Computational Consciousness

Dana H. Ballard
Department of Computer Science
University of Texas at Austin
Austin, TX
dana@cs.utexas.edu

Abstract

One distinguishing feature of consciousness is the ability to simulate possible futures complete with populations of other minds and their motives. Technically, this ability places demands on the brain, which on an evolutionary scale developed structures more suitable for acting in the near term based on concrete stimuli. Thus important insights as to the nature of consciousness can be obtained by considering how it could be implemented within these structures. Exploring the technical requirements of such an implementation results in constraints on how consciousness could appear in the brain's neural substrate.The crucial questions turn on not so much as to *where* is the location of consciousness but as to *what* are the component functions that, taken together, are requisite for consciousness.

1 Why are we conscious?

We all have a feeling of consciousness that we can talk about and share with others. When we are alone most of us hear a compelling 'voice' in our heads articulating our thoughts, and if you are good at mental imagery, you can have a compelling sensation of different images of people and places; all kinds of things. Such shared sensation has led philosophers and neuroscientists to search of *the mechanisms behind the sensation of consciousness*., the newest effort being (Noe, 2009). But despite all the efforts to uncover the 'feeling of what happens,' or pin down a quale, progress has been modest, to say the least.

Since consciousness must reside in the brain, one thought is that progress might be speeded up if it could isolated to a specific place therein, or at least excluded from a place. Crick and Koch famously declared that V1, the cortical area receiving visual input form the thalamus, was not conscious (Crick and Koch, 1995), perhaps bringing hope to acolytes as V1 is a very large cortical area, and therefore the search problem would be reduced. More recently Milner and Goodale have made a more positive assertion, claiming that consciousness might reside in the temporal cortex,

but not the dorsal cortex (Milner and Goodale, 1995; 2006). The thrust of this paper is to argue that this conjecture cannot be true using two primary lines of reasoning. The first is fairly basic: A large body of experimental data simply does not support this claim.

The second line of reasoning is more speculative and rests on claims as to the role of consciousness. This line presupposes that a helpful question might be to ask: *Why* are we conscious? If there was an answer to that question, then the mechanisms that support it could be cast in a different, possibly more constructive light. This is the tack we take, but with a computational focus. We lay out a precis of human brain function from that perspective and then describe consciousness as an important component. The slowness of neural circuitry means that most of the brain is memory and that behaviors involve primarily memory retrievals. Consciousness is a necessary feature for the brain's process of analyzing, storing and acting on *unexpected* data.

Any student of consciousness has to take note of Damasio's *Descartes' Error*. In that groundbreaking treatise (Damasio, 1994), the emotions are described as a way of promoting fast decisions in a brain with slow neural circuitry. Picking behavioral programs amounts to choosing amongst pre-stired ratings gleaned from uses the body's emotional state. Neurotransmitters do this and the emotions felt by the conscious self are the signature that this has happened. So neurotransmitters select amongst unconscious programs. But from Wegner's experiments, we argue that consciousness is just a program too (Wegner, 2002), and very appropriately described as including as its signature 'an emotion of authorship.' For lots of programs the brain selects, the author/world distinction does not have to be made explicit, but for consciousness, it does. We explore the impact of this in the computational support that would be needed to implement this function (Throughout this paper, computation and program are used with their standard definitions in computer science).

2 Trying to put consciousness in its place

One idea that one might have is that consciousness might have a specific *place* in the brain, a kind of consciousness 'pineal gland', but a brief survey of experimental results can easily disabuse us of this notion. Let us start with Milner's patient H. M. who, when operated on to transect his hippocampus, lost the ability for new permanent memories. Since the 1960s when this was done countless additional studies have confirmed this basic result: the hippocampus together with the amygdala form a neural complex that is responsible for abstracting experiences that will be committed to long term memory. Without these components this consolidation does not happen. One would not say that patients with hippocampal damage are not conscious; among other things they can converse in the present about their past and remember prior events they experienced with friends. But these conversations and other new events cannot be committed to permanent memory. The patients are obviously conscious; its just that the ability to encode new memories is gone

and consequently the ability to 'know' about them and consciously discourse about them disappears as well.

Long term memories are coded in the brain's cortex which is organized into specific areas that contribute to specific functions. The role of these areas begs to be misinterpreted by functional magnetic resonance imaging (fMRI) which uses a differential measure of oxygenated blood flow to pinpoint cortical areas that are more active in certain tasks relative to a control, but what we can safely say is that different areas have roles in facilitating the neural activity all over the brain that leads to a reportable percept. A specific example will make this clear.

Patients can suffer damage to their parietal cortex in a way that leads to a deficit in object-centered locations. In eating food from a plate these patients will eat the food in either the left or right half of the plate (depending on whether the damage is in the right or left hemisphere), leaving the other half untouched. This is remarkable as the damage is not a visual field deficit since the patients can move their eyes and will have gazed at the missing pieces. But apparently when the act of eating is engaged this information is not available. The patients are just not conscious of the 'missing' part of their visual field. These patients' deficit is particularly challenging for the idea that its temporal cortex that is the site of consciousness, as it is *dorsal* cortex that is the site of the damage. Before the damage they are conscious of their full visual field and can discourse about objects anywhere in it; afterwards they are unaware of objects in the damaged object-centered field. Numerous other examples can be included here with the same result: consciousness is an add-on to the function of a specific part of the brain and when that part is damaged, the related part of consciousness is damaged also.

The main weight of the argument for the temporal-cortex-site-of-consciousness hypothesis is that, according to experiments done on the single patient DF, these areas are separate and dissociated. DF could post a card in a tilted slot aperture, putatively the exclusive job of dorsal cortex, but was unable to make a judgement of the aperture's orientation with respect to a standard, putatively the exclusive job of temporal cortex . But could it be that in the case of the parietal patients the dorsal cortex serves as kind of a switch? The argument might be that It is still the case that temporal cortex is the separate site of consciousness, but it does not receive its desired input in these cases. Enthusiasm for this line of argumentation has come from experiments that show that normal subjects who saw the Ebbinghaus illusion - a circle surrounded by larger circles looks smaller than the same circle surrounded by smaller circles - yet reached for the central circle with the correct grasp aperture. So apparently the dorsal cortex is unaffected by the illusion. However (Brenner and Smeets, 1996) show that while the grasp for a similar illusion is corrected, the lifting force anticipates the illusory size of an object. So the effects of the illusion are used in the grip calculation.

The backdrop for thinking of cortex as having distinct areas has such long standing support from anatomical and physiological evidence,particularly fRMI studies, and, given all this evidence, it is easy to exaggerate their spatio-tempoaral separateness. Just for fun, conjure up a mental image where cortical areas are French medieval villages where axonal spikes are carried back and forth between them on horseback.

But this illusion is quickly dispelled when realizing that the fMRI measurements between test and control conditions reveal differences of just 1-3 percentage points; 97-99 percent of the activity is common to both conditions. Furthermore spikes travel at relatively high velocities and can communicate between areas in just a few tens of milliseconds. The point is that, realistically, any two different cortical areas are intimately part of any computation that they are both involved in.

The central claim from these observations is that any attempt to put consciousness in a neural 'place' is unlikely to succeed. The arguments for not assigning awareness to just one of the dorsal and temporal corticies can be made for any comparable set of cortical areas and indeed for the rest of the forebrain. The forebrain (amygdala, hippocampus, cortex, basal ganglia and thalmus) is our mamalian heritage wherein any damage to one of its subsystems produces very predictable deficits in awareness. In this respect, consciousness is like Carroll's Cheshire cat in that when a piece of forebrain fails, a facility associated with that piece, that can impact consciousness in a predictable way, vanishes with it. Thus the experience of consciousness, that we perceive as a unity, below our conscious level is very much a composition of a disunity comprised of well defined capabilities linked to cooperating forebrain subsystems.

3 Consciousness meets computation

If consciousness is just going to be a program, then we had better introduce computation and say a bit about programs, in particular, how programs are oprganized. In the 1950s when computers began their transition form special purpose one-off curiosities to mainstream platforms that handled scientific and business calculations, the idea that the brain was some kind of computer surfaced, but was not greeted with much enthusiasm in research communities beyond promoters of 'expert systems (Anderson, 1983; Newell, 1990).' The idea of computation being central to the brain was very much a metaphor that competed with other physics based notions such as the hologram.

Nowadays though the situation is reversed and information based computational theory is seen as the most plausible scientific program in understanding brain function. The computational framework has sensitized us to the fact that most of the brain is about memories. The slow neural circuitry, over one million times slower than silicon, means that at a basic level behavioral responses have to have been pre-computed and looked up by a fast indexing technique. Before Google, we might have been skeptical that this was possible, but know we know that, in a way analogous to web crawling and indexing the brain works over its lifetime to continually sort its behavioral programs so that they can be rapidly brought to bare on the situation at hand.

One indication of the acceptance of computation was the *Nature Neuroscience* journal's endorsement in a special issue devoted to brain computation e.g. (Hinton, 2005). This enthusiasm is hardly a fad though, and is a product of many modeling breakthroughs whereby complex phenomena have simple explanations once the framework of computation is introduced. Let us review two examples.

How do the memory circuits get formed? A computational answer for the peripheral visual circuitry has found wide acceptance and that views those areas as constructing a code for natural images. Any good code should represent the image to be coded, but computational models have introduced a special spin; the image indeed should be coded, but a desirable property of the code is that its neural substrate should use an economical signaling strategy. Neurons signal in discrete spikes and half of the metabolic cost is in spike generation, so codes that use less spikes and thus save energy are to be preferred. Stunningly those codes when created in computational simulations appear very similar to the experimental observations from neural recordings (Olshausen and Field, 1997; Rao and Ballard, 1999). This methodology implies that the brain has refined *Occam's razor*. Simple codes are preferred, but simplicity must balance the accuracy of the code against the complexity of its description.

How are memory circuits used? The visual codes that we have just been talking about allow the description of visual data but they do not not include a prescription for what to do with it. For that we must turn to another set of research results that use trained monkeys as subjects. In primates the eyes have pronounced foveas so that the resolution at the point of gaze is greatly increased over the periphery. In humans the factor is one hundred. As a consequence, almost all the time, primate eyes use rapid ballistic movements termed saccades to orient the high-resolution gaze point over interesting visual targets. However programming this gaze change takes about a quarter of a second. Monkeys are instructed to hold gaze at a fixed point while looking at a computer monitor. Two line segments appear and the monkey must make a saccade to the end of the segment that happens to be connected the gaze point.

The experiment takes advantage of the fact that the neurons in the early visual periphery are sensitive to small precise photometric edge segments at specific retinotopic locations. The hypothesis being tested is that the monkey finds the end-point by mentally tracing the line segment from the fixation point to its end. The experimenters record form a neuron that is coding a point on this route and find that its spike rate increases in the way consistent with the tracing hypothesis (Roelfsema *et al.*, 2003).

The above twoexamples are united by their use of computation as the core concept that guides the interpretation of data, yet describe experiments that span two separate abstractions. When studying memory formation, the use of the memory is postponed. And when studying the use of memory, the existence of the memory is assumed. Keep in mind that these two abstractions are just two of several that must be assumed to handle the richness of human brain computation. Table 1 shows a more complete candidate computational hierarchy.

What about consciousness? Could it be the sole *deux ex machina* that escapes a computational description? Computation is not an all-powerful theory and has well understood limitations, but all of these concern mathematical infinities and are unlikely limitations for a satisficing animal brain. The odds on bet is that when consciousness is understood at some mechanical level, that level will be isomor-

Level	Computational Model
Operating System	Schedule all behaviors; trade-off debugging and runtime behaviors 'if all the operations are succeeding, stay in runtime mode'
Debugging	Analyze unusual events using off-line mode; why isn't the jelly sticking to the knife?'
Runtime	Pick a suite of behaviors to handle the current situation; 'take the lid of the jelly jar,"pick up the knife'
Behavior	Sensory-motor coordinated actions; 'inset the knife into the jar and gingerly bend and remove it'
Routines	Task-specific tests of the environment; 'locate the purple jelly surface'
Calibration	Encode environmental statistics in specific circuits; filling-in phenomena: color constant neural circuitry
Neural	Models of specific neurons in circuits; excittaion and inhibition
Synapses	Models of a neuron's components; roles for different neural transmitters; neural spike spike signaling; basic units of memory

Table 1: A proposal for the brain's use of computational hierarchies. In the task of making a sandwich different sub-tasks are described at different levels of abstraction.

phic with computation. Furthermore, as some kind of program, it must fit in the computational hierarchy.

4 Abstraction hierarchies

As Newell pointed out, any complex system that has admitted of a description has incorporated the use of the notion of hierarchy and the brain is unlikely to be an exception (Newell, 1990). Hierarchies have two prominent properties:

1. As the description becomes more abstract its components necessarily run slower than those of the lower level and take up more physical space.
2. The more abstract description omits details from the lower level.

A ready example from silicon is that of digital circuitry. The dependence of current on voltage across a gate is continuous, but digital circuitry abstracts that into two levels. The speed of such circuitry is governed by the time it takes for the levels to change between one level and another. Here the continuous voltage value is abstracted away into a binary code. In a similar way computational models of the brain are different at different levels as the primitives at each level are correspondingly different too. Just like silicon computing, brain computation has to be organized into hierarchical levels, each of which groups primitives from the level below into equivalences. The average cortical neuron receives input from two to

ten thousand other neurons and yet at any moment summarizes those inputs into a single spike or silence. Higher levels make similar kinds of abstractions. So a key question is: If consciousness is a program, who does that program project into the various computational abstraction levels?

Searle, arguing against artificial intelligence models, and by extension computation, use a now-famous example of the Chinese Room. A person who knows nothing of chinese is in a room where chinese sentences in characters appear at an input slot. While the person knows nothing of chinese, he or she has access to a set of instructions in english that are of the form "if you see these characters output these characters." Supposedly the person is a model for the computer, mimicking linguistic behavior but understanding nothing. But here is where the trouble starts: Is the person in the room actually a human or a computer (or a neuron)? Depending on how one answers this question there are different logical consequences and therein lies the difficulty.

The danger of confusing abstractions is easier to understand if we change venues for a moment and talk about computers and their abstractions. One can program in MATLAB, assembly language or microcode, but each language is cast at a different level of abstraction and each involves a different level of familiarity with the underlying computer that runs the programs. MATLAB requires no knowledge of the machine at all, whereas assembly language requires an elemental model of basic random access architectures and microcode requires an intimate knowledge of a particular machine's low-level hardware instruction set. The point of these levels is that they cannot be mixed. Once an abstraction level is picked one must stay within it; MATLAB and microcode cannot be mixed.

What seems so clear when couched in silicon computational terms somehow does not always survive the translation into philosophical arguments about human computation. To return to our example, Searle's attempt to gain purchase depends on transits across abstraction boundaries. Is the occupant in the room a human or computer? It has to be one or the other; to not insist on a choice leads to what I would term *the abstraction con*, which comes up repeatedly in discussions about consciousness. Consider Escher's famous drawing of hands drawing each other. Is it just a drawing, or is it really a picture of something that could happen? Perhaps when you don't look, the hands transform into versions of television's Addams Family's *Thing* and take a few pencil strokes. But let us not let this kind of magical thinking displace the magical thinking, aka the abstraction con, upon which the drawing's kick depends: "Its a pair of hands" oscillating across the abstraction boundary separating "Its a drawing of hands."

Lets return to consciousness, but now wonder how consciousness gets represented in an abstraction hierarchy. Could it be that A) consciousness is an abstraction that appears at a given level but has no discernible trace at the level below? The alternatives are that B) consciousness is only manifested at lower levels or else C) it has a trace that spans all levels. Most researchers would opt for at least option A given our phenomenological experience of hearing our own internal directive voice, but there is obviously also enthusiasm for option C as well. Koch for one has advocated a search for the *neural correlate of consciousness* (Koch, 2003).

Of course unless one is a dualist there will be a sense in which the answer must be 'C.' But the huge cautionary note from the example of silicon computation is that one needs to understand the various abstraction levels in their own terms as well as the process of translating between them. Failing to keep them separate and mixing abstraction levels only results in confusion. This assertion can be seen as in sympathy with Lamme's view (Lamme, 2006). He points out that the various claims as to consciousness can seem confusing if they do not respect basic neural organizations. Our extension is that those organizations should be organized into computational hierarchies, because the brain is doing computations and hierarchies are the only way we know of organizing complex computational systems.

However, given that the current state of knowledge of the brain provides only an outline of plausible abstraction hierarchies, this leaving an enormous amount of work to do. At this point we will content ourselves by elaborating the differences between what we have termed in Table 1 the 'Runtime' level and the 'Debugging' level, the latter being the level where the contents of consciousness are most evident.

5 Will work for Dopamine

Although what we are calling the debugging level of abstraction will be the most important for understanding consciousness, to appreciate this level it is important to understand the abstraction level beneath it. How do programs get established in the first place? We know a bit about the programs themselves in that, in broad outline the cortex stores elaborate states of the world and the basal ganglia sequences through those states and triggers actions. While silicon computing sequences through stares at an incredibly high 2 gigahertz rate, the basal ganglia probably works a about 0.3 hertz, or a billion times slower than silicon. Of course the bandwidth of the cortex in terms of its parallel processing more than makes up for the slowness. We also know a bit about how programs are formed in that the job of the amygdala is to filter out what is important and the job of the hippocampus is to slice and dice those components into pieces that are compatible with what the cortex and basal ganglia have stored. So although we cannot say too much about the details, the big picture of what programs are is sketched.

Given that we have programs, who - or rather *what* - is the programmer? Answering this question requires understanding at a basic level what programs do for us and that is prediction. Ultimately we need to be able to mate successfully and along the way to survive and the way to do that is to be able to predict the future. And one hugely important way to do that is to save what has happened in the past along with its value so that if the situation recurs one can estimate its outcome and value before its conclusion. The past is prologue.

In this process it helps to keep in mind that the brain's programs are not literal parts of the external world, but just internal models of that world. Ramachandran and Sachs have beautifully describe cases where, owing to some kind of brain or body injury, that model is divorced from the true picture of the world, but the brain's conscious owner is unaware of the schism (Ramachandran and Blakeslee, 1999; Sacks, 1985). The point is that these cases of brain injury tell us about the healthy

brain's structure and that is one of building and maintaining representations of the world and programs for extracting reward from it. This last point needs elaboration, for although we act in the world to successfully survive and mate and along the way achieve measurable rewards in terms of behaviors that satisfy us in one way or another, to accomplish all this the brain's *models* need 'pretend' rewards or secondary rewards that stand in for the real thing. The main neurotransmitter that signals secondary reward is known to be dopamine (Schultz, 2002), which in honor of Europe's common currency, we will term the 'neuro.' Its power is very much experienced by cocaine addicts who engage in behaviors that trigger dopamine release. In fact most addictions can be conceived in terms of behavioral shortcuts to dopamine release. Most of us though are calibrated, in the sense that we can engage in socially-acceptable behaviors that the brain can translate into a dopamine reward estimates. How can one choose between behaviors A and B? Simple. The brain can retrieve their dopamine estimates and pick the most rewarding.

Up to this point we have described the featured non-persona of philosophers and B movies, the *zombie*. The zombie brain has an enormous raft of programs that negotiate to the hypothalamus to be valued in neuros. With this common currency the brain can pick the most valuable for execution; no conscious thought required.

What is the zombie state? Almost all car drivers have experienced it. One drives miles of a familiar route, gets distracted and then at some point is conscious of the driving venue as well as conscious of remembering nothing about a huge driving segment wherein one was guiding the car, obeying traffic lights avoiding pedestrians etc. Perhaps you were engrossed in conversation with a passenger while you were following the familiar route, not the one you needed to follow for the particular passenger. During this period, for the driving behavior, the zombie programs wwere running. There was no use of any special monitoring because the states that were directing behavior had pre-coded expectations of consequences that were constantly being met; it is only when this does not happen does something non-zombie have to be done. Of course in Philosophy there always seems to be someone on the other side of the fence. Searle, again:

> "It is true, for example, that when I am driving my car "on automatic pilot" I am not paying much attention to the details of the road and the traffic. But it is simply not true that I am totally unconscious of these phenomena. If I were, there would be a car crash. "(Searle, 1990)

This was written before the 2005 DARPA Grand Challenge that had five vehicles successfully complete a complicated desert trail loop autonomously. Hopefully we can all agree that those vehicles were not conscious. The point is that the vehicles are essentially limited to zombie driving and did not crash.

David Foster Wallace captures the zombie state brilliantly in his essay *How Tracy Austin broke my heart* (Wallace, 2007). where he asserts that professional athletes have difficulty describing their feats precisely because *the descriptions are no longer accessible* when the over-learned skills are compressed in the zombie brain:

> "The real secret behind top athlete's genius, then, may be as esoteric and obvious and dull and profound as silence itself. The real, many-veiled answer to the question of just what goes through a great player's mind as he stands at the center of hostile crowd-noise and lines up at the free-throw that will decide the game might well be: *nothing at all* ... there's a cruel paradox involved. It may well be that we spectators, who are not divinely gifted as athletes, are the only ones able truly to see articulate, and animate the experience of the gift we are denied. And that those who receive and act out the gift of athletic genius must, perforce, be blind and dumb about it – and not because blindness and dumbness are the price of the gift, but because they are its essence." *pp 154-155*

Make our inner zombie the athlete and our consciousness self the spectator and you have one of the best compact descriptions of the relationship between the two ever written. The quote also highlights another point. The zombie is usually disparaged as inferior, but the metaphor shows off its true relevance: Zombie skills are those that have been honed to near perfection through experience.

The 2005 DARPA Grand Challenge shows that, while the detailed programs that the brain might use for its Runtime abstraction level have yet to be pinned down satisfactorily in the neural substrate, the computation that does the job is fairly well understood and is the subject of standard texts e.g. (Bishop, 2008; Thrun *et al.*, 2005).

6 What is consciousness for?

Zombie programs depend on their model of the world being a very good fit; all the contingencies that can occur have been seen a significant amount of times and their responses are coded. Thus the search for alternatives has been done and remembered. But before that happens, as models are being constructed, the statistics of the model need to be gathered. This is the role of consciousness. The best analogy for this is the process of debugging a program on a silicon computer. Once a program has been debugged it can be used in a zombie state where it is not tampered with further, but before that state is reached, the programmer must stop the program, try alternate versions of it, and test them to see if they behave according to expectations. Like consciousness, this process, is much slower and has substantial off-line portions of time.

However an important contrast is that, unlike debugging, there is no programmer in the human consciousness, just a neural search program that is trying to fit a model to data. In this fitting process, it is central to distinguish between effects that the human agent produces and effects that the rest of the world produces. For this point we return to one of the best characterizations of consciousness, that of (Wegner, 2002). In his famous *I Spy* experiment, the interval of time between when subjects think of an action and when they do it is manipulated. When asked to rate their actions on a fourteen point scale from ' I caused it' to 'It just happened,' subjects rate the period that has the thought preceding the action by a half a second as the

most causal. The stunning suggestion is that the act of consciousness itself is just one more model that the brain uses, and its 'fit' depends on the temporal relationship between the brain's machinery and the body's actions in the external world. Wegner characterizes that result of a good fit between the two as 'an illusion of authorship.' to which I would add: produced and required by the neural program in its search process.

The search process of consciousness has a special and remarkable technical problem to deal with and that is that it must share the same neural hardware used by the zombie programs. This was brought to light by measurements of the firing patterns of cortical 'mirror cells'(Rizzolati and Sinigalia, 2008). Rizzolatti et al discovered when recording from a monkey's cortex in motor area F5, that the cells would respnd when the monkey picked up a raisin but also when another monkey did so or when the experimenter did so. The profound implication of this is that there is finally concrete evidence that the monkey and by extension humans use one set of neural 'hardware for representing all these different events. What this suggests is that a large part of the experience of consciousness may be generated by mental simulation using the same neural circuitry that is used for everyday action, a point made much earlier by (Merleau-Ponty, 1962) and amplified recently by (Barsalou, 1999).

Searching by simulation using existing knowledge has a lot of advantages, the principal one being that, in the act of exploring the effect of changing one variable, the rest of the program can be used as is since it is already in place. Thus the search process can systematically try out slightly different variations of the simulation program without extensive re-programming of all the components. However this boon comes with an attendant disadvantage and that is that some bookkeeping must be done to keep track of what as simulation and what is reality. To go back to Rizzolati's example, the monkey's brain must somehow keep track of the difference between the experimenter doing the action - Debugging mode - or the monkey itself monkey doing the action - Runtime mode.

7 Tagging

The computational role for consciousness is to be the mechanism that does this bookkeeping, but this point requires quite a bit of elaboration. To give it a name lets call this ability *tagging* in the sense that we can tag the actions of the brain according to the particular agent whose activities are being represented. We use the ability to tag in different ways:

1. We may need to reflect on the past or the future. In these cases the same brain hardware handles the visual perception of past and future, but, as they are not in the present, they must be tagged.

2. We may need to reflect on another's motives. In this case the other person must be simulated on our brain hardware, but of course is not us and must be tagged to keep the distinction apparent.

3. Of course it gets complcated if we have to consider what another person is thinking that we are thinking or when we must imagine the other person's future actions, but the overall ability still rests on tagging or the bookeeping to record that the brain's activity is a simulation of reality, not reality itself.
4. Multiple personality disorders are a failure of the tagging system. Often in order to protect itself from the consequences of remembering early abuse, the brain will adopt an alternate personality that did not experience it. In this case the tagging is dropped tp protect the user.

We can appreciate the usefulness of tagging with an analogy to program debugging. When a computer program is running it is in the 'zombie' state where it sequences through a set of instructions until the end of the program. But when something goes wrong the programmer interrupts this sequence, slows down the execution and interrogates the values of particular variables, seeking to explain an unexpected part of the execution. In the analogy, the programmer is the conscious homunculus. (Humphreys, 1992) In Runtime Mode the zombie mindlessly executes a set of responses to coded states of the world and in doing so its programs get translated into all the more concrete models of the neural substrate. In Debugging Mode the identical neural hardware is used and most of the operations are indistinguishable from their Runtime Mode instantiations with one huge exception: the system 'knows,' i.e. has the machinery to understand that, for portions of the Debugging Mode program, it is simulating the results. What can complicate the matter even further is that Debugging Mode may only be required for small portions of a large sequence.

the above observations have a huge implication for understanding consciousness through psychological experiments. Since the conscious mode -Debugging- is utilizing the same underlying neural substrate as the unconscious mode - Runtime - the former is actually being used as a probe to discover information about the various neural circuits rather than necessarily revealing information about itself, a point also made by (Block, 2007).

8 In Summary

Our principal thesis has been that much previous work has concentrated on finding a mechanistic explanation of how it *feels* to be conscious and, specifically, in trying to trace that feeling to the neural substrate. While it might be possible to do this, it also might be an extremely difficult task, equivalent to questioning the value of the mass of a proton. Why is it the specific value that it is? In the same way consciousness, to work, may have to produce a signature, that distinguishes the agent from the surround. Why it produces the feeling that it does as opposed to another may not be an answerable question, however compelling it is to ask it. Ramachandran cannot resist and makes it one of his central unanswered questions about consciousness, phrasing it another way: How can we distinguish the consciousness that we readily accept that others have, which he terms third person consciousness, from the consciousness that we have, which he terms first person consciousness? The difficulty in resolving this distinction can be appreciated by recalling the all-important respect for abstraction boundaries in our discourse. Thinking about another's con-

sciousness is at a more abstract level than living our runtime (zombie) existence. in the same way, if we want to characterize our own consciousness, we have to 'tag' it so that now we are effectively debugging ourselves as a third person simulation. To *simultaneously* engage the runtime environment and experience our own consciousness requires that we blur an abstraction boundary, which we have declared taboo.

Of course taboos can be violated, and even celebrated, as they are in (Hofstadter, 2007). By violating an abstraction boundary one can create an aptly named 'strange loop,' where one ends up unexpectedly at a lower level of abstraction, with the result that strange things happen. Its not that you cannot do these things, but, from a computational standpoint, you should not.

The question of why consciousness exists may have a ready answer when compared to its useful partner, the unconscious. Unconscious programs represent repeated andreliable interactions between the agent and the world that can be coded invariantly for each case. Consciousness is used to direct the search for new programs, and in that search it becomes essential to distinguish the agent from the surround. In a strong sense this paper has advanced nothing particularly new, since its core ideas have been pioneered by many other researchers, such as Wegner, Damasio and Hymphreys. Hopefully, its main value might be in steering the quest for understanding consciousness towards more accessible computational questions.

References

[Anderson, 1983] J. Anderson. *The Architecture of Cognition*. Harvard Univeristy Press, 1983.

[Barsalou, 1999] Lawrence W. Barsalou. Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22:637, 1999.

[Bishop, 2008] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2008.

[Block, 2007] N. Block. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30:481–499, 2007.

[Brenner and Smeets, 1996] E. Brenner and J. B. J. Smeets. Size illusion influences how we lift but not how we grasp and object. *Experimental Brain Research*, 111:473–476, 1996.

[Crick and Koch, 1995] F. Crick and C. Koch. Are we aware of neural activity in primary visual cortex? *Nature*, 375:121–123, 1995.

[Damasio, 1994] Antonio R. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. Hanover, 1994.

[Hinton, 2005] Geoffrey E. Hinton. Computation by neural networks. *Nature Neuroscience*, 3:1170, 2005.

[Hofstadter, 2007] Douglas Hofstadter. *I am a strange loop*. Basic Books, 2007.

[Humphreys, 1992] N. Humphreys. *A History of the Mind*. Vintage, 1992.

[Koch, 2003] C. Koch. *The Quest for Consciousness*. Roberts, 2003.

[Lamme, 2006] V. Lamme. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10:494–500, 2006.

[Merleau-Ponty, 1962] Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge & Kegan Paul, 1962.

[Milner and Goodale, 1995] D. Milner and M. Goodale. *The visual Brain in Action*. Oxford University Press, 1995.

[Milner and Goodale, 2006] D. Milner and M. Goodale. *Epilogue:twelve years on*, volume The Visual Brain in Action (Second Edition). Oxford University Press, 2006.

[Newell, 1990] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

[Noe, 2009] Alva Noe. *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. Farrar, Straus and Giroux, 2009.

[Olshausen and Field, 1997] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

[Ramachandran and Blakeslee, 1999] V. S. Ramachandran and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. Harper-Collins, 1999.

[Rao and Ballard, 1999] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, 2:79–87, 1999.

[Rizzolati and Sinigalia, 2008] G. Rizzolati and S. Sinigalia. *Mirrors in the Brain*. Oxford University Press, 2008.

[Roelfsema *et al.*, 2003] P. R. Roelfsema, P.S. Khayat, and H. Spekreijse. Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences USA*, 100:5467–5472, 2003.

[Sacks, 1985] Oliver Sacks. *The Man Who Mistook His Wife for a Hat*. Simon and Schuster, 1985.

[Schultz, 2002] W. Schultz. Getting formal with dopamine and reward. *Neuron*, 36:241–263, 2002.

[Searle, 1990] J. Searle. Who is computing with the brain? *Behavioral and Brain Sciences*, 13:632–634, 1990.

[Thrun *et al.*, 2005] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[Wallace, 2007] David F. Wallace. *Consider the Lobster*. Abacus, 2007.

[Wegner, 2002] Daniel Wegner. *The Illusion of Conscious Will*. MIT Press, 2002.