International Journal of Machine Consciousness

EMPIRICALLY GROUNDED CLAIMS ABOUT CONSCIOUSNESS IN COMPUTERS

DAVID GAMEZ

Department of Computing, Imperial College, London, SW7 2AZ, UK
david@davidgamez.eu

Received Day Month Year
Revised Day Month Year

Research is starting to identify correlations between consciousness and some of the spatiotemporal patterns in the physical brain. For theoretical and practical reasons the results of experiments on the correlates of consciousness have ambiguous interpretations. At any point in time a number of hypotheses co-exist about and the correlates of consciousness in the brain, which are all compatible with the current experimental results.

This article argues that consciousness should be attributed to any system that exhibits spatiotemporal physical patterns that match the hypotheses about the correlates of consciousness that are compatible with the current experimental results. Some computers running some programs should be attributed consciousness because they produce spatiotemporal patterns in the physical world that match those that are potentially linked with consciousness in the human brain.

Keywords: Correlates of consciousness; functionalism; information integration; experiments; Turing machine; Analytical Engine.

1. Introduction

> "Could a machine think?" My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines. ... No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not. [Searle, 1980, p.424]

Over the last twenty five years research on the neural correlates of consciousness has identified areas of the brain and properties of neural activity that are systematically linked with conscious states [Tononi & Koch, 2008], and there has been considerable discussion about the possibility that artificial systems could be conscious. This article addresses this debate by suggesting that claims about consciousness in computers can be grounded in experimental work on consciousness in the human brain. The first half examines what we know and can hope to know about the correlates of consciousness in the brain. I will then

argue that some computers running some programs will cause spatiotemporal patterns in the physical world that match those that are potentially linked with consciousness in the human brain.

2. The Correlates of Consciousness

2.1. *The Platinum Standard System*

The starting point for research on the correlates of consciousness is a physical system that is known or commonly agreed to be associated with consciousness. While there has been an extensive amount of debate about the importance of the body and environment to thought and consciousness – for example, [O'Regan & Noe, 2001], [Clark, 2008] - it is also clear that the physical brain plays a special role in relation to consciousness. For example, living brains are necessary for consciousness in humans, brain damage leads to partial or total loss of consciousness, and there is a substantial amount of experimental evidence for correlations between consciousness and neural activity [Tononi & Koch, 2008]. The body and environment also have little influence on consciousness during dreams, out of body experiences and hallucinatory states [Gamez, 2007]. All of this strongly suggests that the brain is the part of the physical body that is linked with consciousness.

Although we typically assume that infants and higher mammals are conscious, we are most confident that consciousness is associated with adult human brains. To focus the specification of a conscious system further, it is necessary to state that the adult brain should be normal, i.e. it is undamaged and its functions and measurements fall within two standard deviations for the human species. The brain also needs to be awake at the time of measurement, with "awake" being used as a non-technical indication that the brain is functioning in a way that is typically considered "conscious". This type of wakefulness is distinct from the medical definition, since apparently wakeful states can be exhibited by people in a vegetative state who are unlikely to be conscious [Laureys et al., 2002]. While there will be times when the awake normal adult human brain is not conscious – for example, epileptic automatism [Ramachandran & Blakeslee, 1998] - a science of consciousness has to start somewhere, and the awake normal adult human brain is the physical system that we are most certain is associated with conscious states.

The awake normal adult human brain will be referred to as the platinum standard system.[a] Just as a platinum-iridium bar in Paris was used to define the length of a meter, the awake normal adult human brain is our platinum standard for a conscious system. If this physical system is not associated with consciousness most of the time, then nothing is. In the future we might decide that other systems can be used as platinum standards for research on consciousness, such as a race of aliens based on silicon chemistry. However,

[a] Other parts of the physical body and environment could be incorporated into the platinum standard system if they were shown to be necessary for consciousness.

at the present time there is only one type of platinum standard system, which significantly constrains our ability to precisely identify the correlates of consciousness (see Section 3).

In line with Baars' [1988] notion of contrastive analysis, the normal adult human brain in a state of deep sleep or otherwise unconscious can be used as a platinum standard *un*conscious system, since we have first hand 'experience' of not being conscious in this state. This makes the unprovable assumption that the deeply sleeping or anesthetized brain is unconscious, and not just unable to remember and report its conscious states.

2.2. *Experiments on the Correlates of Consciousness*

To establish which aspects or parts of the platinum standard system are linked to consciousness, we need to measure different properties of the physical brain, measure consciousness and look for correlations between the two. This type of research does not attempt to *reduce* consciousness to the physical brain, nor is it concerned with metaphysical speculations about dualism, epiphenomenalism, supervenience, and so on. It is a purely empirical approach that uses systematic experiments to identify correlations between consciousness and the physical world.

The contents and level of consciousness in the platinum standard system are measured through first person behavioral reports – for example, "I am conscious of a red balloon". Although there are a number of problems with accurate descriptions of conscious states [Gamez, 2006], there is not space to cover them here, and it will be assumed that phenomenological reports from the platinum standard system can be gathered in enough detail for systematic experiments on the correlates of consciousness. The physical states of the platinum standard system can be measured in a variety of ways – for example, using EEG, fMRI or implanted electrodes. By identifying correlations between the phenomenal and physical measurements we can identify features of the platinum standard system that are correlated with conscious states.

A collection of one or more features of the physical world that are correlated with consciousness and which are absent as a collection when consciousness is absent will be referred to as a set of sufficient correlates of consciousness (a SCC set).[b] There might be more than one SCC sets and SCC sets could overlap. The brain contains one or more SCC sets when consciousness is present. Each member of a SCC set will be referred to as a correlate of consciousness and a feature of the physical world that *might* be part of a SCC set will be referred to as a potential correlate of consciousness (PCC). In the example given in Table 1, D is not a correlate of consciousness because it does not systematically co-vary with conscious states, A, B and C are correlates of consciousness, and {A,B} and {A,C} are SCC sets.

Experiments on the correlates of consciousness are attempting to identify spatiotemporal patterns in the platinum standard system that form SCC sets. The current experimental results and our informal observations of strong correlations between

[b] If it turns out that there is only one SCC set, then this SCC set will be necessary as well as sufficient for consciousness. Otherwise one particular SCC set will not be necessary for consciousness because the brain could be associated with consciousness when a different SCC set is present.

physical and phenomenal states suggest that consistent SCC sets are likely to exist. However, it is possible that further research could demonstrate that stable correlations do not exist between consciousness and features of the platinum standard system.

Table 1. Illustrative example of correlations that could exist between consciousness and features of the platinum standard system. A, B, C and D are physical properties of the brain, such as the presence of dopamine, neural synchronization or 40Hz electromagnetic waves. "1" indicates that a feature is present; "0" indicates that it is absent.

A	B	C	D	Consciousness
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	1
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1

2.3. *Functional Correlates of Consciousness*

It has often been claimed that functional features of the brain are linked to conscious states. Theories about consciousness that have a substantial functional dimension include Aleksander's [2005] axioms, Metzinger's [2003] constraints and global workspace theory [Baars, 1988]. A key question in discussions about consciousness and machine consciousness is whether the execution of a function could be correlated with consciousness by itself (could a SCC set consist solely of functions?), or do SCC sets necessarily include non-functional properties of the physical system.

In this discussion I will assume that functional correlates of consciousness can be implemented by programs running on a computer. Programs already exist for many of the proposed functional correlates of consciousness, such as global workspace theory [Franklin, 2003], and even if we don't know how to write a program for a particular function (for example, a complete simulation of the brain), it does not seem unreasonable to assume that such a program could in principle be written.[c]

[c] The limits of computer power will not be considered here.

The operation of a computer program can be described using a state machine in which the nodes represent states of the computer running the program and the links between nodes represent transitions between states that occur in response to particular conditions. An example of a program and its corresponding state machine is given in Figure 1. A state machine description of a program makes it easier to ask whether a physical system, such as the brain, implements a particular program and whether the execution of a program could be correlated with conscious states.

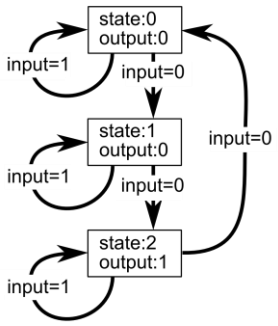
a)

```
//Initialize variables
input = 0;
state = 0;
output = 0;

//Main run loop
while (true){
    checkInput();
    if(input == 0){
        if(state == 0)
            state = 1;
        else if (state == 1)
            state = 2;
        else if (state == 2)
            state = 0;
    }

    if(state == 2)
        output = 1;
    else
        output = 0;
}
```

b)



c)

Time	input	state	output
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	2	1
6	0	0	0
7	1	0	0
8	1	0	0

Figure 1. a) Example program describing the behavior of the wheel discussed in Turing [1950] and Bishop [2002; 2009]. The state of the input variable is updated by the checkInput() function, which could identify whether a brake is on or off. The output variable could control a light. b) State machine describing the operation of the program. c) Single run of the program in which the input varies as shown and the state and output variables are changed by the program in response to the input.

To establish whether a particular function could form a SCC set by itself it is necessary to carry out experiments to test whether it is executed in the brain when consciousness is present, and not executed when consciousness is absent. The first step in these experiments is to define the measureable features of the platinum standard system that will be treated as states in the state machine. For example, the levels of blood oxygenation in different parts of the brain could be treated as states, or the voltages of the neurons. Different definitions of the states of the brain are likely to lead to different interpretations of the functions that are being executed, and experiments will have to be carried out with different state definitions to establish which has the strongest link with consciousness.

When the states of the brain have been defined two approaches can be used to determine if a particular function is being executed by the brain when it is associated with consciousness. The first is to examine the brain's state transitions to see if they contain a particular function that has been proposed to be correlated with consciousness. For example, the state transitions of the brain might contain a state machine that implements a global workspace. If this was executed when consciousness was present and never executed when consciousness was absent, we would have evidence that a global

workspace could form a SCC set by itself. One problem with this approach is that there are likely to be many different ways of writing the complex functions that have been claimed to be linked to consciousness in the brain. A second issue is that over a finite time period the sequence of states that the brain moves through might be consistent with the execution of several different state machines, including the function that we are looking for (see Figure 2). This would undermine the claim that a particular function is being executed while the brain is conscious.

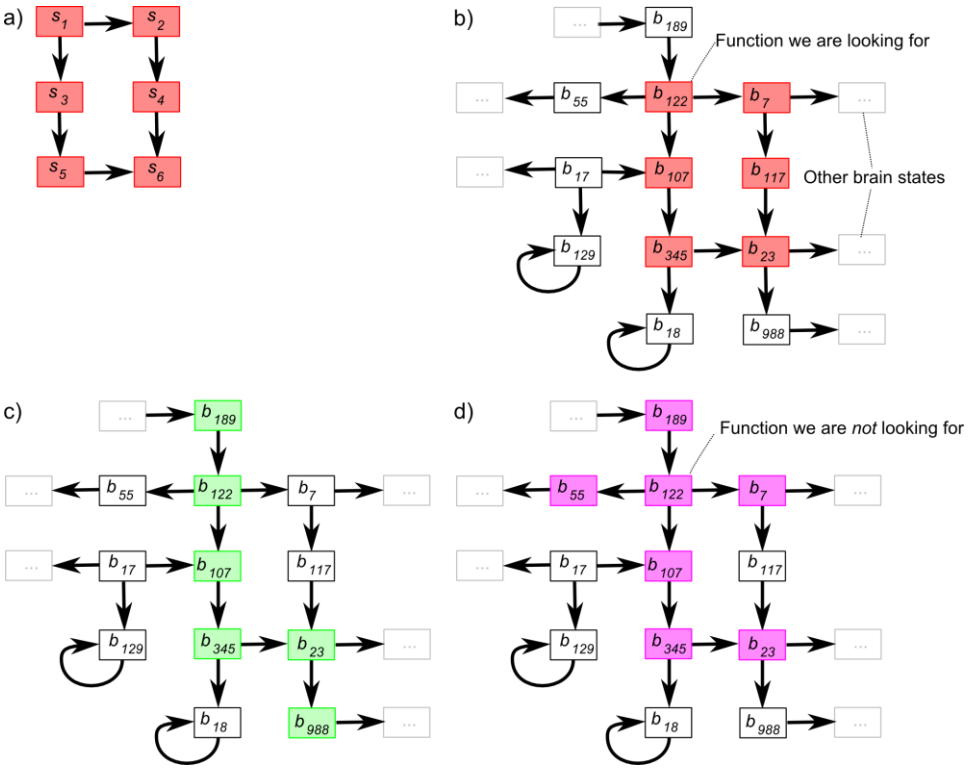


Figure 2. Illustration of the attempt to identify a particular function in the state machine of the brain. a) State machine for a function that has been proposed to be correlated with consciousness. b) Complete state machine of the brain under all possible conditions, which contains the state machine we are looking for. c) Sequence of states entered by the brain when consciousness is observed. Although this sequence of states is consistent with the execution of the function we are looking for, it is also consistent with the execution of many other functions. d) State machine describing a function that we are *not* looking for, which is also compatible with the sequence of states in c.

A more promising way of identifying the functional correlates of consciousness is to monitor the state machines that are executed when the brain is conscious without attempting to match them to particular functions. To understand this in more detail, consider an illustrative experiment in which we divide the brain into three areas and use fMRI to measure the average activity in each area. An area is considered to be in state 0 when its activity is below a threshold and in state 1 when its activity is above a threshold.

The states of the three areas are combined into a single three bit number that expresses the overall state of the brain – for example, 001. Let us suppose that we measure the brain activity under different conditions and obtain the state machine shown in Figure 3a. We then identify the states and transitions that occur when the subject is conscious - a possible result is shown in Figure 3b. Finally we record a particular execution trace from the brain when the subject reports consciousness, as shown in Figure 3c.

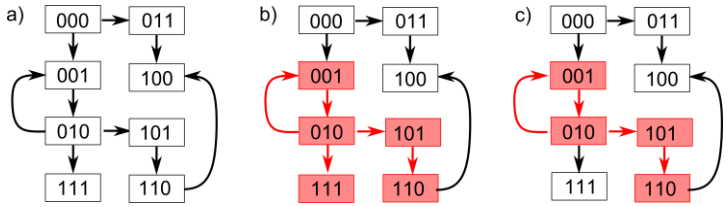


Figure 3. a) Illustrative state machine for the brain under a variety of different conditions. The arrows show the transitions that occur. b) States and transitions that occur when the brain is associated with consciousness. c) Over a particular time period when the subject reports consciousness the brain enters states 001, 010, 001, 010, 101, 110.

At first glance the appearance of a particular set of state transitions whenever the brain was conscious, which were never present when the brain was unconscious, would appear to be good evidence that a function is correlated with consciousness. However, while this would show that a particular function might be *part* of a SCC set, there are a number of reasons for doubting whether the execution of a function could form a SCC set by itself. A first issue is that if it is purely through the execution of a certain function that the brain is conscious, then any system that executes this function must be conscious as well. This leads to consciousness being attributed to implausible systems, such as the population of China connected with radios and satellites [Block, 2006]. More problematically, Bishop [2002; 2009] argues that over a finite time interval any sufficiently complex physical system follows the same state transitions as a function that is correlated with consciousness. Consider the example shown in Figure 3c in which the brain moves through the states 001, 010, 001, 010, 101, 110 over the time interval t_1-t_6. If the execution of a particular function is a necessary and sufficient correlate of consciousness, then any system that follows the same execution trace should be conscious as well. So consider a system, *s*, that goes through a sequence of states s_1, s_2, s_3, s_4, s_5, s_6 (this could be a clock). As Bishop shows, by mapping brain state 001 onto the disjunction of states [s_1 v s_3], 010 onto [s_2 v s_4], 101 onto s_5 and 110 onto s_6, we can interpret *s* as executing the same function as the brain over the time interval t_1-t_6, and so *s* should *also* be considered to be correlated with consciousness over this period. One consequence of this argument is that the functional correlates of consciousness can be found in the unconscious brain over a particular time interval, which eliminates their status as PCCs. Another consequence is that the functional correlates of consciousness can be found in all other systems that have a minimal level of complexity, which leads to an untenable panpsychism. A number of people have responded to this remapping argument by claiming that the target system *s* is missing the counterfactuals of the

original function, which might also be important to consciousness [Chrisley, 1995; Chalmers, 1996]. A full discussion and response can be found in Bishop [2009].[d]

While these arguments suggest that functions cannot form SCC sets by themselves, it is an open question whether the implementation of a particular state machine in the brain could be *part* of a SCC set. It will be shown later how particular functions running on particular computers could lead to spatiotemporal patterns in the physical world that might match the correlates of consciousness in the platinum standard system.

2.4. *Neural Correlates of Consciousness*

Neural activity cannot form a SCC set by itself because neurons are active when the brain is unconscious. However, some features of neural activity, such as synchronization and/or neural activity in particular brain areas, might be consistently correlated with consciousness [Tononi & Koch, 2008].

2.5. *Material Correlates of Consciousness*

Some of the brain's chemicals could form part of a SCC set, although most of them cannot form SCC sets by themselves because they are present when the brain is unconscious. However, it is possible that there could be chemical transformations in the brain that only occur when consciousness is present.

2.6. *Mathematical and Algorithmic Correlates of Consciousness*

Many brain measurements exhibit mathematical relationships that could be correlated with consciousness. For example, a number of algorithms have been put forward to measure information integration [Balduzzi & Tononi, 2008], [Gamez & Aleksander, 2011] and preliminary experiments have demonstrated correlations between information integration and consciousness [Lee et al., 2009; Massimini et al., 2009; Ferrarelli et al., 2010].[e] There is some potential overlap between mathematical and algorithmic correlates of consciousness and neural correlates of consciousness because neural properties, such as synchronization, can be expressed in mathematical terms.

One potential issue with mathematical and algorithmic correlates of consciousness is that information is often treated as if it was an objective property of a system, whereas it is typically a highly subjective interpretation. This problem can be addressed by interpreting information as data, which can be more easily linked to objective physical properties [Gamez, 2011]. A second issue is that it is necessary to specify the *level* at which an analysis is carried out. The mathematical and algorithmic theories that have been put forward are expressed as relationships between elements of a system, which

[d] These arguments against functions forming SCC sets by themselves apply all the way down to the functions of the atoms in the brain and below. The execution trace of a complete simulation of the brain over a particular time period could be remapped onto any other sufficiently complex sequence of states, which would be interpreted as implemented the brain's functions over that time period.

[e] While Tononi [2008] hypothesizes that information integration *is* consciousness, this paper will only consider mathematical and algorithmic relationships to be potential *correlates* of consciousness.

could be brain areas, cortical columns, individual neurons, glia, ion channels, molecules, electromagnetic waves, and so on. It is an empirical question whether a mathematical or algorithmic property of a set of measurements at a particular level or several levels is correlated with consciousness, and experiments need to be carried out to establish what is the case.

Mathematical or algorithmic correlates of consciousness are valid only for the levels of the physical system where they have been demonstrated to exist. Independent evidence would have to be obtained to prove that information or data integration, for example, can form a SCC set by itself, rather than information or data integration at a particular level of the platinum standard system. For instance, an experiment demonstrating a correlation between neural information integration and consciousness could not be interpreted to show that traffic lights would also be conscious if they exhibited a similar level of information integration. Mathematical and algorithmic correlates of consciousness are similar to physical laws that apply to carefully defined measurements of the physical world. Newton's equation for the gravitational attraction between two bodies provides a good approximation when it is applied to the masses of the bodies and the distance between them, but it is nonsensical if the masses or the distance are replaced by other quantities, such as the surface area or color.

3. Indeterminacies about the Correlates of Consciousness

Ideally we would build up a complete picture of the relationship between consciousness and the physical world by identifying one or more SCC sets in the platinum standard system. In practice there is a substantial amount of indeterminacy in our knowledge about the correlates of consciousness, which is caused by the limitations of our experiments and the technology and resources that are available to solve the problem. At any point in time there will be a number of hypotheses about the SCC sets that are compatible with the current experimental results. Our experimental results create *indeterminacy envelopes* (see Section 3.4) that enclose collections of features of the platinum standard system that are potentially correlated with consciousness.

3.1. *Experiments on the Natural Brain*

Many experiments on the correlates of consciousness can be carried out by observing the conscious and unconscious brain in its natural state using measurement techniques that do not alter or affect its functioning – for example, EEG, fMRI or implanted electrodes. One problem with this approach is that it relies on separations between phenomena that happen to occur naturally, which leaves many ambiguities. For example, experiments on the natural brain might inform us that a high level of information integration between neuron firing events was correlated with consciousness, but it would not tell us whether the electromagnetic waves generated by the neurons or the movements of chemicals across the synapses were the levels of the system at which information integration was important.

Natural experiments have the further limitation that they cannot separate out the different material aspects of the brain. For example, all platinum standard systems naturally contain dopamine, are a particular size and operate on a particular time scale, and these factors cannot be experimentally separated out by observing the brain in its natural state. These indeterminacies could only be addressed by natural experiments if we discovered other platinum standard systems with a different material constitution. For example, if we encountered a species of conscious beings that did not contain dopamine, then we would know that dopamine is not always necessary for consciousness.

3.2. *Interventionist Experiments*

An interventionist experiment deliberately alters the brain or the subject's behavior. The distinction between experiments on the natural brain and interventionist experiments is not clear cut because many interventions could be considered to be natural behaviors of the system. Some examples are as follows:

(i) *Behavior interventions*. Ask subject to look to the left, carry out a particular task or report their conscious states.

(ii) *Chemical interventions*. Administer chocolate, alcohol or DMT; replace blood with an artificial blood substitute.

(iii) *Physical interventions*. Damage brain using sharp instrument or blow to the head; alter electric field using TMS or electric shocks; replace hippocampus with a functionally equivalent silicon chip.

Some of these interventions change the behavioral output of the platinum standard system; others leave it intact.

Behavior-changing experiments can play a useful role in the scientific study of consciousness because they can alter the reports of conscious states and produce a measureable result. Prior to the intervention we can measure consciousness in the platinum standard system through a first person behavioral report. Next we can intervene in the brain, by administering chemicals, delivering a blow to the head, and so on. Then by measuring consciousness again through another first person report we can establish whether the intervention had an effect on the consciousness of the subject.

Behavior-neutral experiments are useless because the reports about consciousness from the subject are exactly the same before and after the intervention. This makes it indeterminate whether the intervention has left the consciousness of the platinum standard system intact, or produced a zombie that unconsciously acts in the same way as before.

Suppose we carry out an interventionist experiment in which we replace part of the subject's brain with a functionally equivalent chip. This has no effect on the subject's behavior because the chip is assumed to provide the same input/output function as the brain area that is being replaced. The only thing that can potentially change in this experiment is our *interpretation* of the subject's behavior. Prior to the experiment we interpreted the speech of the subject as a report about conscious states. By substituting part of their brain with a functionally equivalent chip we have transformed the platinum

standard system into a different physical system that can no longer be unproblematically assumed to be associated with consciousness. We can continue to interpret the subject's speech as a report about their conscious states if we assume that biological neurons are not necessary for consciousness, but if we need to make this assumption, then there is no point in carrying out the experiment in the first place.[f] Similar problems occur with other behavior-neutral interventionist experiments, such as the replacement of haemoglobin with an artificial blood substitute or changes to the brain's temporal and spatial scale.

The futility of behavior-neutral experiments limits our ability to precisely identify the correlates of consciousness because many features of the brain can only be shown to be part of a SCC set through experiments that hold the behavior of the brain constant and vary another factor, such as its material constitution or rate of processing. This limitation could only be overcome if we could develop a way of measuring consciousness that does not depend on behavior.

3.3. *Technological, Ethical and Resource Limitations*

Our current resources and technology also limit our knowledge about which features of the platinum standard system are SCC sets. For example, our current scanning technologies fall far short of full access to all neurons in the brain, and so many experiments on the correlates of consciousness cannot be carried out.

There are also many experiments that could in principle be carried out, but have not been done because of ethical constraints or a lack of resources or funding. Many experiments are only permitted on animals; others cost too much; some have not been carried out because no-one has thought about them or had the time.

3.4. *Indeterminacy Envelopes*

Each experimental result is a fixed point of knowledge that limits the hypotheses about the features of the platinum standard system that could form SCC sets. Follow-up experiments can be used to further reduce the indeterminacy about the features of the physical world that are correlated with consciousness. This situation will be described by saying that an *indeterminacy envelope* surrounds the set of hypotheses about the correlates of consciousness that are compatible with an experiment's results. Each indeterminacy envelope contains one or more collections of properties that form SCC sets, as well as properties that have not yet been demonstrated to be uncorrelated with consciousness. An example of an indeterminacy envelope that could result from a neural synchronization experiment is shown in Figure 4.

[f] There has been a substantial amount of discussion about the replacement of part of the brain by a functionally equivalent chip [Moor, 1988; Chalmers, 1996; Prinz, 2003]. While thought experiments, such as Chalmers' fading and dancing qualia, have been used to make hypotheses about the results of behavior neutral experiments, I have argued elsewhere that we cannot use our imagination to tackle this problem [Gamez, 2009].

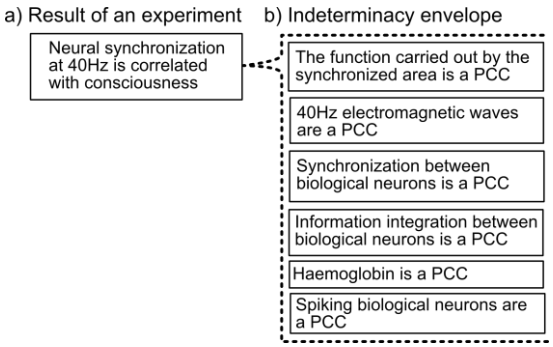


Figure 4. An indeterminacy envelope that could result from a neural synchronization experiment (this example is only loosely inspired by actual experiments). a) An experiment demonstrates that 40 Hz neural synchronization is a SCC set. Whenever this occurs, consciousness occurs; without this form of synchronization no consciousness is reported. b) Indeterminacy envelope enclosing the compatible hypotheses about the correlates of consciousness that arise from this experiment. For example, haemoglobin and high information integration between biological neurons could be sufficient correlates of consciousness.

Further experiments are required to eliminate properties from the indeterminacy envelope that are not correlated with consciousness, and to establish which combinations of PCCs form SCC sets. This process will be referred to as the *reduction* of an indeterminacy envelope and it is illustrated in Table 2.

Over time the indeterminacy envelopes will be reduced as we improve our technology, spend more time and money on the problem and use natural and behavior-changing experiments to identify the links between consciousness and the platinum standard system. However, since some experiments cannot in principle be carried out, it is likely to be impossible to completely reduce the indeterminacy envelopes down to one or more precisely defined SCC sets.

4. Consciousness in Computers

4.1. *The Physical Computer*

Computers are physical machines that can be constructed out of many different materials. For example, Babbage's design for an Analytical Engine was a mechanical system that could be programmed with punched cards, early computers were based around valves and modern computers use circuits etched in silicon to run their programs. When a program is run on different computers it should produce approximately the same output for the same set of input data.[g] As a program is run it produces a particular pattern in the physical world, which varies with the type of computer. In the Analytical Engine a program causes

[g] A program written for one computer might have to be recompiled and possibly rewritten to run on a different computer. A computer's numerical precision will affect the output.

steel and brass rods and cogs to mechanically interact; in computers based on valves or silicon a program produces changing voltage patterns and movements of electrons.

Table 2. Abstract example of the reduction of an indeterminacy envelope. Features A, B, C and D are assumed to form a complete description of the physical system. a) We observe that consciousness is present when physical features A, B, C and D are present. This makes A, B, C and D PCCs and subset(s) of them will be SCC set(s). b) A second experiment is carried out to test whether A is part of a SCC set. Under all possible combinations of B, C and D consciousness is absent when A is absent, and so A must be part of a SCC set, but it cannot form a SCC set by itself because consciousness is not present when only A is present. More experiments are needed to reduce the indeterminacy envelope further and identify which factors in combination with A form one or more SCC sets.

a)

A	B	C	D	Consciousness
0	0	0	0	0
0	0	0	1	?
0	0	1	0	?
0	0	1	1	?
0	1	0	0	?
0	1	0	1	?
0	1	1	0	?
0	1	1	1	?
1	0	0	0	?
1	0	0	1	?
1	0	1	0	?
1	0	1	1	?
1	1	0	0	?
1	1	0	1	?
1	1	1	0	?
1	1	1	1	1

b)

A	B	C	D	Consciousness
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	?
1	0	1	0	?
1	0	1	1	?
1	1	0	0	?
1	1	0	1	?
1	1	1	0	?
1	1	1	1	1

To understand this point better consider the abstract description of a computing machine put forward by Turing [1936]. This consists of a tape of infinite length that can move to the right or left, read and write heads, a processor that maps the input and current state to an action, and an internal memory that holds the state the machine is in (see Figure 5a). The Turing machine is a good abstract model for a computer because according to the Church-Turing thesis, anything that is algorithmically computable can be computed on a Turing machine.

There are many different ways in which the components of the abstract Turing machine can be implemented in a physical system.[h] In this discussion I will focus on different implementations of the read/write heads and the tape that stores the program and is used for output. In a silicon system the tape can be implemented as a number of capacitors that store tiny voltages. The tape can also be implemented as a line of squirrels sitting on a log, with a large squirrel interpreted as a 1 and a small squirrel interpreted as

[h] A video of a mechanical version that uses marker pens and 35mm film is available here: http://www.youtube.com/watch?v=E3keLeMwfHY. A LEGO version is shown here: http://www.youtube.com/watch?v=cYw2ewoO6c4.

a 0. In the squirrel scheme a gripper writes to the 'tape' by moving the squirrels around on the log and a camera system converts the sizes of the squirrels into 1s and 0s (see Figure 5b).

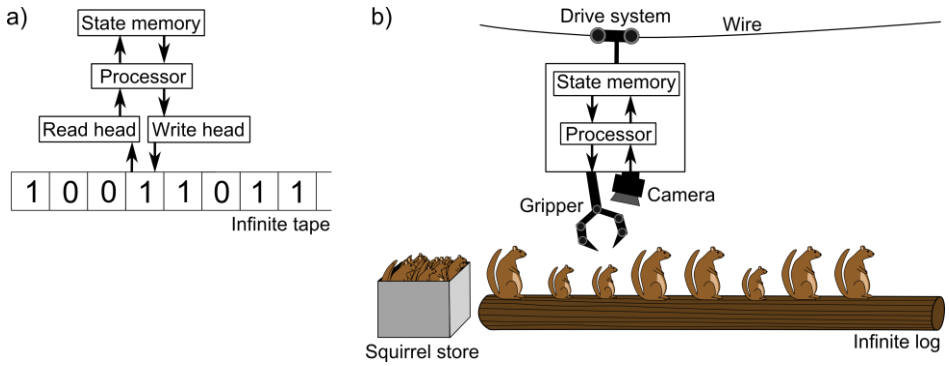


Figure 5. a) Abstract Turing machine; b) Turing machine in which large and small squirrels on a log implement the tape.

Different implementations of the tape lead to different alterations of the physical world as the machine computes - voltages are changed in silicon or squirrels of different sizes are moved about on a log. These changing patterns depend both on the program that is being run by the Turing machine and on the physical implementation of the tape. Other aspects of the computer, such the processor and state memory, can be implemented in many different ways, which lead to further variations in the computer's spatiotemporal physical patterns.

4.2. *Claims about Consciousness in Computers*

Our current experimental results delimit collections of hypotheses about the correlates of consciousness, some subset(s) of which will actually be SCC set(s). In the absence of better information, it seems reasonable to attribute consciousness to any system that fits within the indeterminacy envelopes provided by these experimental results. This thesis is more formally expressed as follows:

> *A spatiotemporal pattern of physical states that matches a candidate set of potential correlates of consciousness within an indeterminacy envelope should be assumed to be associated with consciousness.*

The notion of a candidate set is introduced to handle the fact that some PCCs cannot form SCC sets by themselves because they are also present when the brain is unconscious. For example, a chemical in the brain, such as haemoglobin or NMDA, might be part of a SCC set, but it cannot be sufficient for consciousness by itself because it is also present in the unconscious brain. A candidate set also cannot consist entirely of functional PCCs because of the possibility that any physical system (including the unconscious brain) can be interpreted as executing a particular function over a finite time period.

Suppose it has been shown that high information integration between firing neurons is always present when the platinum standard system reports conscious states and is absent whenever the platinum standard system is unconscious. Furthermore, suppose that the voltage of the neurons and their substrate cannot be experimentally separated out. A neural simulation program is run on a Turing machine whose tape is implemented as capacitors with a 100 mV range. If the execution of the program produced patterns of information integration between the capacitor voltages that were similar to the patterns of information integration between the ~-70 to ~+30 mV neuron voltages in the conscious human brain, then we could justifiably assume that this system is as conscious as a human (see Figure 6). However, on the basis of this experimental result, the same program running on Babbage's Analytical Engine could not be assumed to be associated with consciousness because this would generate a spatiotemporal pattern of physical states that was far outside of the indeterminacy envelope shown in Figure 6a. Babbage's Analytical Engine could only be attributed conscious states if, for example, information integration between punched cards or the rotation of brass cogs, had been shown to be linked to consciousness in the platinum standard system. Since our platinum standard system is constructed out of biological components it is unlikely to be possible to make empirically grounded claims about the consciousness of programs running on mechanically constructed computers.[i]

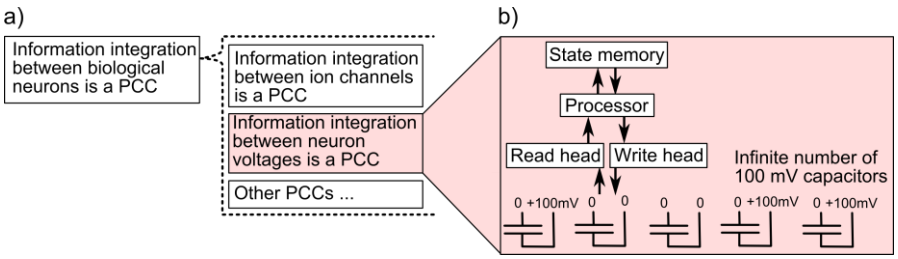


Figure 6. a) Indeterminacy envelope resulting from an experiment demonstrating a correlation between consciousness and information integration between biological neurons. b) A Turing machine whose tape was implemented with 100 mV capacitors could potentially produce patterns of information integration that fit within this indeterminacy envelope.

5. Conclusions

This paper has argued that claims about the consciousness of some machines can be grounded in experimental work on the platinum standard system. Some computers running some programs will create spatiotemporal patterns in the physical world that fit within the indeterminacy envelope of the potential correlates of consciousness in the

[i] The Analytic Engine could be attributed conscious states if data integration was included as a PCC within the indeterminacy envelope, with data being any kind of change within the physical world [Gamez, 2011]. In this case, the Analytical Engine could run a program that produced data integration between the changes in the punched cards that was comparable to the data integration between changes in the neurons in the brain. However, more work is required to determine whether data integration is a form of functionalism or whether it could form a SCC set by itself.

human brain. A claim about the consciousness of a particular machine is not just based on the program that is running, but also on the similarity between the spatiotemporal patterns of physical states in the machine and the experimentally identified correlates of consciousness in the platinum standard system.

I have also highlighted the fact that there is a considerable amount of indeterminacy in our knowledge about the correlates of consciousness in the platinum standard system. This will lead an empirically grounded approach to erroneously attribute consciousness to many systems that are not in fact conscious. As our technology improves we will reduce some of these indeterminacies, but others appear to be fundamental features of our experimental practice and it is not obvious how they could ever be eliminated without a wider selection of platinum standard systems based on different physical and chemical principles or a method for measuring consciousness that does not depend on external behavior.

This grounding of claims about artificial consciousness in scientific data does not prove or imply that particular systems are *not* conscious. While we can only scientifically *demonstrate* that systems similar to the platinum standard system are likely to be conscious, other spatiotemporal physical patterns might also be associated with conscious states.

References

Searle, J. R. [1980] Minds, Brains, and Programs, *Behavioral and Brain Sciences* **3**(3), 417-457.

Tononi, G. and Koch, C. [2008] The Neural Correlates of Consciousness: An Update, *Ann N Y Acad Sci* **1124**, 239-261.

O'Regan, J. K. and Noe, A. [2001] A Sensorimotor Account of Vision and Visual Consciousness, *Behavioral and Brain Sciences* **24**(5), 939-973.

Clark, A. [2008] *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford University Press, New York; Oxford).

Gamez, D. [2007] *What We Can Never Know: Blindspots in Philosophy and Science* (Continuum, London).

Laureys, S., Antoine, S., Boly, M., Elincx, S., Faymonville, M. E., Berre, J., Sadzot, B., Ferring, M., De Tiege, X., van Bogaert, P., Hansen, I., Damas, P., Mavroudakis, N., Lambermont, B., Del Fiore, G., Aerts, J., Degueldre, C., Phillips, C., Franck, G., Vincent, J. L., Lamy, M., Luxen, A., Moonen, G., Goldman, S. and Maquet, P. [2002] Brain Function in the Vegetative State, *Acta Neurol Belg* **102**(4), 177-185.

Ramachandran, V. S. and Blakeslee, S. [1998] *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (William Morrow, New York).

Baars, B. J. [1988] *A Cognitive Theory of Consciousness* (Cambridge University Press, Cambridge; New York).

Gamez, D. [2006]. The Xml Approach to Synthetic Phenomenology. *Proceedings of AISB06 Symposium on Integrative Approaches to Machine Consciousness*, edited by R. Chrisley, R. Clowes and S. Torrance, Bristol, pp. 128-135.

Aleksander, I. [2005] *The World in My Mind, My Mind in the World* (Imprint Academic, Exeter).

Metzinger, T. [2003] *Being No One: The Self-Model Theory of Subjectivity* (MIT Press, Cambridge, Mass.).

Franklin, S. [2003] Ida - a Conscious Artifact?, *Journal of Consciousness Studies* **10**(4-5), 47-66.

Turing, A. [1950] Computing Machinery and Intelligence, *Mind* **59**, 433-460.

Bishop, M. [2002] Counterfactuals Cannot Count: A Rejoinder to David Chalmers, *Consciousness and Cognition* **11**(4), 642-652.

Bishop, J. M. [2009] A Cognitive Computation Fallacy? Cognition, Computations and Panpsychism, *Cognitive Computation* **1**, 221-233.

Block, N. [2006] Troubles with Functionalism, in *Theories of Mind: An Introductory Reader*, edited by M. Eckert (Rowman & Littlefield, Maryland), 97-102.

Chrisley, R. [1995] Why Everything Doesn't Realize Every Computation, *Minds and Machines* **4**, 403-420.

Chalmers, D. [1996] Does a Rock Implement Every Finite-State Automaton, *Synthese* **108**, 309-333.

Balduzzi, D. and Tononi, G. [2008] Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework, *PLoS Comput Biol* **4**(6), e1000091.

Gamez, D. and Aleksander, I. [2011] Accuracy and Performance of the State-Based Φ and Liveliness Measures of Information Integration, *Consciousness and Cognition* **20**(4), 1403-1424.

Lee, U., Mashour, G. A., Kim, S., Noh, G. J. and Choi, B. M. [2009] Propofol Induction Reduces the Capacity for Neural Information Integration: Implications for the Mechanism of Consciousness and General Anesthesia, *Consciousness and Cognition* **18**(1), 56-64.

Massimini, M., Boly, M., Casali, A., Rosanova, M. and Tononi, G. [2009] A Perturbational Approach for Evaluating the Brain's Capacity for Consciousness, *Prog Brain Res* **177**, 201-214.

Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., Tononi, G. and Pearce, R. A. [2010] Breakdown in Cortical Effective Connectivity During Midazolam-Induced Loss of Consciousness, *Proc Natl Acad Sci U S A* **107**(6), 2681-2686.

Tononi, G. [2008] Consciousness as Integrated Information: A Provisional Manifesto, *Biol Bull* **215**(3), 216-242.

Gamez, D. [2011] Information and Consciousness, *Etica & Politica / Ethics & Politics*, **XIII**(2), 215-234.

Moor, J. H. [1988] Testing Robots for Qualia, in *Perspectives on Mind*, edited by H. R. Otto and J. A. Tuedio (D. Reidel Publishing Company, Dordrecht/Boston/Lancaster/Tokyo).

Chalmers, D. J. [1996] *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, Oxford).

Prinz, J. J. [2003] Level-Headed Mysterianism and Artificial Experience, in *Machine Consciousness*, edited by O. Holland (Imprint Academic, Exeter).

Gamez, D. [2009] The Potential for Consciousness of Artificial Systems, *International Journal of Machine Consciousness* **1**(2), 213-223.

Turing, A. [1936] On Computable Numbers, with an Application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* **2**(42), 230-265.