

Minds, Substrate, Measure and Value

Part 6: Further Justification

Paul Almond

<http://www.paul-almond.com> info@paul-almond.com

3 July 2011

This is the sixth in a series of articles about the nature of reality, the nature of consciousness and the relationship between minds and the rest of the world. Previous articles in the series have argued that, for a coherent view of probability, a measure view of consciousness is needed, in which an observer exists with some measure that is substrate dependent. An implication of this is that we must accept that all interpretations of reality which “find” an observer find a real observer, and that an observer exists in any situation with a degree of measure that depends on the substrate and the amount of information needed in an interpretation for it to find that observer. This was extended to be a cosmological position, implying that we are in an extensive multiverse. In this article, further justification is given for the argument. A thought experiment involving gradual reduction of resolution, or detail, in the physical representation of a mind is discussed. Various objections are considered and answered.

1 INTRODUCTION

This is the sixth in a series of articles exploring the relationship between minds and physical systems (substrates) on which they are based. The argument made in the series has now been extended, becoming a cosmological argument. (A list of the previous articles in the series [1,2,3,4,5], and where to find them on the internet, is at the end of this article.)

In the previous articles, an argument was made for the *many-interpretations* view: the idea that all interpretations of a physical system should be treated equally, and that any interpretation, no matter how extreme, that describes a mind should be regarded as corresponding to a real mind. The argument was then widened in scope, becoming the *cosmological many-interpretations* view, which treats all interpretations that describe objects as implying real objects. An implication of the cosmological many-interpretations view is that we live in an infinite multiverse, in which even an interpretation of an object that describes a complete world implies its existence. In this view, everything is merely implied by an interpretation of something else and our local universe, with its space and time, is merely an interpretation of some underlying system. Some readers will notice a resemblance of the argument in this series of articles to the “dust” hypothesis in Greg Egan’s science fiction novel, *Permutation City* [6].

The purpose of this article is to provide more justification for the proposal. This will mainly involve defending it from various objections which have been made or can be imagined. A thought experiment involving gradual reduction of resolution, or detail, of an observer will also be discussed.

We will now start by considering the first kind of objection.

2 OBJECTIONS ABOUT THE MEANINGLESSNESS OF SAYING THAT YOU ARE “IN” A PARTICULAR COMPUTER

Objection 1: If your mental experiences could be due to software running on either of two machines, and there is no way of distinguishing between one machine and another, it is meaningless to ask which machine you are “in”, making the entire discussion of probability when combining computers in the previous articles meaningless.

Answer

People may disagree on whether it is really meaningless, but this can be easily dealt with, as the author has pointed out, by introducing a slight difference between the programs, so that your expectation of some future event is different depending on the program. For example, we

could imagine one of the earlier thought experiments about combining computers as follows:

You are having the experience of standing in front of a box. When you open the box you will see either a red ball or a blue ball. Computer A is running a program corresponding to your mind in a virtual reality in which the ball is red. Computer B is also running a program corresponding to your mind in a virtual reality in which the ball is red. Computer C is running a program corresponding to your mind in a virtual reality in which the ball is blue. The computers have been made as similar as possible in every respect – other than that of the colour of the ball which you will see when you open the box. You do not know whether or not your experiences are due to Computer A, B or C and you have no way of choosing between them. There seem to be two situations in which you will see a red ball and one situation in which you will see a blue ball, so your initial probability of seeing a red ball should be $\frac{2}{3}$. Suppose now that Computers A and B are combined, as previously discussed [1]: the “FlatMech” machine architecture that the author described gives an example of how this could occur [2]. Eventually, you should be in a situation where Computers A and B have become the same machine. Your problem now is that it might look as if you only have one “red ball” situation and one “blue ball” situation, suggesting that your probability should be $\frac{1}{2}$ for each ball colour, but this result is incoherent, because deciding when two computers have been combined is arbitrary, as was shown in previous articles. The only way of resolving this is to adopt a position of statistical substrate dependence, which in turn needs the many-interpretations view to make sense of it.

Describing the thought experiments in terms of probabilities of different outcomes means that it can make sense to ask “What computer you are in?” and any objections to this can easily be met with some situation like this.

Objection 2: In the earlier thought experiments that you gave, you described situations in which identical programs, running your mind in a virtual reality, existed on different computers. As the programs were identical it could be argued that either could be a representation of your situation. When told that it is meaningless to ask which computer you were in, you have modified the

situation to involve different programs. In your above example, one version of you is going to see a red ball when you open the box and one version is going to see a blue ball. This means that, at the very least, there must be enough difference between the two programs to allow that. For example, some variable in each program, indicating the colour of ball that will be seen when the box is opened, would contain either the value “RED” or “BLUE”. This difference between the two programs means that it is not the same as your previous thought experiments and, to try to make it seem meaningful to say which computer you are in, you are cheating.

Answer

One response to this, and not the best one, is that in the above situation with the red and blue balls, there is no difference between one computer and another with regard to simulation of *you* – until you look in the box to see what colour the ball is. There is no need even to simulate the ball itself until you look in the box. Before you look in the box, all that is needed is some variable indicating what colour the ball is going to be. When you open the box, the program code can use this variable to start simulating the ball. The point is that, until you have looked in the box, there is only a minor difference between the two virtual reality simulations, but this difference has had no impact on your experience and there is no difference at all between one version of your mind and the other. Each version of you can be having exactly the same mental experiences and going through the same mental states. If you consider it important for the programs to be identical then, in what you should presumably regard as the most important respect, they are identical – until you look in the box. If you think that it is unreasonable to just compare the parts of the program running your mind, while ignoring parts of the program simulating the environment, then you are getting into an incoherent argument. How far do we go on this? Do we just look at the software running you and the simulated environment, or do we look at any of the software on which this is based? Do we compare the operating systems of the two computers? A problem here is that you do not reach a point where software suddenly becomes hardware.

A better answer to this objection, however, is to reply that it is based on a straw man of the

argument which relies on consideration of identical programs on different computers. The thought experiments in the previous articles involved considering identical programs corresponding to your mind in a virtual reality, but this identical nature was not an important feature of the thought experiments beyond its use in allowing parts of computers to match up when they were combined. The requirement was to have different versions of you which could all be possible candidates for your current situation, so that you knew that a number of computers are doing various things, you knew that some of these things are possible candidates for your situation, and you knew that these possible candidates would therefore be a reference class. In this sense, two programs, on two different computers are identical for practical purposes if they could both be candidates for your mental situation – if you do not know enough about yourself to rule either out.

3 THE OBJECTION BASED ON “INSTANTIATION”.

One objection is that minds, algorithms and various things that we might consider abstractions only become “real” when they are “instantiated”. The objection is basically that all these abstract things are not real because “they have not made themselves evident in the real world” or “they have not entered the real world”. This kind of objection may be used against the general argument that the author has been giving here, but it is particularly likely to be used against the probabilistic version of the progressive encryption of minds scenario described in the previous article [5]. We will now deal with this objection.

Objection 3: To exist, a mind must be *instantiated*. This is an obvious issue with the probabilistic version of the progressive encryption of minds scenario. In this kind of thought experiment, the encrypted minds do not exist because they are not instantiated. Until they are instantiated they are not real. When they are instantiated they are no longer encrypted anyway.

Answer

There are two replies to this objection. The first reply is based on arbitrariness, and will be a familiar one, and the second is based on the

problem of instantiation, as the idea is being used here, having no basis.

Reply 1: The problem of arbitrariness

The objection is assuming that something called “instantiation” is happening without saying exactly what it is. The whole problem is that, in the absence of a many-interpretations type view, it is hard to say when something is instantiated and when it is not. The many-interpretations view, or something similar to it, should seem forced on us due to the problem of arbitrariness, and this objection is just an attempt at hand-waving the problem of arbitrariness away.

Someone making this kind of objection may say that a deeply encrypted observer is “not instantiated” or that an observer or other object that can only be “found” by a complicated interpretation is “not instantiated”, but unless we can be told how to draw the line, the objection is incoherent. It may seem tempting to say that very abstract things are not “evident” in reality, but how do we decide that? If we can apply some interpretation to find some object, or the inputs/outputs of some computer program, then we can argue that they have become evident to us – that they are in fact instantiated. The objection would have us believe that there is “real” instantiation and “pretend” instantiation – and we are not told how to tell the difference, the idea that what constitutes “real” instantiation is obvious being implicit in the question. The actual argument about what is needed for instantiation could relate to one of the objections previously discussed, or it could relate to a new one, or the individual making it may not have a clear idea: we may not be told.

Reply 2: The problem of instantiation having no basis

Suppose, for the sake of argument, that the problem just stated with this objection has some resolution: there is some kind of cut-off point which allows us to say that some things are “instantiated” in the world – they are apparent to us – and other things are mere abstractions. We might say that an object found by some extreme interpretation is not real because it is “not apparent in the everyday world”.

The problem with this idea is that it defines what is instantiated with reference to the everyday world – things need to be apparent in the

everyday world to be instantiated – yet there is no explanation of why the everyday world itself is instantiated out of all the other things that might be instantiated. A thing seems to get defined as being “instantiated” if it interacts in some kind of acceptable way (which we are assuming can be defined in some way that is not arbitrary!) with other things that are already instantiated – yet this does not tell us how these other things get instantiated in the first place. The whole ontological structure – the whole idea of what is real – seems to rest on itself. Essentially, each object in the everyday world is in the everyday world because it interacts with other objects that are in the everyday world – and each of these other objects is in the everyday world for the same reason. The everyday world seems here like an exclusive club to which you can only gain entry on the basis of a recommendation from an existing member – so how does the club get started? We might say that the everyday world is readily apparent to “observers” like ourselves, but this is also circular: we need to know where observers can be found.

As an example of what is meant here, some extreme interpretation might find some object, but the reality of that object might be dismissed on account of it somehow being “too far” from the “everyday world” – on account of it not being instantiated. Suppose, however, that you are standing in front of a table and you see a robot, controlled by a computer program, collide with the table, causing it to shake. It will be obvious that the robot’s program is instantiated: it just caused a collision between the robot and the table – but we can imagine some object found by an extreme interpretation interacting in such an “obvious” way with another object found by an extreme interpretation. This would presumably be considered invalid because the object needs to interact “meaningfully” with something in the real world – like the table. What, however, causes the table to be part of the “real” or “everyday” world? You might say that the table is part of the everyday world because the robot can hit it – but the robot is getting its reality from its ability to hit the table! You might say that both are in the real world because, you, the observer, can see both, but then what makes you real – the fact that you can interact with the robot and the table?

To get more of an idea of why all this is a problem, let us imagine that the “real world” was

not actually real, while much more basic things on which it is based were. What we call the real world has a story of interpretation underpinning it. We can describe how trees, tables and chairs result from simpler things arranged in certain ways, and how those simpler things are emergent from still simpler things, and so on. Suppose things had been different and that all these interpretations were actually the “wrong” ones. Instead, the “real” world was built from the same underlying reality, at some level, but by a different sequence of interpretations, so that a different “real world” with different observers in it resulted. To these observers, our world would not be the real world. If anyone suggested the reality of our world, it could be claimed that the things in it were not “instantiated” – that they are not evident in the “real world”. The question, here, is: what caused what we see around us to be the “real world” instead of a different “real world” found by different interpretations? Without an answer to that, the whole idea of “instantiation” is circular. It should be apparent that it will be a difficult question to answer – and the author suggests that it cannot be answered satisfactorily. We can hardly make an appeal to some physical process that makes one world real rather than another: all the processes themselves are supposed to be features of the worlds in which they are occurring, and so exist at a level above where they are going to be of any use in resolving this. The answer should be clear: demands for “instantiation” resolve nothing.

4 A THOUGHT EXPERIMENT ABOUT RESOLUTION

Suppose that your brain had been scanned at a very high level of detail – with even single molecules being imaged – into a computer and the original brain had then been destroyed. The scan of your brain has been used to make two computer simulations, Simulation A and Simulation B, running on two computers. Both simulations are identical and very detailed, representing the motion and interaction of individual molecules and involving very small time-steps. The computers work in the same general way and apart from possibly running different programs (which has not happened yet) are otherwise identical. Your mental experiences right now are due to one of these computers. Are your mental experiences due to Simulation A or Simulation B? There is clearly no way of

knowing, and the probability of being “in” each computer is therefore 0.5.

Suppose now that, with you still not being sure which computer you are “in”, the level of detail in Simulation B is very slightly reduced. The simulation might statistically represent small groups of molecules, or it may model the system using slightly longer time-steps – or it may do both. Has this changed the probability that you are in Simulation B?

Suppose now that the level of detail in Simulation B is *progressively* reduced in a series of tiny changes. Each change does not have to be a change in the level of detail for all of Simulation B at once. The level of detail with which a single brain cell is represented might be reduced very slightly, and then the same reduction in detail might be made with another brain cell, and so on. After the level of detail has been reduced over the entire “brain”, a further reduction can be applied in the same way. Eventually, individual brain cells will no longer be represented in the simulation. A similar approach can be followed for extending the durations of the time-steps.

The process by which resolution is lost can be made to occur in almost imperceptibly small steps, and yet, after enough loss of detail, something can be left that few people would recognize as a simulation of a human brain. If nothing stops the reduction in detail, the ultimate end to all this would be a brain simulation consisting of a single binary digit, say “1”, and the idea would be that “This describes what your brain is doing for however long that it is supposed to describe it.” Is this a ridiculous idea? Of course, but the problem is that *we can get to it by a sequence of apparently insignificant changes*.

As the resolution is gradually reduced, does the probability change? If so, does it change suddenly or gradually? It will be apparent that this is the same issue that we have been discussing previously. Saying that the probability will *never* change as the resolution is reduced would clearly be wrong: it would imply that the final version of Simulation B – in which you have just been reduced to a single 0 or 1 – is just as likely to be a candidate for you as the high-resolution Simulation A.

We might think that the probabilities start at 0.5 for both Simulation A and Simulation B, and then at some point the probability for Simulation B abruptly becomes 0 while the probability for Simulation A becomes 1. If we keep reducing the resolution in Simulation B, starting from a simulation so detailed that individual molecules in your brain are represented, we will eventually end up with a simulation in which a single number might describe an entire idea or emotion. At some point it will become untenable to think that Simulation B could be you.

Does the probability change abruptly though? This should seem untenable. It would mean that some version of Simulation B would be just as likely a candidate for your mental state as Simulation A and then, with some slight change this would cease to be the case. How could we say when that slight change happens? There is no physical effect or theory on which we could base any cut-off point.

Another answer is that the probability changes *gradually*, never quite being reduced to nothing, but getting very close to nothing as Simulation B starts to have very low resolution. This is what the many-interpretations view would say. You can always abstract your mental state from the computer running Simulation A or Simulation B, but as Simulation B becomes less detailed, the interpretations needed to “find” your mental state in the computer running Simulation B need to be more complex to make up for the substrate providing less detail. The number of such interpretations with a given length of interpretative algorithm is therefore reduced and the measure of versions of you implied by Simulation B, relative to the measure of versions of you implied by Simulation A, is reduced. In practice, of course, when the resolution of Simulation B has decreased below a certain amount we would no longer regard it as worth considering as a candidate for your situation.

One answer to this thought experiment is that it is meaningless to talk about being “in” Simulation A or Simulation B when your experiences are the same in both. This has already been dealt with earlier in this article as a more general objection, and the answers would be no different here. The most obvious answer is that it could be arranged for different experiences to be provided for you in future in Simulation A and Simulation B. In Simulation A you might be about to find a red ball in a box, and in

Simulation B a blue ball, and it would then be an issue of the probability of finding each colour of ball.

This thought experiment will be relevant in the discussion of some of the objections that will follow.

5 FURTHER OBJECTIONS TO WHAT WAS ARGUED IN PREVIOUS ARTICLES

Objection 4: Only *unique* programs are relevant.

Answer

This objection was made previously and the author has already answered it, but it will be dealt with here because it is likely to be a common objection. The idea is that it is fundamentally wrong to count copies statistically at all: the number of computers running the same program is irrelevant with regard to the probability that an observer is existing because of that program. In this view, if you had ten computers running Program A and one computer running program B, and you knew that your consciousness was due to either Program A or Program B, you would not consider either program more likely. All that would interest you is that there are two different unique programs and you have an equal probability (0.5) of being caused by each.

This is a flawed and incoherent view and can be answered in the following ways:

The view that only unique algorithms count suggests that you could have certain probabilities of being in certain situations, due to the activity of a very large amount of computer hardware or *a lot* of some substrate, and then all of these statistics could be disrupted by a very small amount of hardware intervening. For example, you could have one billion computers running Program A and one billion computers running Program B, such that Program A and Program B are both candidates for your situation. With no other information, you should assume you have a 0.5 probability of being in either Program A or Program B. However, if we now introduce a third, single computer running program C, which is also a candidate for your situation, the probability that your situation is due to Program A or Program B has dropped to $\frac{1}{3}$ in each case. The intervention of the single computer running

Program C has upset everything. This should not be very credible. In fact, this seems to imply a strange kind of dualism. Normally, a small change in some aspect of the physical world would not seem to imply that the entire characteristics of the world would change, yet this is what is being claimed here for minds, so they seem to be getting treated in a completely non-standard way. In this kind of view, we only need to be concerned if we have one copy of each version of your mind, presumably anywhere in reality at any point in history. We do not normally base our view of reality on whether at least one of something exists, existed or will exist, ignoring any other versions. The special status this affords minds should look suspiciously like dualism. John Searle has commented that people who take a computationalist view of the mind subscribe to a kind of dualism. While the author disagrees with John Searle in other respects, he has a point here – though the appearance of dualism is coming here from the way statistics is being handled.

A further problem is that trying to refute the many-interpretations position in this way actually needs to assume that the many-interpretations position is wrong to start with, in a way that may not be obvious. If you think you can count the unique occurrences of programs that are running on hardware you are actually claiming to have a formal method for saying which piece of hardware is running which program, to be able to get your unique programs in the first place. There are two obvious ways in which this can be shown to be a problem.

The first way is that when a computer program is running a mind we should ask exactly what is being simulated. This may seem trivial, but it is far from it. The particular digital interpretation of the system that we may make is only one particular interpretation. We may say that two computers are running exactly the same program, but that is only because we have decided to apply the rules which say that particular voltages, for example, correspond to particular symbols. In one computer there may be a voltage of 4.99974 volts in some electronic switch. In the other computer there may be a voltage of 4.99982 volts in the same switch. It is a particular interpretation that says that these correspond to the same digital operation and, therefore, the same computer program. If anyone claims that this is clearly the case, and that two computers, in slightly different physical configurations, are

running the same computer program because it is some kind of objective truth – that is to say we can round off the physical variables in the system in some way – then the author would give you the following problem. Suppose we have two computers, Computer A and Computer B, which are running what most people would regard as the same computer program, except of course that a slight difference between the computers means that there are slight differences in the physical implementation, such as the behaviour of individual electrons in components, or the positions of components, or signal voltages, etc. Someone trying to use this objection would say that these two computers together just represent one unique implementation of the computer program. Now, suppose we write a computer program which simulates the behaviour of Computer A in great detail, down to the level of actual signal voltages, or even individual atoms or electrons. We make a second computer program which simulates the Computer B, again with a huge amount of detail. We now run each of these programs on a separate computer.¹ We now have two computers, each running a different program, each program simulating the physical behaviour of a different computer, but with each of the two computers being simulated running the same program. If you maintain that it is only individual implementations of a program that count, you should have no problem here saying that these would count as individual, unique programs: each program is different. The programs have been made different now, because what were previously just slight physical differences between the two computers running the program have now been made features of the computer program itself. If an atom were in a particular place, for example, in Computer A, that is now reflected in the program simulating Computer A. If a particular voltage were 5.9997 volts in some component in Computer B, then that is now reflected in the program simulating the Computer B. The two programs simulating hardware of the two computers are different, and should count as unique programs. If either of these programs were possible candidates for your mind, you should presumably consider it equally likely that you are being caused by either one of them. However, this should give you a problem.

¹ Actually, having them on separate computers should not really be necessary for anyone, here, but it may help with clarity.

The distinction between hardware and software is actually artificial: it only arises when humans attach symbols to variables describing the physical configurations of the computer system. There is no reason why we should use the particular interpretation that we have been doing. Any computer can be regarded as a simulation of itself. This means that if we regard a computer program simulating the behaviour of a computer as being a unique occurrence of that program, then we should also regard the computer being simulated itself as a unique occurrence of the same program. That is to say, if we have two computers, each slightly different physically, running “the same program”, we should be quite justified in saying that the first computer is running a program which simulates the entirety of itself and the second computer is running a different program which simulates the entirety of itself.

We could also do what has just been discussed in a more general way, so that a computer might be interpreted at various levels of detail, with what was previously regarded as a single binary digit being interpreted as a number of binary digits, the values of which depend on whatever physical value (such as the voltage) was involved. One interpretation might mean that a single bit becomes five bits, another than a single bit becomes twenty bits. This could also be done with the time-steps between states, so that we now take the computational states and find extra states between them. If done with enough detail, this really amounts to what was just discussed – regarding the entire computer as running a simulation of itself – but the point is that this can be done with varying levels of detail – and the level of detail at which we look at “the program that is obviously there” in terms of 1s and 0s is just one of these.

A similar kind of problem arises when we realize that we can look at a computer system “containing” 1s and 0s and combine some of these. We might group the bits in a computer into groups of three, so that each group of three 0s and/or 1s is regarded as representing a single 0/1 value. This would amount to applying an interpretation which reduces the *resolution* of the program. A similar thing could be done with the time-steps. A computer might have a state change every one billionth of a second, but we could say that the *real* state change occurs every thousandth of a second – and we could get each such state by combing all the information in all

the states occurring every one billionth of a second during it. Doing things like this, two computers that previously appeared to be running different programs could appear to be running the same program. This is similar to what was just discussed – looking at the actual values of things like voltages and looking at states in-between states – except that it is now being done the other way round.

The problem should be obvious now, and should be familiar. Just as computers do not have well defined edges, computer programs themselves do not have well-defined edges either. If we think that only unique computer programs should be counted we have the problem that whether we have one computer program or two computer programs depends on the level at which we decide something is a computer program – whether we look at the level at which programmers have been working or whether we regard the entire physical system, described in exhaustive detail, as corresponding to a program.

There is one answer that someone making this objection could give that should be examined. Suppose we have a computer running some program and, as has been discussed, we apply some interpretations that are not very extreme and which give slightly different algorithms: what were previously individual bits have now become multiple bits and what was previously a transition from one computational state to another has now become a sequence of transitions through multiple computational states. It could be said that all these programs will still be “running close to each other” in a way. This is to say that there will always be some consistent relationship between them, no matter how long they run. For example, if we apply two interpretations to a computer that “find” extra states in between what would be the conventionally accepted computational states, it could be argued that the two programs found in this way drift apart while in these extra states, but they will never get too far away from the conventionally accepted states of the program. One problem with this answer is that simply say, “So what?” Just because there may be a lot of parallelism in the behaviour of two objects over time, it does not follow that they are the same object. We can also raise the issue of a cut-off point here. How close does the relationship have to be?

The second way of looking at this is closely related. It also involves focusing on the fakeness of the distinction between software and hardware and also involves the issue that we cannot say where the edges of a computer are. Suppose we have two computers, Computer A and Computer B, running the same program. Computer A is on a red table and Computer B is on a blue table. From one point of view we could say that each computer is running the same program and that therefore we only have one unique program involved. From another point of view, however, we could view the table and the computer together, in each case, as being the computational system. From this point of view Computer A and the red table *together* – both the computer and the table – can be considered to be running a computer program that simulates a computer on a red table. Computer B and the blue table together can be considered to be running a computer program which simulates a computer on a blue table. The differences between these two computer programs mean that we can regard them as unique programs. It should be apparent here that this is arbitrary. It should also be apparent that the colours of the tables were a crude way of getting an obvious difference. It was not necessary to use coloured tables. It was just to make the point this is very closely related to the previously mentioned case, except now, as well as just using physical differences between the computers and arguing that those could be interpreted as corresponding to different computer programs, we are using the physical environment around the computer, based on the idea that there is no natural edge to the computer.

Some readers might ask what would happen if the two computers were absolutely identical. The author might try to claim that this would not practically be the case, or would not happen in the kind of situations we are considering here, or the author may try to use the uncertainty principle. In fact, we need use neither. The second way of looking at this in which we considered the computer and the environment outside the computer as implying the running of a program which describes their behaviour means that two different computers could never possibly be in the same situation when looked at in this way. Even if the computers and the immediate environments were absolutely identical, all we would have to do was keep looking at the environments progressively

further out from the computers until a difference was found.

There is one way you could answer this. You could say, “Okay, let us only count individual programs, but we will do this in exhaustive detail so that every particle of the computer and everything about the situation around counts as part of the computer program.” If you do this, and think this is a response to the many-interpretations approach, you have a problem, because *by doing this you are heading very much down the path of throwing away strong substrate independence yourself*. When you have started doing this you are regarding strong substrate independence as pretty much dead, and you have actually adopted the very position you were trying to argue against. You have adopted a position in which effectively copies do count, because in accepting every small physical detail as creating a difference between one computer program and another you have accepted a situation where any two versions of a computer program, by looking at them in a more detailed way, would end up as individual programs: you are counting copies. That is the problem when faced with issues like this: The position of only counting *unique* programs is incoherent and if you try to deal with the incoherence you actually destroy the position by taking away any idea of uniqueness from it. Issues of consistency and relationship and between programs go away if we allow more extreme interpretations that find programs that can diverge from each other.

Against this, it could be argued that we should only count unique programs, but that this does not apply for programs that describe a mind, the computer on which it is running and the physical environment: instead we should just focus on the “abstract description” of the mind, counting unique versions of such things to the exclusion of everything else. This raises what may not be an immediately obvious problem: how do we draw the line between program code that relates directly to a mind and program code that relates to things “outside” it? The cut-off point may not be obvious. Also, again, this seems to be giving minds such different treatment to everything else that it has the appearance of a kind of dualism.

Even if we accept the idea that we should only count unique programs, it will not always be obvious how *many* unique versions of you exist on a given substrate. For example, with the thought experiment in which an observer is

progressively encrypted [5], we would imagine different programs which can all be found by various interpretations of the substrate, all of them being possible candidates, as far as the observer knows, for his situation. With the thought experiment about resolution, described earlier in the article, in which two versions of you exist and the resolution of one version of you is gradually reduced, there is the issue that the two programs are not the same: the lower resolution version of you is different. Against this, it could be argued that the lower resolution simulation still “represents” the higher resolution simulation in some sense, but if it can do this then it could also stand in for many slightly different versions of you which are all candidates for your situation, and it will not be obvious how many such versions of you there are: rejecting the idea that we can simply declare the low-resolution simulation and the high resolution simulation different forces us into this situation by demanding that we compare what the programs represent. What has been said so far amounts to this, really, but this is the most general expression of it: you can look at a computer and apply various interpretations that “find” different programs. You may have two computers with different hardware, running what appears to be the same program, but various different interpretations might be applied to find different numbers of unique programs, all of which are candidates for your experience, in each computer. Some of these programs might be similar, but others may be very different: while they may all agree with your current experience, they may diverge substantially afterwards.

Objection 5: The objects found by extreme interpretations are just being “made up” by the interpretative algorithm.

If it is possible to “find” any object you want by using a sufficiently complex interpretation, an obvious objection, and one that is typically made, is that the object being found is not really there, but is instead being “made up” by the interpretation. If we imagine actually running the interpretative algorithm, it is obvious that for such extreme interpretations the information in the object being found would be coming from the interpretative algorithm, rather than from the object to which is applied, so a “common-sense” view says that the object being “found” is really just being made by the interpretative algorithm.

This seems to be wrongly assuming that there is some kind of causality involved in how objects get here. From a causal point of view, the information in the interpretative algorithm is playing a large part in “causing” the information in the “found” object to be produced, so it may sound like a convincing argument that it is the interpretative algorithm that is causing it. The problem with this is that it is entirely wrong to mix up the causality involved with the process of “finding” the object – of producing some representation of the object that we can look at it ourselves. People do not see this when they imagine the interpretative algorithm actually being run somehow and producing some representation of the found object. They can see that this representation is being produced by the interpretative algorithm and the fact that the interpretative algorithm needs a lot of information to do this makes this idea stronger. This is only a representation of the found object, however, and the actual object exists independently of any sequence of causation involved in finding it.

Some people will reject this, however, insisting that the objects “found” by extreme interpretations are merely “made-up”. To such people the author would ask, “Where is the cut-off point supposed to be?” When is an interpretation reasonable and when is it unreasonable? If the amount of information in the interpretation is an issue – and that seems to be a concern of many people – when does the amount of information in the interpretation become too much? When does it start to become “cheating”?

Some people may answer this by saying that there is no cut-off, but that objects progressively become less real in some sense, but that in itself is not too far away from the many-interpretations view anyway. “Less real” in itself is not a very coherent idea and the many-interpretations view’s concept of “measure” does better. The many-interpretations view does not claim that all objects are equal, as some critics have presented it, but instead recognizes that objects that can be found by shorter interpretative algorithms will tend to exist with larger measure.

Ultimately, this objection is not attacking what it should be attacking, which is the entire approach of viewing things in terms of interpretations and measure, but is instead just nibbling away at the edges of the issue by trying to attack extreme

interpretations. It fails because there is no clear concept of what an “extreme” interpretation is and why it should be invalid, and this whole approach is not touching the main ideas in the many-interpretations view themselves, but is rather just being incredulous about the more extreme examples, and these extreme cases are not what the many-interpretations view is about.

Some people use a more specific form of this objection, in which they insist that the interpretation becomes “cheating” when the amount of information needed in the interpretative algorithm is equal to the amount of information in the object being found (or exceeds it – depending on who is making the objection). This form of the objection will now be discussed.

Objection 6: If there is more information in the interpretative algorithm than in the object being “found”, then nothing is really being “found”.

Answer

People making this objection view a finding algorithm which contains more information than it actually finds as automatically invalid. This objection has already been answered, but because it is one that people still make after reading the articles, it will now be dealt with again.²

There is no obvious cut-off point and there is nothing special about reaching the point at which the amount of information needed in the finding algorithm is as much as that present in the object itself. People who think that there *is* something special about reaching this point seem to miss this. If we keep increasing the information in an interpretative algorithm needed to find some object, nothing philosophically profound happens when the amount of information in the interpretative algorithm becomes equal to that in the object being found. It is an arbitrary point that only *seems* important to some humans.

² In fact, and this surprised the author, one critic even said that the author had missed this – as if the author were unaware of the “problem” that allegedly arises when the finding algorithm contains as much information as the object being found: It is actually one of the first things you would think of!

To try to get this idea across, the author would ask you to imagine that we have some object containing N bits and we find it by applying an interpretative algorithm, containing F bits to some object described by O bits. If F is much less than N ($F \ll N$), few people would have a problem with the idea that the algorithm had genuinely found something. Some people may demand that the object being found has to be non-trivial in some way for a statement of its existence to be relevant, but we can easily assume that we are finding non-trivial objects. The point is that, if some object can be “found” by applying a finding algorithm containing a much smaller amount of information than is in the object to some other object, and if the object interests us sufficiently (and that is a somewhat subjective condition, which the author does not consider necessary), it would seem reasonable to say that the object really exists.

Suppose now that we increase the amount of information in the interpretative algorithm, F . Provided still that $F \ll N$, most people, again, will have no problem: the interpretation will not seem contrived. At some point, however, F will be regarded as too large relative to N . The finding operation, the interpretation, will then seem to be contrived and people will start to declare it invalid. The problem is that where this happens is arbitrary. One person may say that we reach this point when $F=0.5 \times N$ and another that we reach this point at $F=0.75 \times N$. Why should the situation of $F=N$, when the interpretative algorithm is as long as the algorithm being found, be special? If the interpretation somehow becomes invalid when $F=N$ what about when F is slightly less – when the interpretative algorithm is *almost* the same length as the algorithm being found? It should seem absurd to suggest that one algorithm or other object with 1 billion bits in it exists because it can be found by an interpretative algorithm containing 999,999,999 bits but another with 1 billion bits in it does *not* exist when it is found by an algorithm containing 999,999,999+1 bits. If you say that an object ceases to exist when the amount of information in the finding algorithm is the same as the amount of information in the object, by what authority do you say this? What philosophical principle is being invoked? This is nothing more than application of some kind of common sense, and this common sense, if valid at all, would be equally valid if the interpretative algorithm had one bit more or one bit less.

Objection 7: Thought experiments in your argument have contrived the existence of different versions of you, with different futures, and have asked questions about probability. If there are identical versions of me, with different futures, this does not mean that I have to accept a many-interpretations type view to be able to make coherent statements about the probability of future outcomes. Instead, I can say that the probability of future outcomes for me is undefined. In your red ball or blue ball examples there is no probability because I have more than one future. If I am prepared to declare the probability of my future expectations as undefined, I have no need of some view like the many-interpretations view to make probability coherent.

Answer

The idea that probability values relating to your future situation could ever be regarded as undefined should appear immediately absurd – even if there are multiple versions of you. However, someone did actually make this objection, so an attempt will be made to answer it.

As with other objections, this objection is not really trying to deal with the main idea of many-interpretations. It is instead trying to deal with specific thought experiments in which “identical versions” exist with different computers. An example is the thought experiment about combining computers. The objection does not attempt to deal with the more general problem of arbitrariness of interpretation, of which the results of the strange thought experiments are really just a symptom.

One obvious problem with the objection is that the different versions in thought experiments like this do not have to be *absolutely* identical. They merely have to be similar enough to create a situation which does not allow you to be sure about your situation.

Let us consider the thought experiment about combining computers [1,2] as an example. With two computers, Computer A and Computer B, you are going to see a red ball and with Computer C you are going to see a blue ball. Computers A and B are then combined, and the issue is one of what happens to the probability. Clearly, the versions of you running in Computer

A and Computer B have to be very similar to allow combination. However, these two versions have the same future: seeing a red ball. The other version, which is going to see the blue ball, is running on Computer C. How similar does the version of you in Computer C have to be to the versions of you in Computers A and B? Although it is easiest to think about this in terms of a high degree of similarity, in fact the similarity does not have to be very high at all. It depends on what you know about the versions running on each computer. If you know that you must be one of the versions on these computers, then you may know nothing about the different versions other than that they are minds which could be like yours – and they may not even all match your current experience. All that is required is that the version on Computer A and Computer B matches your experience or the version on Computer C matches it, and you would have the uncertainty about where you are and what colour of ball you are going to see. The version on Computer A and Computer B may just be a human mind in general that is about to see a red ball, and the same may apply for the version on Computer C. Despite this, we will still answer this objection further.

Probability values merely describe what is known about your situation, and if those values are ending up undefined then this should be an immediate warning sign that your philosophy has gone astray – that it is not offering a coherent approach for dealing with reality. Such a view would actually reduce the world to an incoherent mess. It would not only mean that it makes no sense to talk about the statistics of your possible future: it would be highly questionable whether the concept of “your future” would mean anything – and we could always use the philosophical trick of framing questions about probability in terms of wagers to force a decision.

This kind of situation is actually implied by various theories. A “large” enough cosmos in terms of time and space would imply the existence of different versions of you. The many-worlds interpretation of quantum mechanics [7,8] would imply that you are “splitting” many times in each second. An argument has been made by Aguirre, Tegmark and Layzer that the inflationary model of the universe implies that versions of all the many worlds in the many-worlds interpretation exist here, in this world, spatially separated from us anyway [9] and a

related argument has been made by Bousso and Susskind [10]. Any of these scenarios would imply that multiple versions of you would exist. If you think that probability requires the non-existence of any other versions of you to be coherent, then you should have serious concerns about the coherence of probability anyway, because it would seem easy for it to be threatened – either by someone making copies of you or by some scientific theory or another turning out to be true. In reality, issues of probability seem to be fairly easily resolved in scenarios like this. The idea of probabilities of different futures *from the point of view of an observer* is generally regarded as a reasonable way of dealing with the many-worlds interpretation, and the idea that the many-worlds interpretation would make probability incoherent would be generally regarded as bizarre, yet that is basically what is being suggested by this objection.

Suppose someone subscribed to this view of copies making probability incoherent. The idea of considering the probable merits of different courses of action would be incoherent. Such a person would have no basis on which to run his life. Suppose this person was then confronted by undeniable proof that someone had made copies of him, or that the many-worlds interpretation of quantum mechanics is correct. Would such a person immediately disengage from the world, or start acting incoherently or randomly – or do whatever it is that people are supposed to do when they suddenly realize that probability is incoherent? We should, in fact, have a hard time even imagining what the appropriate behaviour *is* when you find out that the idea of probability is incoherent: even starting to behave randomly would imply some kind of decision based on *something*. The reality is that the individual would continue functioning in much the same way, as if some statements about the future were likely and others were unlikely, given various courses of action. That would be an indication that the view would not be seriously maintained in a situation like this. The same would apply when we considered the suggestion that probability is incoherent with the multiple copies of the thought experiments in this series of articles.

A further argument against the idea that the existence of different versions of you would make probabilities undefined can be made using an approach similar to the “mixed wager”

refutation of Pascal's wager [11]. Suppose there is just a chance, no matter how small, that a copy of you exists in a situation like yours right now, but with a different future, or that you have more than one future. This might be because the many-worlds interpretation is correct, or because some evil scientist has managed to make a copy of you. The evil scientist need not even be particularly good at making copies of you: even if he is just guessing about how to make the copy, there is a chance that he will happen to get it right. In fact, to rule this out in principle we would have to be sure that it is *impossible* that some process like evolution, the actions of aliens or future humans, or even the random motion of matter, could somehow create a different version of you with a different future. If there is any non-zero probability – no matter how small – that you exist in different versions with different futures, or that you have different futures – and if you think that such a scenario would imply that probabilities about your future are undefined, this would imply that there is a real probability that probabilities about your future are undefined. It does not matter that the probability of your probabilities being undefined may be extremely small. The mere fact of its possibility – that the probability is there – gets it into probability calculations about your future situation that you perform *now*. Any probability calculation about your future situation would have to take account of this probability that the probabilities for your future are undefined, and clearly it cannot: the mere possibility of this, even in principle, would make all probability calculations about your future meaningless, and all probabilities relating to your future would be undefined. If you think that the objection makes sense, then to be consistent, you should regard any talk of probabilities about your situation as meaningless now – and in that case, it should be hard to imagine planning even the simplest thing.

Objection 8: Any mind you found by an extreme interpretation of some random system, done as John Searle describes, would only last while the random behaviour matched the particular interpretation being used. It would be transient, only existing for a fraction of a second. Such a mind need not be considered of any consequence.

Answer

This objection has been made by people. The argument that has been made in this series is not

primarily aimed at proving some kind of animism – that minds live in chairs, sandwiches and walls (as in John Searle's example in his argument against strong artificial intelligence [12]) because they can always be found by the appropriate interpretation algorithm. The argument is about the relationship between mind and substrate, and the argument has dealt with combination of different computers and similar kinds of issues to show that, at the very least, some kind of weak substrate dependence must be admitted. John Searle's argument about universal realizability is not the general argument. It is only one argument that has been used to show that the general problem – arbitrariness of interpretation – exists. This objection does nothing to deal with other arguments that have been used, such as the thought experiments about different substrates from the first article [1].

The argument about John Searle's issue of universal realizability has been used to point out that the problem of distinguishing between valid and invalid interpretations is not one with an obvious solution: there is no point at which the interpretations suddenly become too extreme to be invalid for some hard, philosophical reason. This leads us into the idea that all of the interpretations should be admitted, so as to avoid any arbitrary selection of interpretations, and we then have all these minds existing with varying degrees of measure relative to each other. None of this is really about minds existing in random motion of matter, but it is instead about where we define the edges: where we say, "Here are the valid interpretations that relate to minds." By extension, unless we want to make a special case of minds – and go over into some kind of dualism – this forces on us the problem of where we define the edges of things in general.

At the same time, the author must admit that the argument does suggest that minds exist in such strange places; however it also suggests that they exist with very low measure relative to that of minds in more conventional situations. This is just an extreme consequence of the argument, and arguing against it is not properly arguing against the argument. If you win on this one, it is likely that you will have merely shown that what might have appeared to be an extreme consequence of the argument is of no real interest. You would still not have addressed the issues created by the scenarios involving such things as combining minds, and you would be unable to say where the edges are.

The objection is that if we use some extreme interpretation to “find” a mind in the random motion of matter in, say, a wall, it will only find the mind for an instant at which the motion of the matter matches the interpretation. We might say that a mind existing for such a brief period is not “real” in any meaningful sense.

A problem with this is that it merely raises the issue of how long a mind must exist to be real: it is attempting to impose an arbitrary cut-off point. We might admit that such minds are real, in some philosophical sense, but that their brief existence does not allow them time to experience anything as we experience it: such a mind may not have time for what we consider a “proper” thought. On these grounds, we might consider such minds to be of no more practical interest than a single atom from a heart would be to a cardiologist. How long then does a mind need to last for it to be considered real? One millionth of a second? One tenth of a second? One second? One minute? Part of the argument for many-interpretations has been given in statistical terms, with a demonstration that probability becomes incoherent if we do not accept other interpretations. It might be thought that the minimum time needed for a mind to be said to exist is the time needed for it to exist to be a candidate for your situation in one of the thought experiments in this series, but how long is this time? This objection is doing nothing but arbitrarily declaring the problem to be disposed of.

Some people who raise the objection of “transient minds” will also subscribe to the idea of “mind uploading” – that a mind’s consciousness can be said to continue if the mind is copied into a different substrate. If this is the case, then we should be able to imagine an extreme interpretation which produces conscious experience for a very short period of time, and another interpretation which also produces conscious experience for a very short time, but which takes up where the previous mind left off, and still another interpretation that could find a short-lived conscious experience and so on. In other words, if you think that transient minds would exist in some trivial sense, and if you think mind-uploading is possible, you seem to have the problem that the conscious experiences of a set of transient minds that you regard as trivial could be regarded as being combined in a sequence to form a conscious experience of non-trivial duration. If you reject this, then you seem

to be rejecting the idea that the conscious experiences of a mind continue if a continuation is made in some other substrate. One reply to this would be that each of the transient minds would be too short-lived for this to work – that mind-uploading only works when a significant amount of time passes in a single substrate, between transfers – but that just raises the issue of where the cut-off point is supposed to be. This issue does raise questions about continuity, probability and measure, and those questions are not answered by the discussion which has just been given, but they do not have to be: the point is that transient minds do not go away as an issue just by being transient.

Someone might actually try to make the complexity of whatever relationship exists between all the individual transient minds – how the continuity between all the different minds is supposed to work – an issue, suggesting that this would somehow give it low measure. That would be futile as no claim that minds associated with extreme interpretations have high measure is actually made: taking such an approach would actually imply some kind of acceptance of a many-interpretations view. Further, as soon as you start to theorize about the “mechanics” of continuation from one mind on one substrate to another, and to suggest that different continuations have different status, due to something about the substrate, the duration for which the mind exists on a single substrate before the transfer, the complexity of the continuation, how often continuation between substrates has to occur, etc. you are actually losing strong substrate independence, because you are having to take account of how things are implemented: you would be getting into a many-interpretations kind of view.

A further problem with the objection based on transient minds is that it is simply not true that all minds found by extreme interpretations of the random behaviour of systems would be transient. You could find a mind which lasts for longer merely by using a more extreme interpretation, and you could find a mind which lasts for still longer by using a still more complex interpretation, and so on. One answer to this that the author is aware of someone using is that the minds found in this way would somehow not be real as a result of progressively longer interpretative algorithms being needed to find progressively longer-lasting minds. This, however, is a metaphysical rule that has been

pulled from nowhere. On what basis is it decided that a mind does not exist if it has a limited lifespan and a longer-lived version of it can only be found by a longer interpretation? We should be able to accept the reality of a single mind, existing on its own terms, with a description of its experiences, for a given length of time without considering what would be needed for a longer-lived version of that mind to exist. Using some interpretation, we might find a mind that exists for one minute, and we may say that a longer interpretative algorithm is needed to find a mind that has the same experiences and then exists for more than one minute, but human minds only exist for a finite period of time. Some interpretation of the substrate on which your mind is based will find your mind, and describe it as existing for some finite period of time, but you do not say that your finite lifespan makes you less real.

We could also approach this issue by combining two of the ideas already used to answer the objection: mind-uploading to go from one transient mind's experiences to those of another and longer interpretations that find longer-lived minds. We might imagine a conscious observer who has mental experiences that lasts for a finite, but non-trivial period of time, and which are found by some long interpretative algorithm, and are then continued as mental experiences found by a different interpretative algorithm that also last for a non-trivial amount of time and so on. If we take a naïve view of mind uploading, there would be no sense in which interpretative algorithms would be becoming longer to do this: we would just be swapping from one to the next and so on. As the author said earlier, the idea of continuity from one substrate to the next in this situation raises statistical issues. In fact, it should become apparent that we are going to have to use some approach involving interpretation and measure to consider coherently the statistics of the transition from one substrate to the next. The point here, however, is that someone who rejects a many-interpretations kind of view really has to reject approaches like that – meaning that there seems to be little in the way of relying on “mind-uploading” type ideas to get continuity of any duration at all without using interpretations of arbitrary length. As has been said earlier, as soon as you start to theorize about the “mechanics” of continuation from one mind on one substrate to another, you are into a many-interpretations type

view – so such an approach will be useless for refuting the many-interpretations view.

So far, we have focused on answers to the objection that deal with the issue of transient minds and it has been shown that a mind being transient is not a problem.

The objection focuses on minds found by interpretations of random features of systems over time, but extreme interpretations might not rely on randomness. Instead, some physical system might imply the existence of some mind in a way that seems unusual, but is consistent over time. For example, if you imagine someone else's behaviour over a period of time, your brain clearly contains something from which the mind of that person could be derived by some extreme interpretation, and it should be possible to imagine the same kind of interpretation doing this over an arbitrarily long time-span. Likewise, some interpretation might be applied to the frames in a film, as each appears over time, to produce a description of the mind of a character in the film: the way in which this works could be quite consistent over time. The thought experiment about resolution, discussed earlier in this article, is relevant here. In that thought experiment, a version of you is simulated on a computer and the level of detail of the simulation is gradually reduced. As the detail of the simulation is gradually reduced, the interpretation needed to produce something that is a candidate for your mental state from the simulation will become gradually more extreme and require longer algorithms to represent it. However, this does not imply that finding a version of you which lasts for longer would require still longer algorithms: this is not like trying to find you in the random motion of the matter in a wall. What might seem to be a low-resolution simulation of you could be interpreted as a high-resolution simulation by applying the same kind of interpretation to it at different times. In fact, John Searle's thought experiment of finding a program running in a wall can be reduced to this, really: while trying to map the random motion of large numbers of atoms onto a program's state over a long period of time may seem to require an interpretation that deals with different behaviour of the wall at different times, we hardly need all these atoms. We could just do the mapping with a small number of atoms in the wall, and use an interpretation in which almost all of the information describing the mind's state is coming from the interpretation anyway.

The final answer that will be given to this objection is that the focus on minds found by interpretation of random features of a system over time is misplaced, because an interpretation might not be applied to some system varying over time in the same way that the mind is experiencing time. The interpretation might describe a mind's experiences varying over time, but be applied to some underlying substrate with the timescale not matching up. For example, the description of a mind experiencing a week's worth of thoughts might be derived from the behaviour of a physical system over a period of one minute. The interpretation may not even map instants of the physical system onto experienced instants of the mind being found in a one-to-one way. For example, the description of a mind experiencing a week of existence might be found by an interpretation of an underlying physical system at all kinds of times, and one instant of the mind's thinking may be taken from the substrate at many different times. The interpretation might not even be of anything varying over time at all. For example, the description of a mind experiencing a week of existence might be derived by an interpretation of an underlying substrate at a single instant. The description of a mind experiencing the "flow" of time can, in principle, even be produced by an interpretation of something in a world in which things cannot be described in terms of time. In fact, the many-interpretations view suggests that every feature of our world that we know is "constructed" and that "time" is merely a feature of our world that is found by interpretation of something that cannot be described in terms of time. Focusing on John Searle's "minds in walls" scenario is a distraction.

Objection 9: Who or what is supposed to be running these finding or interpretative algorithms?

Answer

The interpretation is not regarded as a process that actually has to occur. Rather, an object is regarded as being implied merely by the logical possibility of the interpretation that finds it being performed, whether such interpretation is actually done or not.

A misunderstanding of the argument in this respect is likely to be used by somebody to make a flawed argument for the existence of God. Such an argument for God would rely on the

idea that objects exist if the appropriate interpretation can find them, and therefore that the "process" of emergence needs an observer to make it work. The author has no belief whatsoever in such an idea.

6 OBJECTIONS BASED ON REDUCED AMBITION

Objection 10: Do we need to describe everything formally? Maybe some aspects of reality are beyond formal description?

Answer

This objection was effectively made by Steve Grand in a debate with the author [13]. A reply to this is that it is defeatist. If we are to say that there are aspects of reality that defy formal description, we are effectively admitting the existence of the supernatural. If we are going to take this approach, why bother doing *any* science or philosophy?

A further reply can be based on this, however. Is the concept of the "supernatural" in this sense of the world – of things beyond formal description – even coherent? A formal description of something is merely one with the vagueness removed. We could keep asking progressively more detailed yes/no questions about the world and, assuming that we get an answer to each, our description of the world is becoming more formalized. Even if we do not know the answers, if answers would be there to be known *in principle* our description of the world could be formalized. To deny this would be to say that the world could have some property which can be meaningfully said to exist, yet about which we cannot frame a coherent yes/no question. The claim for such a property should be viewed as incoherent.

Besides, as has been pointed out, all this can be reduced to issues of probability of various expectations in various experiments. If we admit that those probabilities are properly defined, then we are effectively admitting that our situation can be formally described. If we deny that those probabilities can be formally defined, even in principle, then we are making a strange claim. As is often done in arguments like this, we can force a detractor to estimate probabilities by offering a wager, and the detractor would have to use his best estimate of the probabilities. The detractor might claim that there is more to it than

this – that the situation has features which affect the probabilities in some “informal” way – but what does that even mean? Probability values are numbers, and to think about one is to formalize it.

Objection 11: Why should the question of what constitutes a conscious observer have a definite answer? Can we not just accept that there are shades of grey?

Answer

The objection made in this way is somewhat vague. Viewed simplistically, it seems to be implying that the whole idea of trying to get a firm answer on where observers are is flawed. The problem here is that, in a way, this is what the author has been proposing! If a substrate merely implied that an observer is here, or here, and not here or here, and that some object like a tree is here and not here, then the argument would not have got off the ground. The problem starts when we recognize the greyness and then try to do probability assignments in various thought experiments.

We might consider the objection as implying something else. It could be argued that consciousness is not necessarily an all or nothing thing and that an extreme interpretation that appears to find a conscious observer is actually finding what is effectively a very diminished form of consciousness. We might simply say that this tells us nothing profound about reality: we might consider all kinds of things to have very diminished forms of consciousness, and we may regard such things as almost meaningless. The problem with this is that thought experiments can be produced in which some of these other forms of consciousness are actually candidates for your own situation. We then still have the problem of determining what causes one mind to be a candidate for your situation and another not to be. If the interpretation needed to find a mind becomes progressively more extreme, and the mind becomes progressively less conscious, exactly when is it supposed to stop being a candidate for your own situation and experiences? If you claim that there is some cut-off point you need to say where it is – and the author has been arguing over a number of articles that you cannot do so. Alternatively, if you say that the mind remains a candidate, but merely becomes less likely as a candidate, you have essentially turned your idea of diminished

consciousness into one of non-diminished consciousness and diminished probability. Even if you start with a “shades of grey” view of consciousness, you will ultimately have to turn those shades of grey into a statistical concept, and that is all that the many-interpretations view, in so much as it talks about minds, is effectively doing.³

7 CONCLUSION

As has been argued in previous articles [1,2,3,4,5], a coherent view of probability requires us to take a measure view of consciousness, in which any interpretation of a physical system which describes the mental state of an observer is regarded as describing a real observer. That is to say, minds are implied by any interpretations of physical systems that produce them. This is the *many-interpretations* view [1,2,3]. For our approach to minds to be consistent with our approach to other things, this view needs to be extended to include all physical objects, so that any interpretation of a physical system which describes an object actually describes a real, physical object. An implication of this is that we live in a vast, general multiverse which seems to have some similarity with the *mathematical universe* proposed by Max Tegmark [14,15,16]. Even the space and time of our world should be viewed as emergent things, implied by an appropriate interpretation of something else underlying them [4].

Various thought experiments have been described to support the argument. In this article, a thought experiment has been described involving gradual reduction in resolution or detail. Two systems contain representations of your mind, and are candidates for your situation. The level of detail in one of the simulations is gradually reduced and the issue is one of how this affects the probability that it is causing your mental experiences.

³ It should be noted that the many-interpretations view does not say that one mind literally has a lower probability of being a candidate for your situation due to being found by a more extreme interpretation per se: rather, the need for a more extreme interpretation to find something implies that there will be fewer different interpretations to find objects of that kind, so the different versions of that mind will all exist with a lower relative measure.

Various objections have been made, or can be made, to the argument, and some of these have been dealt with.

One objection asserts that it is meaningless to distinguish between various situations in which you may be, but this can be answered by arranging for different experiences in the future in each situation.

Another objection is based on the idea of “instantiation”: it is asserted that objects described by extreme interpretations do not exist until they are “instantiated” in the world. This is answered in two ways: by pointing out that it does not remove the issue of arbitrariness but merely makes the arbitrariness about what is instantiated and what is not and by pointing out that the idea of “instantiation” requires us to have some idea of a “real world” in which things are to be instantiated, and it does not tell us what that is.

An important objection, because of its apparent intuitive attractiveness and the extent to which people will use it, is the objection that only unique programs are relevant. This view is incoherent, and also implies a kind of dualism. The objection is also limited in scope: it does not deal with the general argument, but merely with specific thought experiments in it, and even with this context, it fails to address the issue that many *different* versions of candidates for an observer’s experience could be found by different interpretations.

Another objection that will be made by many people because it seems intuitively “obvious” is that an object does not exist if more information is needed to find it than is present in the object itself. The flaw here is in assuming that there is a special case when the amount of information needed to find the object is equal to the amount of information in it: it would imply that one extra bit of information in the finding algorithm for an object could cause it not to exist.

A further objection is that thought experiments involving multiple versions of you do not need to have any resolution that allows a coherent view of probability: instead, probability values for future outcomes can be left undefined. This is an incoherent view that would actually leave probabilities undefined even now.

One objection focuses on the idea of the mind found in the random behaviour of the wall in John Searle’s argument against strong AI [12] and suggests that minds found by interpretation in such a way would be transient or if they lasted longer then the longer a mind lasted, the more complex the interpretation would need to be. These arguments can be shown to be invalid. It should also be noted that they focus on an extreme consequence of the argument rather than the argument itself, and therefore have little chance of meaningful success anyway.

Another objection involves asking who or what is supposed to be making these interpretations, but this merely involves a misunderstanding of what is being claimed: it is not being claimed that the interpretations actually *happen*, but rather that the *logical capability* of them to be applied implies that things exist.

A further objection is based on the suggestion that we do not need to describe everything formally, and that the issues raised by the thought experiments can be ignored. However, if we cannot describe things formally, even in principle, this implies that things exist about which coherent yes/no questions cannot be asked. This should seem incoherent as an idea.

A final objection is that we do not need to be able to say whether a mind exists in one place or another, but that instead we should accept that shades of grey exist. However, the need to be able to put this into probabilistic terms for answering questions about the likelihood that you are in various situations will effectively turn this into a many-interpretations type view.

PREVIOUS ARTICLES IN THIS SERIES

[1] Almond, P., 2007. *Minds, Substrate, Measure and Value, Part 1: Substrate Dependence*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate1.pdf> or <http://www.paulalmond.com/Substrate1.doc> [Accessed 29 June 2011].

[2] Almond, P., 2007. *Minds, Substrate, Measure and Value, Part 2: Extra Information About Substrate Dependence*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate2.pdf> or <http://www.paulalmond.com/Substrate2.doc> [Accessed 29 June 2011].

[3] Almond, P., 2008. *Minds, Substrate, Measure and Value, Part 3: The Problem of Arbitrariness of Interpretation*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate3.pdf> or <http://www.paulalmond.com/Substrate3.doc> [Accessed 29 June 2011].

[4] Almond, P., 2010. *Minds, Substrate, Measure and Value, Part 4: The Cosmological Many-Interpretations View*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate4.pdf> or <http://www.paul-almond.com/Substrate4.doc> [Accessed 29 June 2011].

[5] Almond, P., 2010. *Minds, Substrate, Measure and Value, Part 5: Cryptic Ontology*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/Substrate5.pdf> or <http://www.paul-almond.com/Substrate5.doc> [Accessed 29 June 2011].

REFERENCES

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

[7] Everett, H., 1957. Relative State Formulation of Quantum Mechanics. *Reviews of Modern Physics*, 29, pp.454-462.

[8] Price, M. C., 1995. The Many-Worlds FAQ. [Online] The Anthropic Principle. Available at: <http://www.anthropic-principle.com/preprints/manyworlds.html> [Accessed 12 December 2010]. (Also available at: <http://www.hedweb.com/everett/everett.htm> [Accessed 12 December 2010] and <http://kuoi.com/~kamikaze/doc/many-worldsfaq.html> [Accessed 29 June 2011].)

[9] Aguirre, A., Tegmark, M., Layzer, D., 2010. Born in an Infinite Universe: a Cosmological Interpretation of Quantum Mechanics. [Online] arXiv:1008.1066v1 [quant-ph]. Available at: <http://arxiv.org/abs/1008.1066> [Accessed 18 June 2011].

[10] Bousso, R., Susskind, L., 2011. *The Multiverse Interpretation of Quantum Mechanics*. [Online] arXiv:1105.3796v1 [hep-th]. Available at: http://arxiv.org/PS_cache/arxiv/pdf/1105/1105.3796v1.pdf [Accessed 18 June 2011].

[11] Duff, A., 1986. Pascal's Wager And Infinite Utilities. *Analysis*, 46, pp.107-109.

[12] Searle, J. R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: The MIT Press. 9th Edition. pp.207-212. (Originally Published: 1992. Cambridge, Massachusetts: The MIT Press.

[13] Grand, S., Almond, P., 2009. *Cryptic Ontology*. [Online] Machines Like Us. Available at: <http://www.machineslikeus.com/forum/cryptic-ontology> [Accessed 31 October 2010]. (A debate between the author and Steve Grand, in which the idea of a multiverse based on all possible interpretations, and the justification for it, was discussed. Steve is sceptical of all this.)

[14] Tegmark, M., 1998. Is the theory of everything merely the ultimate ensemble theory? *Annals of Physics*, 270, pp.1-51. (Also available online at: http://arxiv.org/PS_cache/gr-qc/pdf/9704/9704009v2.pdf [Accessed 4 September 2010]).

[15] Tegmark, M., 2003. Parallel Universes. *Scientific American*, May 2003, pp.40-51. (Also available online at: http://space.mit.edu/home/tegmark/PDF/multiverse_sci_am.pdf [Accessed 5 December 2010].)

[16] Tegmark, M., 2007. The Mathematical Universe. *Found.Phys*, 38, pp.101-150. (Also available online at: <http://arxiv.org/abs/0704.0646> [Accessed 5 December 2010]).