CS-TR-3324 | July 19, 1994
UMIACS-TR-94-91

An Error-Theory of Consciousness

Don Perlis

Institute for Advanced Computer Studies
Computer Science Department
AV Williams Bldg
University of Maryland
College Park, MD 20742

perlis@cs.umd.edu

Abstract

I argue that 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 own errors. This in turn requires a robust self model as part of its world model, as well as the capability to come to see that world model as residing in its belief base (part of its self model), 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, would not be 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.

Note: this paper will appear as a chapter in {\em Toward a Scientific Basis for Consciousness}, edited by Stuart Hameroff et al, MIT Press.

Work done with partial support from the Army Research Office (DAAH0494G0238).

An Error-Theory of Consciousness

Don Perlis

Department of Computer Science

and

Institute for Advanced Computer Studies

University of Maryland

College Park, MD 20742

301-405-2685

perlis@cs.umd.edu

This research was supported in part by a grant from the U.S. Army Research Office (DAAH0494G0238).

Abstract

I argue that 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 own errors. This in turn requires a robust self model as part of its world model, as well as the capability to come to see that world model as residing in its belief base (part of its self model), 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, would not be 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.

1 Introduction

Consciousness serves the function of allowing 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, self-consciousness. This theme will be argued throughout what follows. I begin with what may seem like a very different issue, but which will in fact provide a key: the distinction between a symbol and its referent.

No one mistakes a symbol for what it stands for; we easily distinguish the two. The symbol is something we use in our thinking, hence instances of it occur in us, in our belief base, in our self model; whereas the *symboled* 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 symboled 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-symboled 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.

When an agent's reasoning behavior 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 (Dennett 1991). I suggest that this is a key component of that behavior'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 symboled inside the agent), there is no "I" and no awareness (see below for more on double representation).

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-symboled 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.

In several earlier papers (Perlis 1987, 1990, 1991, to appear) I present various aspects of this theory, but mainly focusing on the problem of intentionality (language and meaning). Here I emphasize instead the themes of mind and consciousness *per se*.

2 Double representation and error

As noted above, people can make a distinction between a symbol and what object-of-reference it stands for, hence our thinking must have distinct representations for things and their representations. Such a distinction is a double-representation, in effect, since when making it people do not have immediate access to the real external objects (we never do). We have a representation for the referent (that representation can be the symbol itself) and another for the symbol when we want to talk about the symbol. Quotation is one device we use, out of many: "Joe" is the name of Joe; "dog" refers to dogs. Another term some may prefer is *reflection*: I can reflect on (some of) my referring inner states *as* inner states instead of what they refer to.

Yet another way to view this is in terms of imagination: when we take something to be in our heads and (not necessarily) out there, we are imagining it. Thus I can imagine a

"Joe" by using some of the internal tags I would use for a real Joe; some but not all, since in imagination I suspend belief, which is to say (so I argue) that I quote (or reflect upon or mention) my internal usages instead of simply using them.

Even though double, the distinction 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.

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 in partway 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. 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 visual) language of thought.

This same line of argument suggests that importance of dividing the world into external reality and internal view, a kind of other-vs-self distinction. Thus I think that a self-notion or self-model is probably of major importance to the study of mind, possibly even that without a self-representation there would be no mind/consciousness. According to this view, 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.

This also bears on Searle's Chinese room (Searle 1980). To understand a word is to tie it to a part of one's world model. Searle's scenario does not consider such double representations. His discussion leads one to imagine that the book of instructions for manipulating Chinese symbols does not involve data structures that represent the room itself or the fact that it is using symbols for external entities. On the error-theory such a limited scenario would in fact not be conscious, nor would it have true intentionality. But this would not be a condemnation of computationalism; it would simply illustrate that the Turing test is not enough to guarantee mental content, that *some* computational models do not have such, and leaving open whether some others might.

How is it that symbols can represent entities at all, especially distant ones beyond the symbol-manipulating machinery (Searle's room), is a distinct question that Searle conflates with those of understanding and consciousness. I address the former below.

3 Intentionality

This leaves untouched the issue of how symbol (belief) and symboled (reality) become linked, beyond "simple" cases such as staring wide-awake at a close bright red dot in a well-lighted room: here the dot-symbol in the brain (in whatever distributed neuronal form it may take) symbols the actual dot. But how about visualizing a dot far off and out of sight? That is much harder to characterize since we do not have ready access to what is symboled, hence no obvious clue that those neuronal processes "mean" a far off dot.

In (Perlis, 1991) I take an even simpler case than the close red dot as a basic one for my analysis: one's own foot. Our foot we can see, and also think about when the visual link is broken (eyes closed, looking up at the sky, etc) But we cannot (without severe damage) break the neuronal links between foot and sensorimotor homunculi, or between foot and tectum. So I propose to make such built-in wirings a key part of a theory of mind (or of symbol-symboled relations), with the concomitant self-other (internal-external) distinctions above: we can imagine our foot to be amputated, or in fact discover that it has been when visual and sensorimotor signals conflict. We quickly realize that the reality is not necessarily the same as the belief, and we struggle to bring the two into accord. Of course, if the neural link is broken, we usually find conflicts between our senses: our eyes tell us the foot is there, but our proprioception tells us otherwise, so we need to employ circumspection (suspending belief, seeing beliefs as possibilities, just in our heads) for a time in order to undertake to resolve the matter (e.g., by trying to wiggle our toes, or whatever). In (Perlis, to appear) I offer suggestions as to how such an account might be extended beyond bodily reference, based on internal geometry and bodily situatedness and recalibration during motion.

This 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 X, we are also conscious of X 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. Without something along these lines, I think that a "neuronal box" would indeed "confuse" symbol and symboled: to it there is no external reality, it has no way to "think" (consider alternatives) at all. Thus I disagree (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.

4 Appearance-reality distinction

I think that reasoning plays an interesting role here, especially in the recent studies of non-monotonic reasoning, 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). To sum up all the above, the ARD is I think an interesting a candidate for much of what passes as "mind" and it is amenable to technical study (in psychology, AI, linguistics, and hopefully neuroscience). (So far it has primarily been studied only in developmental psychology.) The ARD is the capacity to distinguish between how something appears and how it is. This usually is applied to perceptual judgements (that ball looks blue in this light but it is really white); however, the concept makes sense in far broader settings, such as judging that one's belief that John is old is mistaken (and should be revised). In computational terms, such an ability involves distinguishing the belief that John is old from the reality, hence the belief is not (or at least no longer) seen as being the reality: instead the belief is seen as something inside the believer, made up of objects such as the word "John".

An individual with a damaged ARD capacity, would presumably have loss of the ability to distinguish words from their meanings, thus no ability to comprehend that someone has lied, for instance, or that by moving her head she can see something better. I am not aware of such a clinical diagnosis.

There may be a related disorder in visual awareness: someone who cannot distinguish a seen object from how it looks. Such a person may be puzzled at things becoming blurred in rainy weather, for instance (compare to 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, leaving only a mindless and slavish recording of inputs with possible reactive responses (no weighing of alternatives). According to the error-theory, such a person would not be conscious (not have a mind).

5 Conclusions and neural connections

Laying down (or recalling) an episodic store (of event E) may then be the same thing as being conscious of E. At least such would appear to have some components requisite to the error-theory: self model, world model, and appropriately flexible connections between the two, and seems closely linked to the narrative self-history idea as well. Computational studies in progress indicate great advantage may accrue to an agent with such capacities.

Such a conception of consciousness is a bit different from short-term memory or long-term memory or even their conjunction. It involves these two plus a self model, a model of one's process of laying down an episodic store; not in physiological detail of course, but a high level model of the self as an entity that is undergoing an event and recording it as an experience. Such a conception provides room for Dennett's Orwellian and Stalinesque scenarios (as distinct from one another) and also for reasoned change of mind if it is found that the store is inaccurate. Presumably various forms of reasoning can be applied to the store in the process of such recalibrations—category formation and adjustment, nonmonotonic reasoning, and variable instantiation among them.

I am currently working toward a reformulation of the above ideas that might lend themselves to experimental insights, especially using recent imaging techniques. The hope would be that sufficiently fine-grained imaging might be able to isolate brain areas that perform appearance-reality checks, a kind of neural quotation (or imagination) device. The well-known neural maps such as efference copy come to mind as perhaps primitive versions of such a mechanism.

References

[1] F. Crick. *The astonishing hypothesis.* Scribners, 1994.

[2] D. Dennett. *Consciousness explained.* Little, Brown, 1991.

[3] J. Flavell, F. Green, and E. Flavell. Development of knowledge about the appearance-reality distinction. *Society for Research in Child Development Monographs*, 51, 1986. No. 1, Series No. 212.

[4] D. Perlis. Putting one's foot in one's head—part II: How. In E. Dietrich, editor, *From Thinking Machines to Virtual Persons: Essays on the Intentionality of Computers.* Academic Press. To appear.

[5] D. Perlis. How can a program mean? In *Proceedings of the 10th Int'l Joint Conference on Artificial Intelligence*, pages 163–166, 1987.

[6] D. Perlis. Intentionality and defaults. *International J. of Expert Systems*, 3:345–354, 1990. special issue on the Frame Problem, K. Ford and P. Hayes (eds.).

[7] D. Perlis. Putting one's foot in one's head—part I: Why. *Noûs*, 25:435–455, 1991. Special issue on Artificial Intelligence and Cognitive Science.

[8] J. Perry. The problem of the essential indexical. *Nous*, 13:3–21, 1979.

[9] J. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.