
the Wrongness of Searle

By Paul Almond

3 September 2010

Website:

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

E-mail:

info@paul-almond.com

Searle's argument against strong AI is based on the idea that consciousness results from a specific kind of physical process, whereas the same computation can occur in completely different substrates, with completely different underlying processes, making computation interpretation-dependent and not an intrinsic property of a system. This is flawed. There are no obvious limits to the generality with which a process can be described, and a process description might be sufficiently general that it spans substrates and instances of the process that do not "look the same". To demonstrate this, a stand-in process for consciousness is presented, and gradually changed. Expecting all instances of a process to "look the same" is merely demanding that the definition of a process conforms to the limitations of human pattern recognition. An algorithm can be considered to correspond to a physical process defined very generally. The different instances of the process would not "look the same", but this would be irrelevant: They would all be connected by their compliance with the formal description of the physical process. Importantly, in relation to Searle's argument, *the physical process corresponding to an algorithm could easily be defined with such generality that it can occur in human brains and computers as we understand them practically*. In fact, such a process description would tend to be the natural result of describing the process generally. With such a view of algorithms, Searle's objections based on syntax become an irrelevancy: What is involved is merely just another physical process, if a particularly generally defined one. No answers are offered to the questions that Searle raises, but it is merely shown that Searle is not providing any plausible answers.

Table of Contents

1 Introduction	4
2 A Stand-In for Consciousness.....	6
2.1 The Process	6
2.2 Changing the Process	6
2.3 When did the process become something else?	10
3 Consideration.....	12
3.1 How can a physical process be described?.....	12
3.2 Applying a Physical Process Detector to Computers	12
3.3 Wetness and B-Wetness Revisited	13
3.4 Implications for Searle’s Argument	13
4 The Chinese Zoo	17
5 Conclusion.....	20
6 Bibliography	22

List of Abbreviations

- AI artificial intelligence
GPS global positioning system

1 Introduction

Professor John Searle is skeptical of the idea that artificial intelligence (AI) research could eventually produce computers that are conscious. Searle is not opposed to the idea that a computer could duplicate the *behavior* of a human brain: His issue is with what he calls “Strong AI”: the idea that such behavior should be assumed to be associated with a mind, and which he describes as follows.

*“...according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.”*¹

Searle thinks that this is unlikely.

Searle regards the idea that the right program could cause a mind to exist, irrespective of the computational substrate, as a form of dualism. Searle is sometimes accused of thinking that consciousness needs some mystical property, but this is not his position: His position is one of *specificity of process*. Searle regards consciousness as an emergent property of specific processes in the brain, a position he states as follows.

*“Consciousness is caused by lower-level neuronal processes in the brain and is itself a feature of the brain. Because it is a feature that emerges from certain neuronal activities, we can think of it as an ‘emergent property’ of the brain. An emergent property of a system is one that is causally explained by the behavior of the elements of the system; but it is not a property of any individual elements and it cannot be explained simply as a summation of the properties of those elements. The liquidity of water is a good example: the behavior of the H₂O molecules explains liquidity but the individual molecules are not liquid.”*²

Searle says that things that have physical existence are caused by specific physical processes, and that a representation of these things in a computer is not the same. In fact, he says that the *multiple realizability* of algorithms indicates that they are not physical processes at all, but are subject to interpretation, a position he states as follows.

¹ Searle, J.R., 1980. Minds, brains and computers. *The Behavioral and Brain Sciences*, 3, pp.417-457.

² Searle, J.R., 1998. *The Mystery of Consciousness*. London: Granta Publications. pp.17-18. (Originally published: 1997. New York: The New York Review of Books).

“The multiple realizability of computationally equivalent processes in different physical media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all. They depend on an interpretation from outside.”³

Searle says that a rainstorm has specific causality underpinning it, and even if we simulated this in a computer, nobody would get wet. He argues that consciousness should be treated the same way. Consciousness results from a specific, biological process and a computer simulation of this process will not be the process itself, but will instead involve the different physical processes of whatever computer is running the simulation: We should no more expect consciousness to be associated with a computer simulation of this process than we should expect a simulation of a rainstorm to involve real rain.

In this article I will be arguing against Searle’s position. I previously wrote an article making basically the same argument that I am going to make now, but I did not present the argument clearly enough: This article will hopefully correct that.⁴

One thing that will *not* be featuring in the argument is *the Chinese room*.⁵ The Chinese room is a thought experiment by Searle intended to show that the strong AI position is implausible. I regard the Chinese room argument as a device used to argue Searle’s case, rather than the case itself, which is that consciousness is associated with a specific kind of physical process and a computer algorithm is not. The Chinese room is a distraction here, and rather than discussing it I will attack Searle’s central argument directly.⁶

The argument I will make against Searle will look at the issue of what a process is. I will argue that what constitutes a process could actually be very general and that a process may be so general that two instances of that process do not appear the same to humans, and yet are the same because they both satisfy some formally expressed definition of that process.

³ Searle, J.R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press. p.209. (9th Printing, Originally published: 1994. Cambridge, Massachusetts: MIT Press).

⁴ Almond, P., 2008. *Searle’s Argument Against AI and Emergent Properties*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/SearleEmergentProperties.pdf> or <http://www.paul-almond.com/SearleEmergentProperties.doc>.

⁵ Searle, J.R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press. p.45. (9th Printing, Originally published: 1994. Cambridge, Massachusetts: MIT Press).

⁶ This does not mean that I will *never* write about the Chinese room: In fact, anyone who spends enough time arguing with Searle will have to discuss it sooner or later. It is just that I consider it of secondary importance relative to the importance of dealing with the main argument about processes and specificity, and once a clear position on the main argument has been stated, the Chinese room will be easier to deal with.

2 A Stand-In for Consciousness

2.1 The Process

We are discussing Searle's idea that there is a "specific" kind of physical process needed for consciousness, but if there is we do not know exactly what that process is and what we do know about it will be complicated. We need a simple stand-in for the process that is supposed to cause consciousness: something that we can use as a philosophical device. The process I will use is based on an apparatus often used in experiments or demonstrations involving statistics. The process used is as follows.

A piece of glass is fixed against a brick wall, with a gap between the wall and the glass. A set of wooden pegs (made of mahogany wood) are fixed to the wall. There is a mechanism at the top of the wall to release a ball. When the ball is released, it falls between the glass and the wall. On its way down, the ball may strike one or more pegs, with a change of direction being caused each time this happens. When the ball arrives at the bottom, it is collected in one of a number of bins, arranged along the bottom of the wall. An automatic mechanism then takes it back up to the top, to be released again, and so on.

I know that this does not look anything like consciousness, but that does not matter. I am asking you to imagine that Searle is right and to *pretend* that this is the specific process needed for consciousness. We will now look at how specific this is supposed to be.

2.2 Changing the Process

If the process just described is the basis of consciousness, there has to be some way in which minor differences are acceptable: After all, different people have differently built brains, which means they must have slightly different versions of the above process. We will now consider a sequence of changes to the process, as follows.

1. Are the positions of the pegs critical? We *slightly* alter the positions of one or more wooden pegs.
2. We alter the positions of the pegs more.
3. Must the pegs be made out of the specific kind of wood? We replace the mahogany wood pegs with ones made out of oak.
4. Does it have to be wood at all? We replace the oak pegs with metal pegs.
5. Must the wall be made the same way? We replace the bricks with concrete.
6. We replace the glass with acrylic (Perspex).

7. Must the movement of the ball be due to a gravitational field? We put the apparatus in a large, rotating space station, so that the movement of the ball “downwards” is really due to the space station’s simulated gravity.
8. Must there be physical contact between the ball and the pegs? We put some kind of magnetic system on the ball and the pegs, so that when the ball gets close to a peg, it is repelled magnetically.
9. If we are using a magnetic system to make the ball “bounce off” pegs, do we need the pegs there at all? We take each peg away, just leaving the magnetic system to repel the ball if it comes close.
10. Earlier, the gravity that made the ball fall downwards was replaced by the simulated gravity of a rotating space station, but does the external environment need to provide this at all? We put the apparatus back on the earth, but laid horizontally, and instead of gravity to make the ball move, we equip the ball with small thrusters to accelerate it as if it were under gravitational acceleration. (Clearly, the thrust needs to be in the same direction, even if the ball rotates, but we could use some clever system to vector the thrust appropriately.)
11. Now that the apparatus is horizontal, are both the acrylic (Perspex) “front” and the concrete “back” of the apparatus needed? The acrylic (Perspex) front is removed, leaving just the ball moving around being magnetically repelled on the concrete surface.
12. Does a “peg” (or rather, now, the magnetic system that has replaced a peg) need to exert a force *directly* on the ball? We take the magnetic repulsion systems on the ball and the peg away. Instead, where the pegs used to be, there are now colored marks on the concrete surface over which the ball moves. When the ball passes over these, sensors in the ball detect the mark, indicating a “collision” with a peg, and a mechanism in the ball directs the ball’s thrusters appropriately to deflect it as if it had struck a wooden peg.
13. Are the colored marks needed? Suppose we put a global positioning system (GPS) receiver in the ball. As the ball moves through the system, it continually receives radio signals from GPS satellites, which are used by a mechanism in the ball to determine its position. The ball has a list of the positions at which the “pegs” are. When the ball’s position, obtained from the GPS, indicates that it should be striking a peg, the ball’s thrusters are directed to change the ball’s direction appropriately.
14. Is interaction with an external GPS system needed? We replace the GPS system with an inertial guidance system in the ball.
15. Does the system all have to be in the same place? We cut the apparatus in half, to make two tiles, and physically separate them. The ball starts on one tile. When it reaches the edge of that tile, it strikes a device at the edge which records its position, speed and direction. A mechanism then lifts the ball off that tile, transports it to the other tile, and puts it on that tile, with the position,

- speed and direction that it would have had if the two halves of the apparatus were still joined together. The two tiles are arranged, relative to each other, in the same way that they would be arranged if joined together: There is just a gap between them.
16. Does it matter if the above alteration happens to a greater extent? We do not stop at dividing the apparatus in half. Instead, we divide it into one hundred tiles, each with one hundredth of the area of the original apparatus. Each has a mechanism to pick the ball up when it reaches the edge, and move it to one of the other tiles. Again, the tiles are arranged in the same way, relative to each other, in which they would be arranged if joined together: They are just spatially separated.
 17. Must the tiles be arranged in any particular way? The positions of the tiles are changed *slightly* so that their arrangement is slightly different from the way in which they would be arranged if they were still joined together. This does not stop the system working: The machinery that moves the ball from one tile to another just needs to “know” where all the pieces are.
 18. Do we need as many tiles? All the tiles are actually the same: The actual pegs, and any associated mechanisms, have long since been removed, being replaced by the inertial guidance system in the ball which makes it act as if it had struck pegs. The number of tiles is now halved, so that there are now fifty tiles, each with one hundredth of the apparatus’s original area. Each of the remaining tiles now performs the function of two tiles that were originally next to each other near that location. For example, suppose there were two tiles, Tile A and Tile B, next to each other. They are replaced by a single tile, Tile C. When the mechanism that moves the ball from one tile to another is moving the ball to what is supposed to be Tile A, it now moves it to Tile C and, when the ball arrives on Tile C, the mechanism in the ball causes the ball to behave as if it were on Tile A. Similarly, when the mechanism that moves the ball from one tile to another is moving the ball to what is supposed to be Tile B, it now moves it to Tile C and, when the ball arrives on Tile C, the mechanism in the ball causes the ball to behave as if it were on Tile B.
 19. Is there any limit on the extent to which we can rearrange the tiles? The positions of the tiles are changed completely so that their arrangement is now *completely* different from what it would be if they were still joined together. Again, the machinery that moves the ball between tiles just needs to “know” where all the pieces are.
 20. Is there any limit on the extent to which we can combine tiles? Just as the number of tiles was previously halved, with each tile taking over the role of two tiles, all the tiles except one are now removed. The apparatus has now been reduced to a single tile, with one hundredth of its original area. The ball moves across this tile until it reaches the end, where a mechanism measures its position, speed and direction, lifts it up and puts it on the other side of the tile,

- with the ball's guidance system treating this as if it had moved onto an adjacent tile. The ball therefore rolls repeatedly across the same, small tile, with this situation, at any time, representing the ball at some position in the original, full-sized apparatus.
21. The process has previously involved a spherical ball, which rolls along the apparatus, but why does it actually have to be a ball? The ball is replaced by a cube, and the surface over which it has to slide is made a low-friction one.
 22. The tile over which the "ball", now a cube, moves is reduced in size still further, so that is now only slightly larger than the cube itself. As before, this stands in for a piece of the full-size apparatus. When the cube reaches the edge of the apparatus, the mechanism is used to move it over to the opposite edge, where it continues to move as if it had actually moved onto the next tile. The cube is hardly moving at all now, because it only has to move a small amount before it reaches the edge and gets repositioned. This is happening so quickly that it is not worth detaching the repositioning mechanism from the cube. Someone watching all this would simply see a small tile, much smaller than the original apparatus, almost completely occupied by a cube, which has some mechanism connected to it (the repositioning mechanism) and is continually shaking and firing small thrusters.
 23. The cube can move a tiny amount before it reaches the edge, but would it really matter if it reached the edge right away – as soon as it moved? The size of the tile is now reduced further so that it is no larger than the cube. The cube is now in continuous contact with the edges of the tile. The repositioning mechanism is now redundant, as the cube cannot move anywhere. The edges of the cube measure the force exerted by the cube as it fires its thrusters, and a mechanism "tells" the cube where it is at any time.
 24. Are the thrusters necessary? It is not as if the cube is going anywhere anymore. The thrusters are removed and replaced by flashing lights on the cube. The mechanism responsible for measuring the cube's acceleration is "informed" of the cube's "thruster" use by these lights, and it sends information back to the cube "telling" it where it is.

and so on. I will stop there, because it would probably be pointless to continue. By this stage, we have a process which looks nothing like the original system, but it was made from the original process using a sequence of small changes. I could have made the changes smaller still, were it not for the need to prevent this article from being hundreds or thousands of pages long: I think it should be obvious, from the above example, that I could have altered the system as gradually as I wanted. The process with which we end up is not just a random process, though: It still has some relationship with the earlier stages in the sequence, and with the process with which we started. An observer, able to look in detail at what was going on in each process would be able to see that they all had *something* in common.

Searle himself has discussed systems which have some similarity with the brain, but are based on different substrates: for example, plumbing systems that mimic neural networks, in which the signaling between “neurons” involves water flowing through pipes. Searle intends this to make the point that you can mimic a process without having the right underlying causality – without it being the same process. The point I am making, here, is that consideration of different systems like this actually works against Searle, because it raises the idea of changing some implementation of a process gradually while bringing into doubt the issue of where the edges are.

In the example given here, the original implementation process has ended up as one that is a kind of abstraction of it. In fact, I could have described a sequence of changes in which the process retained some of its low-level messiness – yet with all the instances of the process still having some central idea in common. I need not have stopped at the rather abstracted version of the process that resulted, either: After “going up” to the that level of abstraction, I could have gone back down to some more detailed version of the process, but one which did not look like the original version, and yet would have some features in common with it.

2.3 When did the process become something else?

We are using the physical process in the wooden pegs apparatus as a stand-in for the specific physical process that is claimed by Searle to cause consciousness. We just made a sequence of changes to that process, so the obvious question here is as follows.

When, during the sequence of changes, did the process stop being the same process?

Was it when the pegs were replaced with metal ones? Was it when the apparatus was broken up into separate tiles? Was it when the positions of the tiles were mixed up?

Putting consciousness aside for a moment, and just considering this with regard to when a process ceases to be the same process, it should be clear that there is no obvious answer: *It depends on how narrow the definition of the process is.* One process might be defined so narrowly that it demands that the pegs are made of wood, and with that definition it ceased to be the same process as soon as the pegs were replaced with metal ones. Another process’s definition might say that it remains the same process until the positions of the tiles are all mixed up. Another process’s definition might say that it remains the same process right through the sequence of changes.

If a process’s definition is very narrow, then we might reasonably expect different versions of that process to look the same: The definition would be forcing them to be superficially similar to each other. If, however, the definition of a process is more general, different instances of the process may not look similar. This may suggest that they are different processes, but this would be a prejudice due to limitations in human pattern recognition: In reality, the apparently different processes would have some feature in common, even if it is not one that is readily apparent to an observer.

One objection to this could be that all versions of a process must somehow “map onto” each other: that you can somehow overlay one on another in space and time and they will match up. This objection would be naïve: We routinely accept many processes as having versions for which this simply does not happen.

Another objection could be that for two versions of a process to be the same, the “underlying processes or causality” must be the same. In this discussion, however, I have not even separated higher-level processes from underlying ones: We have been considering the entirety of a physical system. In any case, even if we say that there is some underlying process or causality which has to be the same, that just runs into the problems I have discussed here of saying when a process stops being the same process: It just becomes an issue for the underlying process, now.

3 Consideration

3.1 How can a physical process be described?

I have argued that we cannot consider the description of a process as being in terms of how it “looks”: Some processes might be defined far more generally than such a description would allow. This raises the issue of exactly how we *are* supposed to express the description of a process.

Any process should be formally describable in principle. In a previous article I described a *process detector*. This is a philosophical device intended to detect the existence of a particular physical process. The process detector has probes which it can move around to measure physical variables in space, and an internal computer with a program, designed especially for the particular process which it is supposed to detect, that controls the movement of the probes, obtains measurements from them as inputs, and indicates when it has found an instance of the process which it is supposed to be detecting. For example, if a physical process detector were designed to detect digestion, it might indicate a positive result when placed near humans who were digesting food or animals that were digesting food: Its program would look at what was going on around it, and provided that this satisfied its criteria, it would indicate that it had found digestion.

The idea of the process detector suggests that, for a physical process, the main requirement should be as follows.

A physical process is that which can be formally described, *in principle*, by the program in a process detector.

Anything else is just human-prejudice and expectation that reality complies with the limitations in our pattern recognition.

3.2 Applying a Physical Process Detector to Computers

There is no reason why a physical process detector could not also be used on computers. Suppose we have a particular chess program and we want to detect any computers that are running it. The physical process detector could be programmed to indicate a positive result when it has found, in a physical system, a process satisfying the criteria for some general process that we define as “running the chess program”.

As I have described things so far, the physical process detector program to find a chess program would have to be defined in much more general terms than the program needed to find something much more specific: We can understand chess programs to be

running in many different ways. This, however, does not make the situation *qualitatively* different to the finding of any other physical process: It is just a matter of degree.

In fact, there is a way in which the definition of a physical process may be able to describe it in many situations with radically different causal underpinning without having to be explicitly very general itself: It might use *abstraction*. We tend to define processes in terms of underlying processes. A physical process detector may have a program to find a process that works by referencing processes that can be detected by physical process detectors with different finding programs. If some underlying process(es) is/are specified sufficiently generally, a program to find a higher-level process may not need to be very general, as it can exploit the underlying generality.

3.3 Wetness and B-Wetness Revisited

The argument presented here has some relevance to the concepts of “wetness” and “b-wetness” which I discussed in the previous article about Searle’s argument.⁷ I said that we could define a property, “b-wetness”, in a more general way than standard “wetness”, such that the property could be shared by “wetness” and a computer running a simulation of “wetness”. In fact, a property called “c-wetness” could be defined, which is not significantly more general than “wetness”, but is based on lower-level properties that are, on some level, defined sufficiently generally that *they* could be found in a conventional situation where wetness occurs and in a computer running a simulation. This would mean that a process which, at a high level is not substantially different from wetness could be found in conventional situations and in computer simulations. This then raises the issue of whether we should require that the description of “wetness” specifies what it has to be like “all the way down”, or whether it allows “wetness” to be described very generally at lower levels or to be based on processes which are themselves described very generally (both of which amount to the same thing). If we allowed this, then it would suggest that we might take the view that wetness=c-wetness, implying that real wetness could occur in a computer simulation. I will not be taking a position on this in this article.

3.4 Implications for Searle’s Argument

All this amounts to a refutation of Searle’s argument against strong AI.

Searle accuses advocates of strong AI of a kind of dualism. Admittedly, that may *seem* to be the case due to the way the strong AI case is presented: Advocates of strong AI tend to talk about consciousness being about “information processing” and resulting when “the right computation is being performed”. This, however, should hardly surprise us: It

⁷ Almond, P., 2008. *Searle’s Argument Against AI and Emergent Properties*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/SearleEmergentProperties.pdf> or <http://www.paul-almond.com/SearleEmergentProperties.doc>.

is the sort of language used in everyday computing. The charge of “dualism” results from taking it as a formal statement of the strong AI position, when in fact, it can be taken to be a somewhat badly worded statement of a position which does not involve dualism at all.

The strong AI position is coherent, and avoids any idea of dualism, if we consider it in terms of *physical processes*. For any specific computation, the description of that computation actually corresponds to a description of a *physical process* which can be found to occur in physical systems that may appear completely different to normal, human intuition. A process could occur with such generality in one of two ways.

- The description of the process may itself be very general, allowing many physical systems that *appear* different to have that process.
- The description of the process may reference lower-level processes which are themselves, at some level, defined *very generally*, allowing many physical systems that appear different to have those processes.

Which of these is used is really a matter of convenience: They amount to the same thing and it is just an issue of how the description is stated.

Against this, advocates of Searle’s position may reply that *I don’t get it*: that all the different ways in which a computation could be done involve completely different processes. The argument already made in this article should have dealt with this: Requiring physical processes to be narrowly defined in ways so that all instances of that process “look the same” is just naively expecting things to conform to human pattern recognition. Physical processes can be defined with different degrees of generality, and a physical process could be defined with such generality that it can be found in what appear to be different kinds of physical system. An algorithm could be viewed as an informal way of stating such a definition.

Advocates of Searle’s position could respond that different implementations of an algorithm cannot possibly involve the same physical process because the “underlying process” or “causal underpinnings” are different. This would achieve nothing, because it puts us back in the same situation, where the underlying process can be defined with sufficient generality that it can occur in systems that appear different.

Dualism, then, is not an issue: All of this comes down to physical processes, just as it does with most other things that we might want to discuss. Any algorithm can be viewed as an informal description of a physical process. Advocates of Searle’s position could say that the description of the algorithm alone does not tell us how general the description of the physical process should be. Humans tend to treat this informally because “everyone knows”, for practical purposes, when a particular system is supposed to be running a particular program. How is this different, though, from how things would be with any other process – with Searle’s specific, biological process? Our idea of many processes involves informal edges, and to make a particular issue of them

with regard to the kinds of processes that occur in computing would achieve nothing. If necessary, we could formally define a physical process that we would regard as running a particular algorithm, such that it could allow the process to occur in what appear to be different systems, just as a specific, biological process could be accurately defined. The only difference between the two situations is the *degree* of generality.

Searle argues that consciousness must result from a specific physical process. The problem, now, is that *Searle has no way of showing how specific the physical process for consciousness needs to be*. We might think that the process can only occur in human brains, or we may accept it could occur in some other “biological” systems. On the other hand, we might define the process for consciousness with the generality of an algorithm: We might describe it in such a general way that it can occur in systems with completely different appearances. Nevertheless, all instances of consciousness would still be due to the same kind of process: It would just be a very generally defined one.

One reply to this could be that it is implausible for the process responsible for consciousness to be so generally defined, but this would just be an appeal to incredulity. It is, in any respect, not what Searle’s argument is intended to show: Searle is trying to show that the running of algorithms on different substrates implies *qualitatively* different processes. The word “different” in front of “substrates” is actually loading the issue here: Again, human pattern recognition might find two instances of processes to be “different” that can actually be connected by a unified, formally expressed description.

Advocates of Searle’s position could point to the issue of *universal realizability*. Searle has argued that whether an algorithm is running in a given system is an issue of interpretation: In principle, we could interpret any system in such a way as to find an algorithm running in it. Searle states this as follows.

“For any program and for any sufficiently complex object, there is some description of the object under which it is implementing the program. Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic to Wordstar, but if it is a big enough wall it is implementing any program, including any program implemented in the brain.”⁸

The problem with trying to use universal realizability is that, if we are regarding algorithms as corresponding to descriptions of physical processes, it amounts to nothing more than saying that it is not obvious where the edges of such processes should be drawn. The algorithms with which we interact in everyday life – commercially created programs in computers for example – would be relatively easily found by short programs in process detectors. It may be that more extreme interpretations are

⁸ Searle, J.R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press. p.208-209. (9th Printing, Originally published: 1994. Cambridge, Massachusetts: MIT Press).

possible, but these would correspond to longer finding programs, or would need to be based on more general descriptions of the processes underpinning them. It would not be the case that all algorithms, found by all algorithms would be equal. The issue is where to draw the line: how generally we should define the physical process corresponding to a particular algorithm. I do not intend to try to resolve this issue in this article: I have given my response to it in writing elsewhere. Rather, I would say that this is nothing more than the issue of how specific the process should be, again, and this issue is going to arise with the description of any process that is supposed to correspond to consciousness. *There is nothing about general views of processes that allow computers to be conscious that makes this issue qualitatively different to the issue presented by Searle's more specific process.* (Incidentally, Searle has said that he does not regard universal realizability as a serious problem, and that it can be resolved by "tightening up the definition of computation".⁹)

The idea of regarding an algorithm as an informal description of a very generally described physical process is actually a quite simple one, and we might reasonably think it has much to recommend it as a way of viewing consciousness on that basis. If we define the process very generally, it is natural that the description ends up like this.

⁹ Searle, J.R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press. p.209. (9th Printing, Originally published: 1994. Cambridge, Massachusetts: MIT Press).

4 The Chinese Zoo

Some people will insist on criticizing what I have said by hand-waving and claiming that *processes involving different substrates do not have the same causality*: that they are not the same process. I have argued in some depth that this objection does not work, because there is no reason to expect the edges of a process to be simply described and to correspond to human intuition. To make this point, I will introduce what I will call *the Chinese zoo*.¹⁰

The Chinese zoo has some similarity to “Biomorph Land”: the genetic space of all possible creatures described by Professor Richard Dawkins.¹¹ The Chinese zoo contains an example of every machine that can be conceived. By “machine”, I simply mean a physical system that performs some process or processes.¹²

For any machine in the Chinese zoo, its neighbors are only slightly different from it. If you go away from a machine in each dimension, you will encounter a neighbor that is different in some different respect. If you look around a machine in all directions, you will see all the ways in which a slightly different machine can be made.

The Chinese zoo is a hyper-dimensional structure, but to make things easier to think about, let us ignore this and, instead, imagine the Chinese zoo existing in just two dimensions, so that we could walk around and view the different machines – just like the exhibits in a real zoo.

Exploring the Chinese zoo, we find a machine that is conscious. It has some process going on in it that is known to correspond with consciousness. The machine’s neighbors are running the same basic process, so they are conscious too, and we might expect their neighbors to be conscious. Of course, it would be absurd to go on forever like this. We should accept that, whatever the process is that corresponds with consciousness, it must have some edges: If we keep changing a process enough, in some ways, eventually it will not count as the original process. We should therefore expect the conscious machines to occupy a region of the Chinese zoo. We might not necessarily expect this region to have “hard edges”: It could be that, near the edges, machines become less conscious. Suppose that the Chinese zoo’s managers have put red tiles on the floor under each conscious machine (or each machine with a reasonable amount of consciousness), identifying it as conscious, and where two conscious machines are next

¹⁰ Yes, I just copied the “Chinese” part of the name from Searle’s “Chinese room” here for no obvious reason other than it being generally related to the same argument.

¹¹ Dawkins, R., 1991. *The Blind Watchmaker*. London: Penguin Books. pp.66-73. (Originally published: 1986. London: Longman).

¹² We can dispense with any idea of machines as necessarily artificial or simple in nature, and if you are not prepared to do that, just use some other word instead of “machine”.

to each other, the red-tiled regions of the floor connect to each other: You can explore the collection of conscious machines by staying on the red tiles as you walk around.

The question we should be asking is: “What conditions must the pattern of red tiles on the zoo floor satisfy?” This is just another of asking: “What conditions does a process’s description need to satisfy?” It is not up to us to impose rules without justification.

We might see a relatively small, simply-shaped area of red tiles. This would correspond to Searle’s view, in which the members of the collection of conscious machines are all superficially quite similar to each other. (I am not accusing Searle of limiting conscious to human brains or biology, here, by the way: It is just obvious that he will demand some level of similarity that different computational substrates are unlikely to deliver.)

The layout of red tiles may not appear so simple, though. It may have an apparently complex pattern, with paths of red tiles running through the zoo, twisting and turning in apparently complex ways to connect different regions of the zoo. By staying on the paths of red tiles, you could find yourself walking around regions of the zoo with machines that look completely different. You might walk from a region of biological machines, to a region of electronic machines, and then to a region of machines made of wood and string.

Now, suppose an advocate of Searle’s position turns up at the zoo, sees an apparently complex pattern of red tiles meandering about all over the place and connecting all these different looking machines. He may protest that the zoo’s managers have it all wrong, that consciousness is associated with a specific process, and that this process cannot be going on in machines that are so different. The zoo’s managers might explain how the red tiles have been placed according to some *description* of the process and that the machines with red tiles under them comply with that description. The Searle advocate could hand-wave and claim that this is wrong: that it is not the same process, but all he would be asking for, really, is a process that is easier to understand, so that you can easily see, by looking at a machine, whether it should have red tiles underneath it.

This analogy may give an idea of what is really meant when I talk of a “more general” process description. It might seem to mean one that would include more machines, and with regard to consciousness this is a quite likely outcome, but it really means one that includes machines that are less superficially similar – and which is expressed in a more abstract way and which is more removed from human ideas of what “similar” means. In principle, a more “general” description could result in the red-tiles covering very little of the zoo’s floor, but they could be in patterns that wind and twist all over the place, so that most of the zoo is not far from a red-tiled region.

I have charged Searle’s argument with demanding that the instances of a process are superficially similar. Advocates of Searle’s argument may say that I am misrepresenting Searle in saying this and, indeed, I am not aware of Searle *explicitly* saying this. It is

implicit, however, in Searle's view that it is implausible to associate an algorithm with consciousness, and therefore that it is implausible to associate an algorithm with a physical process. If we put red tiles under all the machines that were running some algorithm, assuming that our description of the algorithm were extended to include criteria that must be satisfied by the physical nature of a machine that is running it, then we should have a well-defined physical process – and it is not as unreasonable as Searle thinks to view it as a candidate for consciousness. In fact, if we decided to adopt a more general process description, one based on algorithms would be a natural result.

5 Conclusion

Searle's argument against strong artificial intelligence (strong AI) is based on the idea that consciousness should be expected to result from a specific kind of physical process, whereas the same computation can correspond to completely "different" underlying processes in completely different substrates, making computation dependent on interpretation and not an intrinsic property of a system.

It has been argued that this is flawed. There are no obvious limits to the generality with which a process can be described, and no dividing line between a narrowly defined process and a generally defined one. We might consider the generality of a process description as being its "bandwidth" and a process might be defined in sufficiently general terms – with sufficient bandwidth – that the description spans different substrates and there are instances of the process that do "not look the same". Expecting all instances of a process to "look the same" is merely demanding that the definition of a process conforms to the limitations of human pattern recognition.

An algorithm can be considered an informal, and somewhat vague, way of describing a physical process. This physical process can be considered to be defined in very general terms. The different instances of the process would not "look the same", but this would be irrelevant: They would all be connected by their compliance with the formal description of the physical process. That this formal description of the physical process tends to be missing from an algorithmic description of the process should not cause us to infer anything philosophically profound: Vagueness, to some extent, appears in the descriptions of most processes. Alternatively, a process may be defined in a relatively narrow way, yet still be capable of implementation across a wide range of substrates, by making the process description dependent on lower-level processes that are themselves defined very generally. Importantly, in relation to Searle's argument, *the process corresponding to an algorithm could easily be defined with such generality that it can occur in human brains and computers as we understand them practically*. This would be a relatively simple way of defining a process for consciousness. Further, it would be a *natural* way of defining a particularly general process to be associated with consciousness.

Advocates of Searle's position may object that different instances of an algorithm in different physical media cannot possibly involve the same process, because the underlying processes or causality are different: This objection would not help, as the idea of defining a process very generally can be used on any underlying processes.

With such an understanding of an algorithm, Searle's objections based on syntax become an irrelevancy: What is involved is merely just another physical process, if a particularly generally defined one.

This does not resolve the issue of *how* generally the physical process associated with an algorithm should be defined. However, what of that? That is going to be an issue irrespective of whatever process is causing consciousness, and Searle is not providing any answers here. He is merely demanding a certain level of specificity in the process description, while implying that the idea that computers could be conscious is incoherent, when in fact it would be entirely consistent with a higher level of process generality – or a process based on sufficiently general, underlying processes.

Searle has been accused of “carbon chauvinism”, a charge which he denies, and which is unfair: He does not claim that conscious can only be caused in biological systems. However, I suggest that his position can be legitimately described as “*narrowly-defined-process chauvinism*” or “*low-bandwidth-process chauvinism*”.

6 Bibliography

Almond, P., 2008. *Searle's Argument Against AI and Emergent Properties*. [Online] paul-almond.com. Available at: <http://www.paul-almond.com/SearleEmergentProperties.pdf> or <http://www.paul-almond.com/SearleEmergentProperties.doc> [Accessed 3 September 2010].

Dawkins, R., 1991. *The Blind Watchmaker*. London: Penguin Books. pp.66-73. (Originally published: 1986. London: Longman).

Searle, J.R., 1980. Minds, brains and computers. *The Behavioral and Brain Sciences*, 3, pp.417-457.

Searle, J.R., 1998. *The Mystery of Consciousness*. London: Granta Publications. pp.17-18. (Originally published: 1997. New York: The New York Review of Books).

Searle, J.R., 2002. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press. p.45, pp.208-209. (9th Printing, Originally published: 1994. Cambridge, Massachusetts: MIT Press).