbitrariness in deciding which algorithms correspond to mental processes.

I see no reason to think that the sort of arbitrariness which causes problems in interpretation is also an issue in deciding what algorithms are equivalent to minds. We currently lack a formal method of deciding which algorithms are equivalent to minds, or which are equivalent to our own situation at any time, but we can get one later and in the meantime we can make common-sense judgements. With the arbitrariness of interpretation problem, common sense is hard to defend because different interpretations can produce algorithms that might all seem to match our experiences, but when it comes to deciding which algorithms validly correspond to mental experience some will appear clearly wrong.

Someone could argue against this that you could take an algorithm that appears clearly wrong and apply some interpretation to it to get a valid algorithm that seems to match your mental situation. For example, you could find that some interpretation algorithms applied to a wall give the *WordStar* algorithm with some measure. You could then apply some interpretation to the *WordStar* algorithm which produces an algorithm that seems to correspond to a conscious mind and argue that, because the *WordStar* algorithm is running with some given measure, every version of the *WordStar* algorithm should be regarded as equivalent to a mind and that, therefore, we still have some arbitrariness. The problem with this argument is that, even if we accept a second stage of interpretation like this (where we can apply interpretations to algorithms found by interpretation of physical systems), it does not follow that any algorithm found in such a way gives its characteristics to the algorithm which was interpreted to get it. It would simply mean that the same issues of measure arise for the algorithms produced in the second (and any subsequent) stages of interpretation.

Objection 15: You assume that consciousness can be produced by algorithms. Penrose argues that consciousness cannot be produced by algorithms and gives a

mathematical proof of this. Further, if consciousness cannot be produced by an algorithm you should not expect an algorithmic interpretation to find it. Looking at this infinity of interpretations is looking at the wrong thing.

Answer

Penrose's argument is controversial and many arguments have been made against it. I have written an article refuting his "proof" [17] and articles by others are available. I have written about algorithms because this is how the strong AI hypothesis is viewed, however if it turns out that non-computability is needed to understand nature and minds this does not necessarily change the main point of my argument. Even Penrose does not try to suggest that minds are not formally describable, so any situation in which you exist as a conscious being should be formally describable. This means that we have to decide whether our reference class of situations should consist of every situation that is formally describable by an algorithm or every situation that is formally describable in some non-computable way. If we decide on the latter then I would suggest this merely means that an infinite set of interpretative algorithms becomes an infinite set of "non-computable somethings" and the problem of arbitrariness of interpretation does not go away. Further, this could be awkward for Penrose's argument because Penrose suggests that specific physical processes are needed to have minds. If non-computability can be formalised, however, we should be able to take any physical system and make some non-computable interpretation of it which produces some non-computable formal description of a mind, although this may occur with low measure relative to that with which it occurs in systems such as brains. Even if non-computability is important to physics and study of the mind, we should be able to expand the argument made in this article to incorporate it.

It may be that a consideration of the reference class for our situations, what would need to be included in it to give a complete reference class (whether it would need to be made from algorithms or some other entities that include algorithms as a special case), and what measures apply for different situations could give an answer on whether or not we should take Penrose's non-computability seriously.

Objection 16: The abstraction and interpretation sets, and therefore the reference class of possible situations, depend on the formal language used to describe them, so they would be subjective.

Answer

If this were really the case it would be futile to try to use any formal methods in probability, as probability itself would become subjective. I have previously written an article dealing with this issue with languages [24].

Objection 16: It sounds like a mystical, New Age idea.

Answer

I have pointed out that ideas in this article are open to misrepresentation. Even if an argument may sound attractive to some people, or may have some superficial similarities with their claims, that does not make it wrong. Any flaw must lie within the argument or its assumptions.

Conclusion

Searle, in his criticism of the strong AI hypothesis, has a point with the objection of multiple realizability that needs answering. Multiple realizability, or arbitrariness of interpretation (as I refer to it) is the issue of it being possible to extract any computational interpretation you want from any physical system. If we say that performance of particular computations on physical substrates causes minds, therefore, we are just applying what may seem to be an arbitrary computational interpretation to that system and the same interpretation could be applied to any other physical system. It may seem therefore that the strong AI hypothesis requires us to choose particular interpretations with no obvious criteria. Searle's answer to this is that the strong AI hypothesis is incoherent.

One way of arguing with Searle would be to try to show that some interpretations are the "correct" ones. I think such an approach would be invalid and that none of the reasons for viewing particular interpretations as correct would be on firm ground. There is a different alternative to accepting the strong AI hypothesis as incorrect. This is to accept all computationally expressed interpretations as valid. This "many-interpretations" view is not a change to strong AI: it is merely a clarified version of it that deals with the issue of interpretation. With such a view the strong AI hypothesis is coherent and different computations are performed with different degrees of measure by different physical systems.

A conceptual device, an algorithm detector, and some related devices which all do basically the same thing were described. The algorithm detector concept formalizes the idea of interpretation by using an interpretative algorithm known as a detection program, with a set of physical probes, to apply a particular interpretation, or set of interpretations, and extract an algorithm from a physical system. With such devices, the problem of arbitrariness of interpretation is formalized as the problem of determining what detection program should be used to detect various computations being performed in physical systems. The problem is that any computation can be found to be occurring in any physical system if the detection program is permitted to be sufficiently sophisticated – and long.

The "many-interpretations" view deals with this by permitting all detection programs that could be placed in devices like algorithm detectors and stating that any computation found to be occurring by any valid detection program should be regarded as actually

occurring for the purposes of constructing a reference class (and this is an informal way of putting it). Because any computation can be found in any physical system, given a sufficiently long detection program, this implies that all computations should be regarded as actually occurring in any physical system. This may appear to be a nonsense result, but the issue of *measure* needs to be considered. Different computations occur with different degrees of *measure*. The relative measure of two different computation, or two general types of computation, is determined by counting and comparing the total numbers of computations in each case as the maximum length of the detection program tends to, *but does not reach*, infinity.

The nature of the physical system will influence the amount of information needed in the detection program to apply interpretations that extract particular algorithms. This agrees with the “common-sense” view of the strong AI hypothesis. Those computations which would normally be intuitively regarded as being performed on a physical system are those which, in the many-interpretations view, will tend to be extracted from the physical system with high measure.

This may seem to be an attack on the strong AI hypothesis. This clarification of it is not a weakening of it. This is an *extreme* form of the strong AI hypothesis – one in which the *association* between symbol processing and minds is taken so literally that any interpretation of a system which gives the correct symbol processing is recognized as corresponding to a real mind.

The concept of *measure* in the many-interpretations view deals with the allegation of observer subjectivity that Searle makes against computational views of the mind: the measure with which any particular algorithm or type of algorithm is running is independent of any observer, but instead depends on the combined effects of all possible interpretations.

The issue of uncertainty about your own mind should be acknowledged when dealing with many-interpretations and measure. You do not know the exact formal description of your mental state and all you can say, therefore, is that there is a group of possible, formal descriptions of your mental state, or a description of the general type of mental state to which yours belongs. This uncertainty is relevant in making any assessments of probability.

Two issues need to be considered with regard to formal descriptions of your mental situation. One issue is that of the formal description of your mental state. This is a description of your mental experiences and is completely abstract: it is separate from any physical considerations. The other issue is that of the formal description of how various features of the physical world relate to your mental state – the issue of the formally expressed interpretation of the physical world from which your mental state is derived. You will always have some uncertainty about each of these, so they can be considered in reference class terms. There is a set of possible, formal, abstract descriptions of your mental state – the abstraction set – and for each of these abstract descriptions there is a reference class of possible, formally expressed interpretations of the physical world from

which your mental state follows – the interpretation set. This article deals with the issue of interpretation and the interpretation set. Any situation that you could be in involves a formally expressed, abstract description of your mental state and a formally expressed description of how that mental state relates to the physical world – an interpretation from that mental state’s interpretation set. The abstraction set and interpretation sets therefore determine the reference class of possible situations in which a thinking being could exist and its is made up of abstract description and interpretation pairs.

There is an issue of whether or not the many-interpretations position is a computational view of the mind. This depends on what “computational view of the mind” means. The many-interpretations view could be regarded as non-computational in the respect that no explicit assumption is made that systems are somehow “doing computation” or “processing symbols”. Instead, computation is associated with the *description* of mental states and the relationships between mental states and the underlying physical world. We might, however, still informally say that minds are algorithms that are running on some physical substrate.

This series of articles has relied a lot on thought experiments involving the probability of you being in different situations, in which you exist on different physical substrates. This kind of consideration allows the strong AI hypothesis to be viewed in reference class terms. A reference class view of the strong AI hypothesis would include physical substrates that would generally be considered as “computers” within the reference class of possible situations in which you might exist, without treating them as philosophically special, and without assuming that physical substrates such as brains must somehow be “computational”. Instead, it is merely assumed that any mental experiences that you are having, and how these relate to the physical world, must be capable of being formally described, and that such formal descriptions could relate minds to substrates such as “computers” so those sorts of situations should be in the reference class. This may appear to be nothing more than an attempt to get a computational view of the mind in by the back door, and in a way it is a computational view of the mind, but no explicit assumption that the mind is computational is needed: the only assumption is that the relationship between mental states and physical reality can be formally described. This has relevance for two controversial philosophical arguments: the simulation argument and Searle’s argument against strong AI. The simulation argument is stated as relying on an assumption of substrate independence, but a reference class view of the strong AI hypothesis makes it clear that no such explicit assumption is needed – although this does not necessarily mean that the argument is correct. Searle’s argument against strong AI maintains that we should not have any good reason to think that computers could have minds because some specific physical process is needed to cause them, but a reference class view of the strong AI hypothesis would suggest that you have no reason to exclude any possible, formal description of our situation as a member of the reference class of possible descriptions of your situation – and some of those descriptions would involve what would be considered computers, but without needing any particular assumptions about computing and minds to get them into the reference class.

One issue that has often been considered, particularly with regard to AI and “mind uploading” is that of what would be needed in physical reality for a mind to exist. If a computer running the correct description of mental processes could cause a mind to exist then what about a book? Do mental experiences exist just by being described or is some sort of change or motion needed in the description? Does a mind need to interact with the external world? With how much detail does some physical substrate need to describe a mind to cause it to exist? The many-interpretations view provides a coherent way of dealing with these questions. The way to approach questions like this is in terms of the reference class of possible situations which is constructed by knowing the abstraction and interpretation sets and the approach should be one of not unnecessarily restricting membership of these sets: if you know that some member of, for example, the interpretation set for a given mental state is inconsistent with your experiences you should discard it, but you should not discard on other grounds. This deals with the question of whether or not movement or change is needed in the physical substrate. There is no need for such change. All that is needed is the logical capability for some interpretation to extract the appropriate algorithmic description of your experiences from it. The issue of how much detail is needed is also resolved. There is no minimum amount of “detail” for the physical substrate to be associated with a mind, because there will always be interpretations that can produce it. This does not make the physical substrate irrelevant. The more detail in the physical substrate, the less the amount of information is that is needed to interpret it as an algorithm corresponding to your mental state, and the higher the measure will be. The amount of detail with which a mind is represented in the physical substrate therefore determines its measure. The issue of whether or not interaction with other parts of reality should be demanded is similarly dealt with by applying the principle of not restricting the reference class without cause. The strangeness of this situation is brought into focus by considering books containing descriptions of the working of minds. According to the many-interpretations view such a book increases the measure with which interpretations can extract that kind of mind from the book. If this seems strange, we should consider that any idea that computers could have minds, or that humans have minds due to “the right sort of processing” is similarly based on the idea that minds exist when the appropriate narrative exists: this is the natural end-point of such an idea. This has implications for mind uploading. It suggests that the measure with which an uploaded mind would exist due to being simulated on a computer would depend both on the physical characteristics of the computer and the amount of detail in the simulation – and computers tend to be built to reduce inefficient use of matter as much as possible, which would decrease measure. This means that mind uploading could have issues of preservation of measure, as well as the existing issues about continuity. Whether or not measure should be a concern for anyone in the future about to rely on mind uploading for continuation of consciousness is a matter for philosophical enquiry beyond the scope of this article.

The “haunting of matter” is a strange consequence of the many-interpretations view. If every interpretation of a physical system that describes a mental state really does correspond to a real mind, then this implies that each of the infinity of possible ways of interpreting any physical system as a mind actually corresponds to a real mind. Rather than trying artificially to get rid of this, I suggest that it should be accepted as reality and

that this infinity of minds is implied by relating minds to formal descriptions of physical systems. This does not mean that all minds are equal, and that the nature of a physical substrate is irrelevant, however: measure is an issue and some types of mind would be produced by some types of physical substrate with higher measure than others.

Some aspects of what I have said here, in particular the “haunting of matter” idea, might be viewed by some people promoting pseudo-scientific views as supporting their claims. Any mention of other minds here is likely to be of philosophical interest only and, if we are philosophically obliged to accept the existence of minds associated with extreme interpretations of physical systems, no meaningful communication or interaction would be possible with such minds, as might be claimed to be had with spirits, ghosts, etc. I hope readers would expect my likely disagreement with the use of this article to support such ideas.

I have mentioned the implications of the reference class view of the strong AI hypothesis for Searle’s argument against strong AI. The general idea in this article also raises issues for Searle’s specific argument of multiple realizability that he uses against strong AI. Multiple realizability is what I have referred to as arbitrariness of interpretation. Searle suggests that the possibility of making any computational interpretation of any physical system means that computation is subjective and that a particular computation cannot be said to happen, or not to happen, in a given physical system. Searle claims that this makes the strong AI hypothesis incoherent, as it cannot be said that a mind that is dependent on a particular sort of computation occurring exists in any objective sense, given the subjectivity of the computation. The many-interpretations view does not have this problem: accepting *all* interpretations removes any subjectivity in *choosing* a particular interpretation.

The many-interpretations view could be considered in two ways – as the *strong many-interpretations view* and the *weak many-interpretations view*.

- The **strong many-interpretations view** would be the position advocated in this article. Any interpretation of physical reality associated with a mind is regarded as being associated with a mind that really exists.
- The **weak many-interpretations view** would not make claims about what minds do or do not exist. It would simply use the many-interpretations view to construct the reference class of possible situations in which you could exist.

It might seem natural to assume that the many-interpretations view advises us to base a reference class of possible situations on those interpretations that can be made of basic particles of matter, but this may not be the case. In the many-worlds interpretation of quantum mechanics (MWI), a controversial theory proposed to explain strange experimental results, reality is really quantum wavefunctions and the appearance of particles at defined positions is due to decoherence of these wavefunctions – experienced by us as splitting of worlds. If MWI is correct then the interpretations should be based on these quantum wavefunctions and the resulting “thin slicing” of worlds as time progresses could be of relevance to some philosophical issues.

The method of reference class construction proposed by the many-interpretations view could be used as the basis for formal, computer assisted investigation of the statistics of possible situations in which we exist. This means that we can expect one day to have numerical probabilities on various “strange” ideas – such as simulation scenarios or other sceptical hypotheses.

The discussion raised in this article has some similarity with “the dust hypothesis” – the idea of a fictional character in Greg Egan’s novel *Permutation City*. However, this article is not merely *Permutation City* expressed as a philosophy article: there are notable differences between this article and Egan’s novel. In particular, in *Permutation City* the main justification for accepting different interpretations appears to be that different interpretations are logically possible, but in this article there is the different, more compelling justification that, without many-interpretations, arbitrariness of interpretation prevents establishment of a reference class of possible situations in which we might be and therefore makes probability incoherent, as shown in the thought experiments. Further, the physical nature of a system is related to the statistics of interpretations with the observation that some physical systems will imply some algorithms, due to interpretation, with greater measure than others. This agrees with our common-sense expectations about the sorts of algorithms we would naturally associate with computers or brains, for example, and is of particular relevance when minds are involved.

References

- [1] Web Reference: Almond, P. (2007). *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence*. Retrieved 12 September 2007 from <http://www.paul-almond.com/Substrate1.pdf>. (Also at <http://www.paul-almond.com/Substrate1.htm>).
- [2] Web Reference: Almond, P. (2007). *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence*. Retrieved 13 September 2007 from <http://www.machineslikeus.com/cms/minds-substrate-measure-and-value-part-1-substrate-dependence.html>. (A copy of the article in Reference [1]. Includes reader criticism of the article).
- [3] Web Reference: Almond, P. (2007). *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence*. Retrieved 3 November 2007 from <http://www.paul-almond.com/Substrate2.pdf>. (Also at <http://www.paul-almond.com/Substrate2.htm>).
- [4] Web Reference: Almond, P. (2007). *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence*. Retrieved 10 November 2007 from <http://www.machineslikeus.com/cms/extra-information-about-substrate-dependence.html>. (A copy of the article in Reference [3]. Includes reader criticism of the article).
- [5] Searle, J. R. (1997). *The Mystery of Consciousness*. London: Granta Books. 1998.

(Originally Published: 1997. New York: The New York Review of Books. Also published by Granta Books in 1997.)

[6] Ibid. Chapter 1, pp14-17.

[7] Searle, J. R. (2002). *The Rediscovery of the Mind*. Cambridge, Massachusetts: The MIT Press. 9th Edition. Chapter 9, pp207-212.

(Originally Published: 1992. Cambridge, Massachusetts: The MIT Press.)

[8] Web Reference: Bostrom, N. (2005). *Quantity of experience: brain duplication and degrees of consciousness*. Retrieved 13 April 2008 from

<http://www.nickbostrom.com/papers/experience.pdf>.

(Also published in *Minds and Machines* (2006) 16:185-200).

[9] Ballantyne, T. (2007). *Divergence*. New York: Bantam Spectra. Constantine 6, 2252. (Fiction).

[10] Searle, J. R. (1980). Minds, brains and computers. *The Behavioral and Brain Sciences* 3:417-457.

[11] Russell, S. J., Norvig, P. (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall.

[12] Egan, G. (1994). *Permutation City*. London: Millennium. (Fiction).

[13] Web Reference: Strout, J. *Mind Uploading Home Page*. (2002). Retrieved 22 June 2003 from <http://www.ibiblio.org/jstrout/uploading/MUHomePage.html>.

[14] Web Reference: *Mind Uploading Research Group*. (2002). Retrieved 22 June 2003 from <http://minduploading.org/>.

[15] Ballantyne, T. (2004). *Recursion*. London: Tor UK. (Fiction).

[16] Penrose, R. (New Edition, 1995). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. London: Vintage. Chapter 2, pp64-76.

(Originally published:1994. Oxford: Oxford University Press).

[17] Web Reference: Almond, P. (2004). *A Refutation of Penrose's Godel-Turing Proof that Computational Artificial Intelligence is Impossible*. Retrieved 24 October 2007 from <http://www.paul-almond.com/RefutationofPenroseGodelTuring.htm>.

[18] Web Reference: Almond, P. (2005). *Occam's Razor Part 4: An Overview of How Occam's Razor Works*. Retrieved 24 December 2005 from <http://www.paul-almond.com/OccamsRazorPart04.htm>.

[19] Hofstadter, D. R. (1999). *Godel, Escher, Bach: an Eternal Golden Braid*. London: Penguin. Chapter 20, pp697-698, p706.
(Originally Published: 1980. New York: Vintage.)

[20] Bostrom, N. (2003). Are you living in a computer simulation? *Philosophical Quarterly*, 2003, Vol. 53, No. 211, pp 243-255. (Bostrom circulated a draft of this paper in 2001).

[21] Web Reference: Bostrom, N. (2003). *Are you living in a computer simulation?* (An online version of the article in reference [5]) Retrieved 8 September 2007 from <http://www.simulation-argument.com/simulation.html>. (Further information about this subject by Bostrom and others is at <http://www.simulation-argument.com>.)

[22] Web Reference: Almond, P. (2003). *Indirect Mind Uploading: Using AI to Avoid Staying Dead*. Retrieved 9 August 2003 from <http://www.paul-almond.com/IndirectMindUploading.htm>.

[23] Everett, H. (1957). Relative State Formulation of Quantum Mechanics. *Reviews of Modern Physics* 29, pp454-462.

[24] Web Reference: Almond, P. (2005). *What is a Low Level Language?* Retrieved 17 July 2005 from <http://www.paul-almond.com/WhatIsALowLevelLanguage.htm>.

Appendix: Differences Between this Article and Greg Egan’s *Permutation City*

In Greg Egan’s novel *Permutation City* [12] a cosmology known as the “dust hypothesis” was proposed, based on the idea of all interpretations being valid, but this article has not looked at this as a cosmological issue (although I will be looking at that later). The focus here has been on computation and minds.

In *Permutation City* the main justification for accepting different interpretations appears to be that different interpretations are logically possible. In this article a different justification is provided, and one that I think is more compelling – that, without the many-interpretations view, arbitrariness of interpretation prevents us from establishing a reference class of possible situations in which we might be and therefore makes probability incoherent.

The arbitrariness of interpretation in *Permutation City* seemed to be based on different ways of sticking pieces of matter together. The arbitrariness of interpretation is more general, relying merely on the logical possibility of an algorithm making a particular interpretation.

This article involves the idea of measure and discusses the issue that statistics would apply to interpretations.

The physical nature of a system is related to the statistics of interpretations with the observation that some physical systems will imply some algorithms, due to interpretation, with greater measure than others. This makes sense of our common-sense expectations about the sorts of algorithms we would naturally associate with computers or brains, for example, and is of particular relevance when minds are involved.

This article has pointed out that the many-interpretations view is, at least in part, a clarification of the strong AI hypothesis and could actually be viewed as an extreme version of strong AI.

This article has pointed out that the many-interpretations view, if correct, makes Searle’s “multiple realizability” argument against the strong AI hypothesis invalid.