Consciousness as self-function

Article *in* Journal of Consciousness Studies · January 1997

CITATIONS
32

READS
49

1 author:

Donald Perlis
University of Maryland, College Park
149 PUBLICATIONS **1,627** CITATIONS

Journal of Consciousness Studies, **4**, No. 5–6, 1997, pp. 509–25

CONSCIOUSNESS AS SELF-FUNCTION

Donald Perlis, Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA. Email: perlis@cs.umd.edu; http://www.cs.umd.edu/~perlis

Abstract: I argue that (subjective) consciousness is an aspect of an agent's intelligence, hence of its ability to deal adaptively with the world. In particular, it allows for the possibility of noting and correcting the agent's errors, as actions performed by itself. This in turn requires a robust self-concept as part of the agent's world model; the appropriate notion of self here is a special one, allowing for a very strong kind of self-reference. It also requires the capability to come to see that world model as residing in its belief base (part of itself), while then representing the actual world as possibly different, i.e., forming a new world-model. This suggests particular computational mechanisms by which consciousness occurs, ones that conceivably could be discovered by neuroscientists, as well as built into artificial systems that may need such capabilities.

Consciousness, then, is not an epiphenomenon at all, but rather a key part of the functional architecture of suitably intelligent agents, hence amenable to study as much as any other architectural feature. I also argue that ignorance of how subjective states (experiential awareness) could be essentially functional does not itself lend credibility to the view that such states are not essentially functional; the strong self-reference proposal here is one possible functional explanation of consciousness.

Introduction

This paper outlines the beginnings of a theory of mind based in large part on the notion of self. However, I do not take self as a fundamental irreducible notion, rather I seek to elucidate self in computational terms. The picture I wish to present has the following outline: Mind, consciousness/awareness, and qualia, are notions coherent only in relation to the concept of self, which in turn can be given functional and computational characterization. On that basis can be built the beginnings of a general theory that at least is not obviously incapable of explaining the former notions, and that holds suggestions for how to go about finding such explanations.

My theses in this paper are, roughly, (i) that consciousness is synonymous with self, and self with a special sort of self-modelling I call strong self-referential computation; (ii) that there is an indivisible 'something it is like to be' a strongly-self-modelling (or referring) entity, constituting a sort of ur-quale,* and without which no experience, no subjectivity, is possible; and (iii) out of the ur-quale can arise fancier sorts of ineffable qualia: colours, emotions, and so on.

Since there is so much disagreement on basic terminology, it will be helpful to set out at the beginning how I understand certain terms.

Paradigms and Definitional Gestures

Concepts mature as we learn more. We cannot expect to grasp the nature of the mind at the outset, nor to have adequate definitions. As we study conscious systems, e.g. brains, we may find out about new structures and processes that will utterly amaze us because they are so different from anything we had imagined before. Who in 1900 had thought of self-replicating molecules? — yet in principle it would have been possible to do so. Who in Flatland thinks of 3-dimensional space? — yet it is possible

* 'Ur' is used here in the sense of *prototypical* or *fundamental* or *primitive*.

in Flatland to do so, indeed to give a mathematical description of its properties, once someone has the idea.

I conjecture that we may find in the brain special amazing structures that facilitate true self-referential processes, and that constitute a primitive, bare or ur-awareness, an 'I'. I will call this the *amazing-structures-and-processes paradigm*. I take it that it is shared by many working in the allied computational cognitive neurosciences, e.g. Baars, Crick, Damasio, Edelman, Harth; but not Block, Chalmers, Deikman, Dennett,[1] Penrose. Such entities once understood may indeed 'stand up and grab us' as obviously self-experiencing and hence possessed of experiential awareness. Such an outcome would then close the explanatory gap.

In order to relate the above to other pieces of the consciousness debate, I provide some definitional guides. Block (1995) notes that the term 'consciousness' is used in many distinct ways; it is Block's *P-consciousness* — i.e., phenomenal consciousness, or subjective experiential awareness — that is the subject of our concern. It is worth noting that Block, along with Crick and others, seems to regard it as almost obvious that P-consciousness is not the same thing as self-consciousness, in apparent strong contrast with the position I shall argue. One small part of my argument is a rather trivial one that I shall reveal here: subjectivity involves a subject, a self, a 'me', simply on the terminological face of things.

We can provide another characterization of P-consciousness in a paraphrase of Nagel (1974): An entity is conscious if it is in a state such that it is 'like something' for it to be in that state. This seems to help us separate examples of consciousness from examples of non-consciousness. Rephrased again, consciousness is experiential awareness, and experience is like something for the experiencer; what it is like, how it feels, is the experience. This may seem a bit circular, but it is usefully suggestive. We will see something analogous (strong-self-reference) come up in an important role in what follows.

Qualia are the individuating aspects of experiences that allow us to distinguish experiences from one another; e.g. we distinguish red from blue by its redness. We distinguish a square from a triangle by its four-sidedness as an aspect of the experience of seeing a square. These examples illustrate that some qualia are partially effable (e.g. square-percepts) and some are not (e.g. redness). Qualia are not restricted to visual experiences: they are found also in emotions, touches, smells, sounds, and so on. They may also occur in thoughts, since a thought has a particular aspect that distinguishes it from, say, a touch: to use Nagel's terminology, it is like something to be thinking, it feels different from not-thinking.[2] A quale is an aspect of an experience such that it is recognizably different (i.e. like something different) for it to be absent.

So defined, qualia occur only as aspects of consciousness; but consciousness might not, at least prima facie, be accompanied by qualia. We will explore this further below.

[1] Dennett (1991) however is hard to pin down. He undoubtedly agrees that the brain achieves an amazing feat of information-processing, and that this is all there is to consciousness. But he also argues that consciousness is an illusion, leaving the impression that once all detail is worked out nothing so very novel will have been discovered.

[2] Thinking feels different from, say, drinking, or from listening to music, or from writhing in pain, or from lying awake but unfocused. And thinking about a calculus problem in a general way feels different from trying to solve it, and in turn different from comparing two solutions.

Functional explanations

Let us recall the challenge by Chalmers (1995, loosely paraphrased): where's the beef (qualia) in many of the proposed 'scientific' accounts of consciousness? He in particular argues that consciousness, unlike the usual objects of study in science, is not itself a behaviour (function or process), and that attempts at scientific explanation simply replace consciousness by some function or process that may result from or contribute to — but is itself not — consciousness. It is the subjective sense of awareness, the qualitative 'feels' of consciousness, pains and pleasures, vivid experiences, that seem to be lacking in the processes or functions.

But why is consciousness not a process, simply an amazing one well beyond the poor pale processes we are currently able to envision, much as a living cell is an amazing but physical structure-and-process far beyond what any chemist could have envisioned in 1900? To be sure, consciousness is something special, beyond cellular chemistry. It's amazingness will be different, perhaps far more astounding, than that of the cell. But we should not assume we currently see the ultimate limits of what processes or functions can entail.

Chalmers argues that past perplexities as to the nature of living things, for instance, were ones of behaviours and thus did not present the same kind of fundamental challenge as does conscious experience: to be a living thing is to perform certain kinds of function, such as reproduction, adaptation, metabolism.

But it was not always thus: the apparent purposiveness of (many) living things did not once seem to be a function. It is only now with the enormous success of the evolutionary, biochemical and computational paradigms that we can at last see biological purposiveness as a kind of evolved electro-chemical computational process: the wasp builds a nest 'purposively' but not in ways that call for explanations beyond ordinary causal mechanisms. Similarly, conscious experience might turn out to be a function, such as an appropriate form of self-modelling, which we might come to understand when we are further along in the quest. How could 'mere' self-modelling have a feel? That remains to be seen, and I will make some tentative suggestions here.

Looking ahead to the thesis — defended below — that there is an ur-quale necessary and sufficient for consciousness, and that it is a special but effable sort of strongly-self-modelling computational process, we can ask: How are fancier qualia to be recaptured, how are the ineffable to be added-on to the effable ur-consciousness? How is it that a presumably mechanical process of distinguishing self from other, itself based ultimately on geometric (spatio-temporal) distinctions in the nervous system, can be green-perceiving rather than red-perceiving? How can geometric distinctions amount to the differing feels of red and green?[3]

But perhaps such feels can be found in a deeper analysis of colour experiences, as based on the self, and on emotional factors such as fear, envy, rage, despondency (yellow, green, red, blue). Emotions[4] in turn might prove to be bodily conditions that are also self-based (fear might involve a condition of unwanted reduction in self-governance). Wants might involve recognition of physiological needs and what might satisfy them. Needs may be perceptions of built-in drives as well as of the organism's inability to act so as to disobey those drives.

[3] Not to be confused with mere wavelength differences.

[4] See O'Rorke and Ortony (1994) for a distinct suggestion.

Another challenge to the amazing-structures-and-processes paradigm (see McGinn, 1995; Shear, 1996) is that whereas physical (process) phenomena occur in spatial arrangements, subjective phenomena do not. A full discussion would take us far from the main theme of this paper, but the following suggestions may serve to indicate that there is more spatiality to subjectivity, and less to physicality, than meets the eye. Our percepts very often have spatial arrangement: my tooth-ache does seem higher than my stomach ache. And while my thought that Nixon was a scoundrel may not be above or below my thought that he was a Quaker, there is a sort of metric tying them together: I turn attention from one to the other, then back again, as if moving through a mental space. Moreover, a thought is not indivisible, it is a complex built out of parts arranged among themselves, e.g. subject and predicate, with a specific linkage that may be metrical in significant ways. Finally, the physical world is not all spatio-temporal: gravity is weaker than the electromagnetic force; but gravity is not above or below electromagnetism. Physically real entities of suitable abstraction need not be spatially arranged.

The amazing-structures-and-processes paradigm then is directly in opposition to Chalmers' view. We proceed to explore the paradigm by reconsidering the role of qualia in consciousness.

Getting by without qualia?

It may be that qualia are not essential for consciousness after all. This can be argued by direct appeal to our experience. We can certainly be conscious with our eyes closed, or indeed with no eyes at all. We can lack a visual cortex, auditory cortex, and certain other portions of our brains, and still be conscious. And even without missing brain parts, we can simply be in a state of not having any of the qualitative experiences so common in discussion of consciousness: touch, sight, sound, smell, taste, pain, pleasure, and so on. For any particular experiential quality we may mention, it seems that we can quite clearly be without it and yet be conscious. If so, then why cannot we also be without any qualia at all and yet be conscious? Is there perhaps a special sacrosanct quale that must remain, an ur-quale? If so, the ur-quale would be the only quale necessary and sufficient for consciousness, all others being contingent. To be more specific, suppose an experience with quale Q to be modified so that Q is missing from the experience. For instance, suppose the redness of an apple-perception disappears and the apple is seen as a shade of grey. There are still qualia present in the modified experience, namely brightness and shape qualia among others. Now suppose the brightness and shape qualia absent as well; in fact suppose the experience is 'reduced' simply to that of knowing there is an apple ahead. Still qualia remain in the experience, for it is like something to experience knowing an apple is ahead, even without seeing it; and we can distinguish knowing about an apple from other experiences. What if now we remove that knowing as well, and are left with bare experience with no distinguishing features to single out: no apple, no thoughts, just bare experiential awareness, pure consciousness.[5]

Let us pursue this a little further. Suppose such a state of experience is possible. Then can it too be distinguished from other experiences? Can one imagine it removed? Is it imaginably absentable? Since all experience would be gone then, it

[5] See Deikman (1996) and Shear (1996) for evidence of such a state coming from various cultural traditions. Below I offer some contrasts of detail in our respective views.

would seem that it is not like anything at all for such a pure consciousness to be gone. But if we cannot even imagine it absent, i.e. what it is like for that experience to be absent, this seems to fly in the face of the property of distinguishability stated as our definition of qualia. This might mean that pure consciousness, if it exists at all, is not a quale. But how can an an experience be like something and yet not be distinguishable from other experiences (i.e. possessed of qualia)?

We are at a seeming impasse. Yet there is a way out: it may be that an experience can be distinguished in and of itself, by its own intrinsic character, rather than by comparison to something else. While everyday qualia such as redness and squareness, perhaps even thinkingness, are distinguishable from one another and also by their presence or absence, perhaps bare consciousness is in and of itself a self-distinguishing process, a process that takes note of itself. If so, it could still be considered a quale, the ur-quale, what it's like to be a bare subject, and distinguishable from other fancier experiences simply in virtue of the additional qualia attending the latter but not the former.

What might this be? That is unclear, and yet it has a certain familiarity to it, in the sense that our experiences are, after all, known to us to be ours, private, personal. If we strip away incidental properties due to our situated histories, do we end up simply with a 'me', a bare awareness, not in touch with objects or environment, but simply having a self-presence *simpliciter*? What is it for a process to distinguish itself, and from what is it distinguished? We will return to this below. The potential beauty of this is that it is not totally implausible that such a kind of self-perceiving process may be computational. Much of the rest of this paper is a tentative exploration of that possibility.

Robots, perceptual awareness, perceptual management, mantises

Some (e.g. Crick) hope to find keys to the nature of conscious experiential awareness by studying particular behaviours such as visual perception. While such work is valuable and important, I think there is serious question whether it alone will get us very far in this issue. One reason is that consciousness can go on quite nicely in the complete absence of vision. Another is that visual processing can go on without consciousness. I think that expressions such as 'perceptual awareness' are risky ones that mislead us into thinking, for example, that visual processing in and of itself is a kind of consciousness. When we are conscious, there can be visual qualia present as part of that experience, but that is not to say that visual processing constitutes visual qualia, the 'what it's like' to be seeing.

Indeed, robot vision systems today routinely perform complex tasks of visual perception, even visuo-motor coordination, binocular focusing of cameras, motion-tracking, and so on. It is startling to observe such systems in operation, hard to avoid the uncanny sense of being watched, their paired robot eyes swivelling suddenly as you walk across the room. Yet no one seriously regards these as in any way conscious or aware of anything at all. I suspect that the impressive physiological work being done on the vision systems of mammals is going to show us structures and processes much like those of today's robots, and little at all about awareness.[6] This is not to say, of course, that consciousness is not a physiological phenomenon: it is, but one that is

[6] That is, one needs to go to much deeper and higher levels of processing to get to consciousness.

at quite a different level from processing of perceptual data. I prefer to call the latter 'perceptual management' rather than perceptual awareness.

Another interesting perception management system is that of the praying mantis. The mantis has in the order of 100,000 neurons,[7] roughly half of which are grouped in two large clumps, one behind each eye. The mantis has excellent visual abilities, and can utilize these as well as auditory processing in navigating a powerdive to avoid bats which feed on them. The mantis can launch an attack of its own as well, for instance cannibalistically on its own species. Thus if mantises are not conscious — and I am not taking sides on this — then a high degree of sensori-motor facility need not endow consciousness even in biological systems.

Such systems nevertheless can have a high degree of self-modelling; e.g. the mantis does not mistakenly attack itself instead of another mantis. But when self-modelling is present in *sufficient* degree, it may confer, or may simply be, what consciousness is. Just what that degree might be is considered below.

Self: A Hypothesis

I will set out a (possible) function of consciousness that I think might in fact *constitute* consciousness. I state this baldly as a hypothesis; later we will have to refine it a bit: *Consciousness is the function or process that allows a system to distinguish itself from the rest of the world*, conferring a point of view on the system, hence providing Perry's essential indexical 'I' (Perry, 1979); this plays an important role in error-correction, and bears on the problem of intentionality. Consciousness is then, first and foremost, a special kind of self-reference.[8]

Moreover, I think that this particular function is one that may be able to bear the weight of qualitative demands; at least the notion of self seems like a hopeful start: to feel pain or have a vivid experience requires a self. There is no such thing as pain *simpliciter*, or experience *simpliciter*, in the absence of an agent that is (has) a self, an 'I' to be the feeler (as in 'I am feeling pain').[9] Thus I think that recognition of self (personal identity) is an essential ingredient in conscious experience; I think it may even be *what it is* to be consciously experiencing. Note that recognition of self can go on in many particular contexts, some of which would be pain experiences, some colour-perceptions, some ruminatory excursions, and so on.

It is fair to ask, however, what good self-modelling is. This brings us to the general issue of error-recognition and repair, which means we must talk about meaning and reference.

Intentionality and self

No one mistakes a symbol for what it stands for; we easily distinguish the two.[10] The symbol is something we use in our thinking, hence instances of it occur in us, in our

[7] Compare to the 100 *billion* or so in the human brain.

[8] Refinements to come below include the idea that non-self can be one's own past remembered self, so that no *external* perception is needed.

[9] Try to imagine a system noting 'there is pain' but not that it is *its* pain. If not its, then whose? No-one's pain? Or consider visual experience: a scene appears as seen from a direction and at a particular distance, namely from the position and at the distance of the observing agent.

[10] Voodoo dolls and cave-paintings notwithstanding. The belief in a deep causal connection between two objects, or even that they are two aspects of the same thing, is entirely consistent with — and even built on — the ability to distinguish the one from the other.

belief base, in our self model. By contrast the *symbolled* is in the world, and merely represented by the internal symbol in our self-model. We have direct control over the one (the internal symbol) and not the other (the symbolled world). Thus we can alter our images or ideas or words: we alter the expression 'this is a dog' to 'this is a wolf' at will (whether for whim or speculation or to correct a false belief), but we do not so easily change a dog into a wolf. This symbol–symbolled distinction suggests several things, which I will detail in what follows. But I will note first that this rather obvious distinction is not currently put to much use in artificial intelligence systems, nor in psychology, linguistics, and neuroscience; it has been largely ignored, except in developmental psychology, where it surfaces in the appearance–reality distinction. I suggest that it may in fact play a very key role in intelligence and consciousness. Its proper handling requires the self-*vs*-world models as stated above, and can be seen in computational terms in part as a kind of quotation mechanism, i.e., 'Ralph' is a word in my thoughts and stands for Ralph in the world. Here we see the beginning outlines of our computational theory of consciousness.

When an agent's reasoning behaviour is reflected into its self-model, then it has become recorded as part of its narrative self-history, a term suggestive of Dennett's interno-phenomenological report.[11] I suggest that this is a key component of that behaviour's being conscious: it takes its place in episodic memory, as something that occurred in or to the agent. Without this double-layer of representation (as being outside the agent and also symbolled inside the agent), there is no 'I' and no awareness.[12]

Thus for a brain structure to provide consciousness, it must be complex enough to be able to provide a self-in-the-world, a symbol-to-symbolled tie that links a self model to a world model and can adjust the latter if errors are encountered. Various neural maps come to mind here, that may be part of a larger system of self–world representations: tectal maps, efference copies, thalamic maps, sensori-motor homunculi.

The above ideas with respect to language are further developed in several papers (Perlis, 1987; 1990; 1991; 1994; 1995). Newton (1988; 1992) develops a similar line of argument. We look more closely at this now, since it further illustrates some of the computational/quotational thesis.

Double representation and error

Even though double, the distinction between symbol and symbolled is useful, perhaps crucial, for it allows us tremendous flexibility to reconsider our beliefs, to see our beliefs as mere beliefs rather than brute truths: it allows us the wisdom that we are after all holders of imperfect views of reality, and the further wisdom that we can try to improve our views by finding our errors and correcting them. It allows what at one moment is a pure symbol undistinguished from what it stands for, to become at a later moment quoted or otherwise seen as an object of thought, something inside and not the outer reality.[13]

[11] A special self-reporting case of his hetero-phenomenological report (Dennett, 1991).

[12] The self/non-self or inside/outside distinction will be refined below, however, bringing it in line with the idea of the ur-quale.

[13] For a similar view see Humphrey (1992).

To relate this to a familiar subjective sense: We find ourselves engaged in a nearly constant back-and-forth between naive belief and circumspect self-querying, as we go through the day thinking about things. We are aware of thinking, aware of time passing, of ourselves with goals and being part-way through an ever-evolving effort. This can be the profound wisdom of a philosopher; or the profane wisdom of a raccoon rubbing water out of its eyes, not long mistaking its still-watery view with the dry world it has struggled to from the lake.[14]

We are constantly bombarded by such clashes in our perceptions, and we iron them out by noting, first of all, that we are possessed of views and that not all of them are correct (if they are in mutual conflict). This I think is a very basic phenomenon, not requiring explicit human-style language, but more like a very primitive (perhaps mostly bodily-and-visual) language of thought.

I suggest that an agent G cannot be conscious of event Y unless G represents an intentionality relation between G and Y: G must record the *fact* of its representing Y by means of a symbol (or image) 'Y' that is inside G. G not only represents Y with 'Y', G also represents the relationship between Y, 'Y' and G itself, along with means to adjust it. Thus G's situatedness in the world that includes Y is central to this notion of consciousness. There can be no box of pure unsituated consciousness, no box of 'perceiving redness', without an observer that is itself part of what is observed.[15] Again then we come to the idea of self as central to consciousness, and self-referral as ur-consciousness: Y and 'Y' are absentable, but not G's self-representation.

In Perlis (1994) I offer suggestions as to how an account based on self might be given for bodily reference and beyond, based on internal geometry and bodily situatedness and recalibration during motion. This yet again fits into my claim above that self is crucial: meaning is measured by reference to the agent's own body, e.g., via homuncular and other cortical and tectal maps, and involving that body's situatedness in the environment: this pain is in my leg; that red ball is in front of me. When we are conscious of Y, we are also conscious of Y in relation to ourselves: it is here, or there, or seen from a certain angle, or thought about this way and then that. Indeed, without a self model, it is not clear to me intuitively what it means to see or feel something: it seems to me that a point of view is needed, a place from which the scene is viewed or felt, defining the place occupied by the viewer. Thus I question (e.g. Crick, 1994, p. 21) that self-consciousness is a special case of consciousness: I suspect it is the most basic form of all.

Appearance–reality distinction and self

Error-recognition has ties to nonmonotonic reasoning (Perlis, 1990; Alchourron *et al*., 1985) in which reasoners may change their minds based on finding conflicts in their beliefs. I think that this too can be seen as an appearance– (or belief–) reality distinction (ARD, see Flavell *et al.*, 1986 and Gopnik, 1993). The ARD provides an interesting handle for studying much of what passes as 'mind' and it is amenable to

[14] The reader wary of my presumptive claims about raccoons, may simply substitute humans.

[15] Thus quotation or some similar device for internal referring may be a key ingredient in the processes by which an entity may be a self, i.e. a self-distinguishing self-presence. More will be said on this below. Note the double-representation implicit in representing an intentionality relation: this is precisely a matter of representing a representation. But it need not require a third level, let alone an infinite regress; we return to this below as well.

technical study (in psychology, AI, linguistics, and — hopefully — neuroscience). (So far it has mainly been studied only in developmental psychology; but see Miller, 1993.) The ARD is the capacity to distinguish conceptually between how something appears and how it is. This usually is applied to perceptual judgments (that ball looks blue in this light but it is really white); however, the concept makes sense in far broader settings.

Consider the example of having the belief that John is old (you see that he has grey hair). Later you discover that he is 25 years old and prematurely grey. Then 'John is old' comes to be seen by you as a belief or appearance, out of line with reality. As a result your beliefs change as they form a new current view of reality. So, there is a loop of belief-to-reality updates. The ARD then is in effect simply the self-*vs*-world modelling discussed earlier.

Note that ARD can involve temporal information: *that* is how it appeared to me (how I thought it was a moment ago) but *this* is how it is. Such reasoned change in belief involves recognition of passage of time and with it a passage of belief-state. Note also that the ARD applies equally both to perceptual judgments and to perceptual experiences. One can judge a past judgment to be in error, and so may one judge a past perceptual experience to be in error. Just as one's judgment or belief that a blue object is directly ahead may later be rejected, so may the experiencing of blueness be rejected as an error: did I really experience that, or is my memory fooling me?

Gopnik (1993) discusses an interesting study of 3-year-olds that bears on our claims. When questioned as to what they think is inside a closed candybox, they state it has candy; when shown that inside are pencils and asked again, they state it has pencils; and when then asked what they had thought it contained before it was opened, they state (falsely) 'pencils'. On the other hand, 4-year-olds do not make such mistakes. There are many subtleties to the design and interpretation of this and related studies. However, on the face of it, my theory might be taken to suggest that 3-year-olds are not conscious of seeing the pencils; or do not consciously see pencils: or perhaps are not conscious of meaning pencils by 'pencils'; or of having seen anything ten seconds ago as opposed to now. That is, they do not seem to distinguish the (former) appearance (a box with candy) from the reality (a box with pencils). The simplest explanation, perhaps, is that they do not remember what they had thought at first; this of course does not entail that 3-year-olds are not conscious. Thus the theory of consciousness I am proposing is not contradicted by ARD data. It is noteworthy that the inside of the closed box is not available in appearance, yet it is believed to be there by the three-year-old. It is unavailable 'perceptions' that seem to present the difficulty. To what extent then must cognitive self-modelling occur, to count as conscious? I have been urging at least some form of this, but it need not extend to time periods long enough to be captured in language.

Consider an individual unable to distinguish a seen object from how it looks. Such a person may be puzzled, for instance, at things becoming blurred in rainy weather, (compare the raccoon example above), or in their disappearing as night falls. This would, to say the least, be a very severe disorder of thought. If I am right, it would amount to the loss of thought altogether — at least if it extended to all modes of representation rather than visual alone — leaving only a mindless and slavish recording of inputs with possible reactive responses (no weighing of alternatives).

According to the theory being advanced here, such a person would not be conscious at all.[16]

We have been discussing self-*vs*-world modelling at some length, but now we must ask what constitutes a self, and how it can be distinguished from non-self. This will add a further dimension to the quotation–computation mechanism.

Strong Self-reference

If it is like something to be conscious, then that something, that experiential feel, is not imaginably absentable, i.e. it is not like anything at all to be without that feel. How then can it be noted, be a part of awareness? How can we note something without thereby noting a difference from an absence of that something?

Ordinarily we may distinguish experiences by differences, but perhaps this is not essential. Perhaps certain notings can be done in such a way that they can only occur positively, never as an absence. In particular, an inherently self-noting process may be exactly that: not imaginably absentable. Whether such occurs in the conscious brain, and whether we can discover such computational processes, is an empirical matter.

Why would such a not-imaginably-absentable feature be important? What is its functional role? Here we come to the crux of the debate, and the crux of this paper. The forms of self-reference most widely cited and studied, from antiquity to the present are weak forms. They tend to come in two types: delegated self-reference[17] and meta-self-reference. Delegated self-reference has been made famous in the sentence 'This sentence is false' as well as others such as 'This sentence has five words' and 'This sentence no verb', not to mention 'This proof-system is consistent'. On their own, such sentences express nothing; it takes a linguistic community to interpret them and close the loop, so that 'this' comes to mean that very expression itself.

Meta-self-reference is another kind of weak self-referring, most easily described with the help of a robot, Ralph. Suppose in Ralph's knowledge base (KB) are various sentences, including 'Sue is Canadian' and 'Ralph is American'. The latter does not in itself amount to Ralph's referring to himself, i.e. it does not form a closed loop back to Ralph (without delegated help from us), unless Ralph also has further sentences or processes that do just that: link the name 'Ralph' to Ralph. Replacing 'Ralph' with 'I' will not in itself achieve this; a special treatment of 'I' is needed (Perry, 1979). Links that tie 'I' to Ralph's own body are a beginning, permitting Ralph to order replacement parts for his broken arm. But he could do the same for robot Sue's broken arm, from knowing Sue was built at a certain Canadian factory. The fact that in the former case Ralph is replacing his own arm, as opposed simply

[16] This hinges crucially on the phrase *all modes of representation*, including self-representation. Such a person would not have a self in the sense argued here. We discuss this at greater length below.

[17] Self-reference has an illustrious role in intellectual history, from antiquity (the Liar paradox) to modern times (Cantor's, Gödel's, and Turing's theorems). However, the form of self-reference in these cases is a delegated one: the actual action of referring is done by an interpreter outside the supposedly self-referential objects (sentences). Moreover, such delegated-self-reference is, when treated with technical care, quite well-understood, not quite tail-chasing after all despite how it may seem to beginning logic students. This alas is not enough for our purposes; no one proposes that, for example, a formal system of arithmetic prone to Gödelian incompleteness is in any sense conscious, or is even an active entity that can perform or partake in processes.

to the arm of a robot named ‘Ralph’, is irrelevant. We can keep adding to Ralph’s KB: ‘I am Ralph’, ‘ ‘‘I’’ means the robot with serial number xyz’, etc, in a hierarchy of referrings, but none seeming to get to a final self-contained self. What is of interest for us is not such meta or delegated self-reference, but rather entities that self-refer all on their own.

Why do we need a self-contained self, where referring stops? Negotiating one’s way in a complex world is a tough business, for a robot or for a biological system, and complex behaviours have come about as a result. Dealing with the inevitable errors that crop up is one big part of the problem, necessitating commonsense reasoning, as above in the case of Ralph noting the need to order a new arm. But now something interesting happens. Suppose the new arm is needed within 24 hours. He cannot allow his decision-making about the best and quickest way to order the arm get in his way, i.e. he must not allow it to run on and on. He can use meta-reasoning to watch his reasoning so it does not use too much time, but then what is to watch the meta-reasoning? Since he is a finite system, his resources are limited and he cannot do all kinds of reasoning simultaneously. He must budget his time. Yet the budgeting is another time-drain, so he must pay attention to that too, and so on in an infinite regress. Treating his planning as one thing and his time-tracking of his planning as another, and so on, by separate modules responsible for each level of reasoning, clearly will not work. Somehow he must regard it all as himself, one (complex) system reasoning about itself, *including that very observation*. He must *strongly self-refer*: he must refer to that very referring so that it’s own time-passage can be taken into account.

Do we ever find ourselves having such a ‘conscious’ state? I think we do so all the time, it is the essence of barebones consciousness: ‘here I am’. Not ‘here is Joe’ and ‘Joe is me’ and ‘me is the person who just thought his name is Joe’ and so on. We catch ourselves in the present, in a strongly self-referring (SSR) loop. It is the recognition that ‘this is now’, where ‘this’ is my present experience that this is now. Circular, yes, but not quite paradoxical.

Now we can look back and say that even a sentence such as ‘I am Ralph’ can strongly self-refer in the appropriate system in which the pronoun ‘I’ is treated in a special way. So, what is *strong self-reference*, what is that special way? I do not have a technically precise answer, but I do claim that this problem is a technical one, not a philosophical one. Robots, like humans and many other biological entities, need this ability, and it is one that is functionally defined. Moreover, it is not so apparent that it does not have, in and of itself, an attendant quality, a something-it-is-like-to-be. It might be a nearly qualeless-consciousness, but with a bare ‘I am here’ aspect to it that is distinguishable even though it is never noticeably absent: the ur-quale.

In light of the above, let us now again ask, what is a self?

I suggest that a self is best thought of as an entity G that can refer to G as that entity doing that very referring. This might for instance be associated with the gloss ‘here I am now thinking about myself’. There is a peculiar kind of tail-chasing mind-bogglingness to such a description. It is this that I suggest is at the heart of self and therefore of consciousness.

I will now advance a tentative semi-technical definition of strong-self-reference. An entity G strongly-self-refers by an action A if:

1. G models the performance of A;
2. that same modelling is part of that very performance of A;
3. this reflexive aspect of the modelling is itself part of the modelling.[18]

These three 'axioms' are admittedly not as clear as one would like. I present them as very rough guides to further study. However, one thing seems clear: time and memory must play very special roles in this, for it is a self-modelling *process* we are dealing with, not a frozen formula. Perhaps these models run on a very fast basic time cycle, perhaps a few milliseconds long, in which there is a blurred notion of the present, i.e. in which there can be several things occuring that manage to refer to one another.

I will now offer several observations that appear to be in line with the idea of a self-referring final-self.

First, our earlier comments that qualia must be *someone's* qualia, that to be in pain one must take the pain to be one's own, does not sit well unless there is a final self. A hierarchy of selves, each referring to the one below, does not self-refer, and so does not take anything to be its own. To say 'the system' as a whole feels the pain, or is aware, simply backs away from the problem. Maybe a system as a whole can be aware, but we need an account of how that might be.

Second, the only kind of reference that does not pass the buck is reference to itself, i.e. a referring that refers to that very referring. This sounds very odd, but we have seen examples: 'this current action of expressing' or 'I am now referring to my referring'. While strained (unfelicitous) these are still intelligible. I am proposing that something akin to this goes on literally all the time when we are conscious, and that this *is* our consciousness. It is of course not usually spoken, and probably is on a much faster time-frame, and would not normally be under our control. On the contrary, it is the very matrix of awareness that gives us control over slower behaviours.

Third, it seems to me that we *explicitly* do something very much like this at times when we think about ourselves. For example, in making the utterance 'I am now speaking English', we refer to our very referring. This I think satisfies the three axioms above; in particular, the reflexive character ('I . . . now') is what makes it be us and not Joe Blow that we refer to. Note that this example exploits a time cycle well beyond a few milliseconds; but it is not a blur, since we easily pick out earlier and later parts of it. But there may be an elementary 'I'-cognition that has no observed subparts: it is observation at its most primitive.

Fourth, we need to do this, at least in deadline-coupled situations. Here is a more difficult example, but one that makes the point.[19] I decide 'I'll get on with things', implicitly meaning not only to stop whatever I had been doing, but also to pass beyond that very decision and on into some other action. Here the decision seems in part to refer to that very process of decision, i.e. to get on past even it. There may appear to be an infinite regress of meta-levels, but I believe this is incorrect and that we do in fact refer to our own very act of referring. Otherwise there is nothing in the represented pattern of thought that ties it back to the thinker's ongoing actions. That is, we might actually either (i) get into an infinite regress and never come to a full stop to get on with other tasks, or (ii) we might simply stay at a particular meta-level

[18] Here 'modelling' is an ambiguous term perhaps nearly synonymous with 'referring' or 'representing'. Presumably the utility of modelling is that it allows the individual the ability to draw inferences and make plans, especially in deadline-coupled situations.

[19] A bit along the lines of the earlier one of Ralph seeking a new robot arm.

and never note that it too is keeping 'us' from those tasks. Somehow we must (and do) tie our ongoing sense of time passing to our ongoing planning and acting, and get the right things done at the right time (sometimes!).

A question then is how can an active system G genuinely self-refer? Does it take something more than information processing? And does it confer consciousness? On the latter, since we do not currently have an independent definition, we are left with intuition. I think that only by careful examination of human behaviour and the design of smarter robots will we be able to position ourselves to have more than merely prejudicial intuitions. At present I simply offer this as a tentative characterization of consciousness, namely: a process of self-referring that satisfies the three axioms above.

No one, to my knowledge, has built, or even tried to build, strongly-self-referring machines. This in large part is due simply to the fact that no one has tried to build robots that can do very much reasoning, or even that can do very much common-sensical self-protection in a complex world. But strong self-reference is what an intelligent robot needs, to avoid the infinite meta-regress, as well as to appropriately take action to protect itself, say, when it infers that 'it' is in danger.

One may retort that although *at times* one is aware mostly of oneself and no more, this is more often not the case. One may think about the Moon, and not oneself. But this is a misunderstanding of my point. The strongly-self-referring ur-quale (which we might give as the gloss 'here I am') is always there, whether or not the 'here' includes the Moon or anything else as part of it. There can be many types of contents to consciousness; the ur-quale is always among them even if it is not in central focus, and indeed it might never be so.[20] It is perhaps better put 'here I am as this noting of things including this noting'[21] or more simply 'this is itself a noting of XYZ going on'. One's activity keeps bordering on focusing on itself and then (necessarily) getting pushed aside by its own activity; and yet this very fact is somehow recorded or observed as part of that activity. Very puzzling stuff, but we should not assume a physical device cannot do just this.

Discussion

The above presents a number of complicated notions that require further comment to appreciate their interconnections. I specially wish to consider some particular areas of possible misunderstanding of my intent. In this I avail myself of some very helpful comments and questions by the editors, Jonathan Shear and Shaun Gallagher.

I have argued that consciousness involves a self/non-self distinction, and then I assert that the ur-quale, the essential ingredient of consciousness, is devoid of the usual cognitive modalities (vision, touch, and even thought with external content) by which we know non-self. This seeming inconsistency touches on a key refinement of my initial definition: the self is also non-self when it is remembered as one's past self: it no longer is the self of the moment — subject becomes object — and this 'sliding along' in self-observation is another way to describe strong self-reference. This need

[20] This is one way of reading of Searle's claim that 'I cannot *observe* my own subjectivity' (1992, p. 99).

[21] Grice's views on speech acts (Grice, 1975) are similar to this, as well as more recent work on mutual knowledge (Barwise, 1989). Both involve self-reference not unlike the strong sort here.

not be a rich memory of years gone by; it is enough that it be a memory of one's immediate past activity, even if that activity is simply internal self-observation, an ongoing 'here I am'.

What is the computational mechanism I have promised, as a possible basis for the ur-quale? It is strongly self-referring computation, probably facilitated by some sort of quotational syntax. However, it is a computational research paradigm, not a precisely defined notion at present. I have presented examples intended to show the need of such a thing in intelligent behaviour, especially deadline-coupled planning in complex environments. This in turn suggests two places to look in evolutionary terms for the appearance of consciousness: (i) where behaviour of that sort does or does not arise[22] and (ii) where there are brains with suitable processing power to allow such strong self-reference.

Note, however, that very likely in evolution, the processing of external perceptual data became important early on, and so the first appearance of the ur-quale may well have coincided with the arrival of 'fancier' qualia. That is, the devices for processing perceptual data likely were well in place long before the ur-quale appeared and made possible the 'translation' of that perceptual processing into fancy qualia. It seems unlikely that evolution would have wasted the energy to build self-meditating worms that could not utilize that ability to better survive. But when deadline-coupled planning becomes essential to survival, when planning and acting need to be subtly dove-tailed and what has just been done needs to be factored into what is to be done next, yet without letting that deliberation take too long, we may be nearing a strong self-loop of now-into-then processing tantamount to the ur-quale.

Thus the ur-quale probably evolved in conjunction with very complex external perceptual processing. Still, it need not be tied to the latter once it is present. This is not to say, however, that the ur-quale is something simple. It will be a complex process, one requiring memory (of itself) and temporal processing. Quite possibly the ability to access the ur-quale in isolation, as in a meditative state, is an accidental by-product of evolution; at least I see no survival value in being able to strip away fancy qualia altogether, despite possible philosophical, psychological and aesthetic value.

With the refinement above, we can now reinterpret my argument for a double-layer of representation (both outside the agent and also symbolled inside) as being outside the present activity of the agent and yet also symbolled in that present activity, namely as one refers to one's immediate past. Thus my theory does not require, for consciousness, sophisticated views of external reality found in, say, adult humans but not in three-year-olds. It is enough that a now–then distinction be made, even on a short time-cycle, enough for self-representation as an ongoing-ness of the self from present into future.

Moving from appearance to reality with regard to conscious experience makes sense precisely in terms of time-passage. We access the process of a moment ago, taking it as the present (i.e. as reality) until we reject it and take the next moment as the present, and so on. We are caught permanently in a now-to-then loop. How such

[22] The Sphex wasp, for instance, seems not to be able to distinguish very well what it has done from what it must still do: if its multi-step routine of stocking its nest with supplies is even slightly disrupted, it begins all over again, repeating many unnecessary steps. This suggests that it has little if any internal model of its own behaviour.

a loop can also refer to itself, i.e. to its own self-referring processing, is an open question. But we have seen reasons to believe something of that sort may well be requisite for survival in complex entities such as ourselves.

Comparisons

Deikman (1996) argues a related position: content is not enough, there must be a self (which he calls the 'I', reserving 'self' for more incidental aspects of the aware agent, such as personality). This deep inner 'pure' self is bare awareness in itself, as suggested by the answer 'yes' that one is likely to give to the question, 'Are you conscious?'

He further says we *are* awareness and do not need to observe awareness; being awareness is a different kind of knowing awareness, from the inside. But what does this mean, and why is it not also a kind of self-observing, perhaps different from but related to other-observing? He seems to suggest self-knowing occurs in a largely different sphere from that of space and time, but this is a large leap that may not be needed.

Contrary to Deikman, I suggest the observer can be and is observed by itself, and so can be content as well as observer. Awareness does always have an object, but that object can be pure awareness of itself.

The SSR theory being advanced here has some common ties with the higher-order theory, HOT, of Rosenthal (1986). However, the latter (actually more meta-theory than higher-order theory) is not genuinely self-referential, and thus cannot avail itself of the suggestive hints we have urged here, toward closing the explanatory gap on awareness and qualia. Rather, HOT postulates distinct levels or layers of representation directed from one to another. By contrast, SSR postulates a single reflexive level.

Rosenthal distinguishes creature-consciousness from higher-order consciousness; the former may come close to what I above called perceptual management, while the latter, a form of self-consciousness, is proposed as the consciousness of interest, Block's P-consciousness. However, it is not defined that way by Rosenthal; it is defined as a kind of meta-propositional information, about creature consciousness for instance. Thus the information that 'one is hungry' — itself distinct from the gastro-intestinal facts of the matter — is a higher-order piece of information a system may have about itself. This in turn may be further modelled at a yet higher level as 'I have the belief that I am hungry'.

Harth (1993) espouses a recursive notion of awareness as a process in which successive passes of processing provide a deepening of representational 'bias'. However, the self-reference described in his account does not appear to be that of representation of the process to itself; there is only content, no subject. What we need, according to the SSR theory, is a genuinely self-reflecting loop, one that takes its own activity into account, that sees itself as a self-seeingness.

There is a frequently-heard view (e.g. Crick, Block) that self-consciousness is a special and unusual form of consciousness. We suggest a distinction between strong-self-reference on the one hand, and introspective consciousness on the other. The former is always present in a conscious system, on my theory: it is consciousness. But the latter is an additional feature, in which the noting includes, say, historical information about oneself, such as 'I tend to be shy'. Here the 'I' reveals SSR at work, and the rest is introspection or meta-knowledge. As such the latter adds much to the

cognitive repertoire of the system, but little at all to our understanding of the nature of consciousness. It is not so much a special kind of consciousness as simply a special kind of information. Indeed, many introspective mechanical systems have been built, but none that are conscious.

To sum up: consciousness is the function of strongly-self-noting that allows a system cognitively to get out of its own way, to avoid an infinite regress.

So, does this 'stand up and grab us' as being obviously right, obviously a fount of an inner life of the mind? I cannot claim so. But I do think that it is at least not obviously wrong, that there is something to the idea of a bare, stripped consciousness that only knows its own knowingness, and that such would not be vividly populated with colours and smells and urges. And I think it also is at least not obviously wrong that it is like something to be in such a state.

Conclusions and Neural Connections

Much of the consciousness debate hinges on qualia — the felt qualities of experiential awareness: colours, pains, moods, what it feels like to do or be or undergo this or that. Yet one can be colourblind and assuredly conscious. While not denying the philosophical challenge that qualia present, we might still consider whether consciousness itself is something more basic than qualia. If we strip away colour experience, pain experience, emotional experience, and so on, is anything left? Is it like something simply to be conscious, and if so, what is it like? Intuition suggests it is like something, but perhaps a very primitive something. Might this not be simply strong self-reference? Note that complex time-situated and memory-bound processing must occur as part of strong-self-reference. It might not be like very *much* to be a pure ur-consciousness/self/strongly-self-referrer: no personal feelings, no goals, no cares. But it is not so evident that it is like nothing, surely not so evident as that it is like nothing at all to be a rock or a Macintosh computer.

Where are we to look in the brain for such amazing structures and processes? A camera can take a picture of itself (via a mirror, and can even take a picture of itself that includes the mirror); this is an elementary example of self-reference. But there may be far more subtle ones in the brain. Known neural loops are a start, from efference copy in VOR to the reentrant loops emphasized by Harth and Edelman. But that's only a beginning. We'll need far better models of strong self-reference, self-modelling temporal loops that take now into then on and on, while also being able to get out of their own way. Perhaps the diagonal method of Cantor, used so well by him and Gödel and Turing in explicating self-referential mysteries of mathematics and computation, has yet more in store for us in the brain.

This paper has sketched one possible 'scientific' (function or process) theory of consciousness. To be sure, I have not given a detailed account of exactly how subjectivity might arise in systems with the functional capacities I describe; but this I think is not to be expected in advance. It is far too early to give up on traditional 'function or process' modes of scientific inquiry regarding consciousness. My hope is that the amazing-structures paradigm will little by little lead to just that, in computational, cognitive, and neuroscientific terms.

References

Alchourron, C., Gardenfors, P. and Makinson, D. (1985), 'On the logic of theory change', *Journal of Symbolic Logic*, **50**, pp. 510–30.

Barwise, Jon (1989), *The Situation in Logic*, chapter 9: On the Model Theory of Common Knowledge. CSLI Lecture Notes: Number 17. Center for The Study of Language and Information.

Block, N. (1995), 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18** (2), pp. 227–87.

Chalmers, David J. (1995), 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2** (4), pp. 200–19.

Crick, F. (1994), *The Astonishing Hypothesis* (New York: Scribners).

Deikman, Arthur (1996), ' "I" = awareness', *Journal of Consciousness Studies*, **3** (4), pp. 350–6.

Dennett, D. (1991), *Consciousness Explained* (New York: Little, Brown).

Flavell, J., Green, F. and Flavell, E. (1986), 'Development of knowledge about the appearance–reality distinction', *Society for Research in Child Development Monographs*, **51**, No. 1, Series No. 212.

Gopnik, A. (1993), 'How we know our minds: the illusion of first-person intentionality', *Behavioral and Brain Sciences*, **16** (1), pp. 1–14.

Grice, H.P. (1975), 'Logic and conversation', in *Syntax and Semantics 3: Speech Acts* (New York: Academic Press).

Harth, Eric (1993), *The Creative Loop* (New York: Addison-Wesley).

Humphrey, N. (1992), *A History of the Mind* (New York: Simon and Schuster).

McGinn, Colin (1995), 'Consciousness and space', *Journal of Consciousness Studies*, **2** (4), pp. 220–30.

Miller, M. (1993), 'A view of one's past and other aspects of reasoned change in belief', PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland.

Nagel, T. (1974), 'What is it like to be a bat?', *Philosophical Review*, **83**, pp. 435–50.

Newton, Natika (1988), 'Machine understanding and the Chinese Room', *Philosophical Psychology*, **1**, pp.207–15.

Newton, Natika (1992), 'Dennett on intrinsic intentionality', *Analysis*, **52**, pp. 18–23.

O'Rorke, P. and Ortony, A. (1994), 'Explaining emotions', *Cognitive Science*, **18**, pp. 283–323.

Perlis, D. (1987), 'How can a program mean?', in *Proceedings, International Joint Conference on Artificial Intelligence*, Milan, Italy, 1987.

Perlis, D. (1990), 'Intentionality and defaults', *International J. of Expert Systems*, **3**, pp. 345–54 [Special issue on the Frame Problem, ed. K. Ford and P. Hayes].

Perlis, D. (1991), 'Putting one's foot in one's head — part I: Why', *Nous*, **25**, pp. 435–55 [Special issue on Artificial Intelligence and Cognitive Science].

Perlis, D. (1994), 'Putting one's foot in one's head — part II: How', in *From Thinking Machines to Virtual Persons: Essays on the intentionality of computers*, ed. Eric Dietrich (New York: Academic Press).

Perlis, D. (1995), 'Consciousness and complexity: the cognitive quest', *Annals of Mathematics and Artificial Intelligence*, **15**, pp. 309–21 [Special issue in honor of Jack Minker].

Perry, J. (1979), 'The problem of the essential indexical', *Nous*, **13**, pp. 3–21.

Rosenthal, David (1986), 'Two concepts of consciousness', *Philosophical Studies*, **49**, pp. 329–59.

Searle, John (1992), *The Rediscovery of the Mind* (Cambridge, MA: MIT Press).

Shear, Jonathan (1996), 'The hard problem: Closing the empirical gap', *Journal of Consciousness Studies*, **3** (1), pp. 54–68.

Paper received June 1997