Operationalizing Consciousness

Don Perlis[1] and Justin Brody[2]

[1] University of Maryland, College Park MD 20742
perlis@cs.umd.edu
[2] Goucher College, Towson MD 21204
jdbrody@gmail.com

Abstract. David Chalmers (among others) is fond of saying that consciousness has no function; it can be there or not – it makes no difference to behavior. In that sense, it supposedly is not like a pumping heart that helps keep one alive. Here we argue to the contrary: that consciousness has a critical function, and one that AI will be forced to deal with as a practical matter, as we probe more deeply into realtime commonsense reasoning. We will draw on a broad range of work – philosophical and otherwise – in making our argument.

Keywords: Consciousness · Intentionality · Self

1 AI and Consciousness

The topic of consciousness tends to lead to two kinds of claims: positive claims about what it is, and negative claims about what it isn't. The latter include Chalmers' claim that consciousness has no function, no physical consequences – it is an epiphenomenon [8]; and Searle's claim that subjective experience (in the form of intentionality) cannot be achieved simply in virtue of a system's executing a (formal) program [26]. Part of our purpose here is to examine – and disagree with – both of these negative claims.

Positive claims attempt to characterize the nature of consciousness. These include Brentano's notion of intentionality [2] – that the mental is characterized by its directedness toward objects of thought: to be conscious (i.e., in a mental state) is to have thoughts (or feelings or attitudes) about something. Another positive claim is due to Nagel: a conscious being is an entity that it is like something to be [15]. This latter notion essentially characterizes consciousness as having a qualitative subjective experience, something happening to and in oneself.

So Nagel and Searle address similar notions of consciousness but with different aims: Nagel says what it is; Searle says that it cannot occur via formal computational processes alone (and in part bases his argument on a Nagel-like experiential character). And Brentano provides a functional role for mind (the relation of aboutness between a thought and its meaning), whereas Chalmers denies any such functional role.

We too seek to characterize consciousness positively, in terms of particular processes. Nagel's characterization is less useful here than Brentano's. But the two can be regarded as taking similar positions: being conscious amounts to having (internal, qualitative, subjective) thoughts and feelings, and thoughts and feelings are necessarily about something. Further, it is a common view in Buddhism that consciousness is “that which is aware of objects” – seemingly combining the two (Nagel’s awareness and Brentano’s aboutness) [28]. So it is tempting to consider whether a suitable subjective form of aboutness is an essential ingredient of consciousness. The aboutness relation – as argued in [19], [20], [21], [22], [17], [16], [18] – not only connects symbols “in the head” to (usually) external meanings but also is a key role of the self: the self is what “intends” to refer to Joe Smith in employing the symbol “Joe”. This will take us on a somewhat meandering tour of various issues, from zombies to language to robots to knowledge. Far from being a strict epiphenomenon, consciousness seems to tie into a wide range of behaviors.

1.1 Zombies and Reflexivity

A philosophical zombie (or just “zombie” if there is no confusion) is a molecule-for-molecule identical copy of a normal human, subject to exactly the same physical laws and thus producing indistinguishable physical behaviors; but (by definition) a zombie has no subjective experience, it is not like anything to be a zombie [8]. The question then is: *are zombies impossible?* This is equivalent to the question: *is consciousness a physical process (i.e., something that performs a physical function)?* Chalmers has argued that zombies are possible ; we shall not repeat his complex argument here (it is essentially based on the idea that we seem to be able to imagine zombies), but rather present a counter-argument: Suppose you have a zombie twin and each of you suddenly says “wow, I’ve got a painful toothache and its getting worse.” In your case, this is because you in fact feel that toothache; but the zombie cannot feel pain (or anything else). Yet identical brain processes are occurring in both brains (by definition of zombie). That is, whatever physical process led to your utterance also led to the zombie’s. Thus your utterance cannot have been based on (caused by) your feeling the pain after all. This is a contradiction, so the possibility of such a zombie is ruled out. The hidden premise here is that when we make a decision to honestly report on a pain (or other subjective experience) it is in part dependent on there really being such an experience. To deny our argument is to reject this highly intuitive premise.

But maybe subjectivity is an after-the-fact event: for instance, one makes an utterance and then comes to feel whatever the utterance was about. In that case, the zombie-twin might simply lack whatever (non-physical) competence is involved in coming to have a feeling. This of course still flies in the face of our intuitive premise and so does not seem much of an argument. But it brings us to an important distinction, between reflexive and reflective notions of self. Looking back on an event and then forming a conclusion about it, is a process of reflection. Thus, I can reflect on the toothache I had yesterday. But my toothache today is even worse. I am not reflecting on this but rather am reflexively knowing the immediate pain itself. This is very easy to confuse. One knows one’s pain reflexively, simply in virtue of there being pain (in oneself); it isn't first there and later (reflectively) known. More generally, subjective experience is experienced (known) in and of itself, directly and immediately as part of being an experience; there is no additional process that turns it into knowledge. The experience is the experiencing of it – to have an experience is to know that experience.

But this sounds very strange. How can something be its own experience? [3] And yet this is close to the mystery that seems to lie at the heart of consciousness studies. There is a "sense of agency" (discussed more below) that is part and parcel of being an agent. When performing a voluntary ("conscious") action, one knows one is so doing; that knowledge does not arrive later on [4]. But an action that is *known simply in virtue of its being performed*, will be complex enough that it already constitutes a kind of (self-knowing) agent. While space will not allow a thorough treatment of reflexivity, [12] offers a fuller discussion.

This may seem to be going in circles. But we are edging toward an unearthing of less mystery and more practicality.

2 Operationalization

As noted above, an intentional agent will have its intentionality grounded in a reflexive model of a self. Our approach to operationalizing consciousness is via operationalizing intentionality, and this will mean giving precise enough definitions of these terms so that they can be implemented. In this section, we will report on preliminary work in both defining and implementing these concepts.

2.1 Enactive Minimal Self Models

A *minimal self* is roughly a minimal process which could be said to constitute an aware subject. The details vary on what precisely this entails; see (e.g.) [22] and [29] for two different but overlapping approaches to this idea. Our intention is to model subjectivity in very short time-scales and ask what phenomena might be constitutive of such; we explicitly leave out phenomena associated with more reflective notions such as a narrative self [25].

Enactive cognitive science views cognition as something that occurs in embodied agents which act on their environments; further this action is fundamental to the extent that perception is itself an act and we perceive our world according to our ability to act on it. Drawing on the work of Varela and Maturana[30], some variants of the tradition further view the determined boundary between an agent and its environment as the ground of meaning. As such, the tradition

[3] It may be worth noting that the Mahayana Buddhist tradition has debated this questions vigorously, with one party emphasizing the paradoxical nature of any notion of self-knowledge and the other emphasizing that this is precisely what constitutes mental life [14]

[4] This is of course not to say that one is only aware of what one is doing during voluntary action; we thank the anonymous reviewer for pointing this out.

offers a number of useful insights which we draw upon to develop an operational notion of self.

We will focus our discussion on endowing bodily selves with senses of agency and ownership and reflexive self-awareness. We have identified and worked on a number of other essential features of computational selves, but omit discussion of these in the interest of space [5].

In accordance with the enactive tradition, we view selves as agents that act in a world and have knowledge of themselves as such. Two fundamental forms of knowledge will then be what Gallagher has termed a *sense of ownership* and a *sense of agency* [9]. These refer to agents' awareness that their actions are done by them and that their body is theirs. For example, when I move my arm I know that *I* caused it to move and that it is *my* arm that is moving. As our discussion of Alice the robot below will illustrate, such knowledge is essential not just for theoretical reasons but for basic functioning in the real world.

We would like to give these notions a formal treatment – this is useful because it grounds the philosophical concepts. By specifying precisely what we mean by, say, sense of agency, we enable an analysis of the concept and an exploration of the role it plays in a computational model of the self. It also allows a set of criteria against which we can test an implementation; if we argue that an agent endowed with a sense of agency has particular properties then it will be critical that any implementation meet our definition if it is to similarly posses said properties.

We ground our definitions of ownership and agency in the neuroscientific concept of an *efference copy*. This is a copy of a motor command that is thought to be kept by an agent so that a prediction of the command's effect on the world can be compared to observed effects. For example, before moving my hand a half inch left my nervous system might make a prediction about what my hand should look like after the action is complete. Such forward modeling is thought to be the neurological basis of a sense of agency[6] [9]. Thus when a change in my hand's position corresponds to the expect effect of a self-initiated motor command, I will have a sense that I moved my hand intentionally. Conversely, a change which does not correspond to such a motor command will have me looking for an external cause for my hand's movement.

We generalize this story somewhat by allowing for a sense of agency to arise from any kind of "full" representation of an agent's actions with respect to its body. Ideas in [4] and [17] suggest a representation of an action as a mapping of the environment (as reflected in the agent's sensory state) onto some internal state so that when the environment changes the internal state will change accordingly. A straightforward example would be an internal image of the agent's

[5] Some of these are that a self should be: *cognitively situated* with a *first-person perspective*; *reflexively self-aware*; *immediately self-aware* in an *essentially temporal way* and *synchronically and diachronically unified*

[6] This need not be a literal copy of the command, but could (and arguably is) rather some sparse representation of the command that allows for some kind of forward modelling of the command's effect.

body that shifts according to actions taken. If a single action (say rotating left 1 radian) has a consistent effect on the representation (even that effect is rotating *right* one radian) then it will be a representation of the action. However, notice that mapping all of our sensory information onto a single point will work as well; our actions will end up being represented by that single point but this representation is still consistent (if trivial). It is not particularly useful however, so we also insist that as much information as possible be preserved; this can be made formal by either invoking set-theoretics notions like bijectivity or the concept of mutual information. Employing the latter concept gives a differentiable notion that can be deployed in machine learning algorithms. It is worth noting that some philosophers of mind (especially in the analytic tradition) take representation as constitutive of intentionality [27].

We have implemented such senses of agency and ownership in two different projects. The first of these used an analogue of efference copy to allow our robot Alice to recognize when she (as opposed to another agent) is making a particular utterance [7], [6]. When Alice initiates a speech act A, that fact is recorded in her knowledge base, and her perceptual apparatus monitors what happens for comparison with expected results from the success of A. Furthermore, the monitoring and the performance of A are iterated in parallel over tiny time-steps so that ideally there is strong covariance between the two. Thus as Alice speaks, she starts to speak and hears her voice, continues to speak and hear, and then hears her voice stop as she finishes speaking; and in all this she simultaneously knows she is so engaged. Such behavior is not a formal nicety, but rather is central to intelligence. Imagine that a robot hears the utterance “Can you help me?” – it will be crucial to its proper understanding and subsequent behavior, whether it takes this to be an assertion made to it, or by it. Note that [3] takes a very different approach, in having a robot infer that it is speaking based on recognizing the sound of its voice – not on direct knowledge of ongoing voluntary activity.

Another implementation of agency and ownership used the ideas about representations of agency outlined previously to force a deep neural network to represent it’s own agency and body while learning to play Atari games [5]. This resulted in qualitatively sparse representations (over all representations, not just the feature trained to recognize the agent’s body) and improved game-play.

2.2 Self-Awareness and Self-Modifying Utterances

Agency, ownership and situatedness are fundamental properties of enactive minimal self-models which are easily thought of in terms of lower-level, sub-symbolic processing. Phenomenologically, subjects are also essentially characterized by their self-awareness, and this seems better characterized in terms of symbolic processing. Following Husserl (as relayed by Zahavi [31]), we take this self-awareness to be grounded, reflexive and occurring in “thick time” (see below).

The temporal nature of self-awareness was analyzed extensively by Husserl, who argued that awareness is not a phenomenon that unfolds instant-by-instant; rather it is an extended but unified whole that consists of “retention, protention” and “primal impression”. Consistent with this view, and to avoid paradox, we posit that a moment (of awareness) is not the durationless instant of physics, but is rather an interval with small positive duration. This will allow actual processing to occur, and corresponds to what Humphrey [10] calls “thick time” and William James refers to as the “specious present” [11]. By allowing moments to have duration, we are given the opportunity to have something like first-order cognition and something like meta-cognition interact with sufficient resolution to be *interdependent*; we are developing a ”diasynchronic logic” mechanism for this based on the Active Logic formalism with a built-in Now(t) predicate which gives agents an evolving representation for the current time [1], [23], [18].

We are exploiting the features of Active Logic to model an agent’s capacity to reason about their own and others’ ongoing inferences in real time while unifying these into discrete statements. Such agents will be able to reason with changing circumstances and the logical consequences of their thoughts and utterances, knowingly speak truly or falsely, and reason with “benignly self-referential” sentences. In particular, such an agent can utter sentences which self-modify as they unfold, potentially modeling the thought process of a person who is speaking in Spanish, notices her audience seems not to be following, and switches to English, saying (truly, and simultaneously knowing it) “I’m now switching to English.”

The basic mechanisms of diasynchronic logic are intended to model sentences as 1) unfolding over time and 2) demarcated by a self-determined end point. The latter property (modeled on Maturana and Varela’s notion of autopoiesis [13]) allows for logical sentences to be self-unifying – the sentence itself can specify where it stops. The former property allows us to take some of the mystery out of self-reference and ground sentences in their own logical values. And this hints at a resolution between our view and Searle’s: special reflexive processing at multiple and overlapping timescales may be the juice that pulls action and perception, semantics and syntax all together into one self-interacting cognitive whole. Of course, almost everyone who suggests an approach to understanding consciousness seems to arrive at a point where some ”magic” is appealed to. But we claim that our approach can be pursued at a practical – even computational – level.

3 Conclusion

We are in agreement with [24] in that consciousness will become more and more central to AI as the latter pushes deeper into the nature of intelligence. This is especially the case regarding recognition and recovery from errors, which in turn require a detailed and real-time representation of self. Thus, far from being an epiphenomenon, consciousness is part and parcel of what it is to be intelligent: reflexively knowing oneself to be engaged in ongoing processes (that same knowing being among those same processes). And the nature of knowing will be revealed as central and complex, well beyond a mere collection of data.

References

1. Anderson, M.L., Gomaa, W., Grant, J., Perlis, D.: Active logic semantics for a single agent in a static world. Artificial Intelligence **172**(8-9), 1045–1063 (2008)
2. Brentano, F.: Psychology from an empirical standpoint (1973)
3. Bringsjord, S., Licato, J., Govindarajulu, N.S., Ghosh, R., Sen, A.: Real robots that pass human tests of self-consciousness. In: Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on. pp. 498–504. IEEE (2015)
4. Brody, J.: An enactive self-model for sparse representations and improved performance. In: Brazilian Conference on Intelligent Systems (2017)
5. Brody, J.: An enactive self-model for sparse representations and improved performance. In: Intelligent Systems (BRACIS), 2017 Brazilian Conference on. pp. 73–78. IEEE (2017)
6. Brody, J., Perlis, D., Shamwell, J.: Who's talking? efference copy and a robot's sense of agency. In: 2015 AAAI Fall Symposium Series (2015)
7. Brody, J., Perlis, D., Shamwell, J.: Who's talking?efference copy and a robot's sense of agency. In: 2015 AAAI Fall Symposium Series (2015)
8. Chalmers, D.J.: The conscious mind: In search of a fundamental theory. Oxford university press (1996)
9. Gallagher, S.: How the body shapes the mind. Cambridge Univ Press (2005)
10. Humphrey, N.: Seeing red. Harvard University Press (2006)
11. James, W.: The perception of time. The Journal of speculative philosophy **20**(4), 374–407 (1886)
12. Janzen, G.: The reflexive nature of consciousness, vol. 72. John Benjamins Publishing (2008)
13. Maturana, H.R., Varela, F.J.: Autopoiesis and cognition: The realization of the living, vol. 42. Springer Science & Business Media (1991)
14. Nagao, G.M.: Madhyamika and yogacara: a study of Mahayana philosophies. SUNY Press (1991)
15. Nagel, T.: What is it like to be a bat? The philosophical review **83**(4), 435–450 (1974)
16. Perlis, D.: I am, therefore i think. In: APA Newsletter on Phil and Computers. The American Philosophical Association (2017)
17. Perlis, D.: Five dimensions of reasoning in the wild. In: AAAI. pp. 4152–4156 (2016)
18. Perlis, D., Brody, J., Kraus, S., Miller, M.: The internal reasoning of robots (2017)
19. Perlis, D.: Putting one's foot in one's head–part i: Why. Noûs pp. 435–455 (1991)
20. Perlis, D.: Putting one's foot in one's headpart ii: How? In: Thinking Computers and Virtual Persons, pp. 197–224. Elsevier (1994)
21. Perlis, D.: Consciousness and complexity: the cognitive quest. Annals of Mathematics and Artificial Intelligence **14**(2-4), 309–321 (1995)
22. Perlis, D.: Consciousness as self-function. Journal of Consciousness Studies **4**(5-6), 509–525 (1997)
23. Purang, K.: Alma/carne: implementation of a time-situated meta-reasoner. In: Tools with Artificial Intelligence, Proceedings of the 13th International Conference on. pp. 103–110. IEEE (2001)
24. Reggia, J.A.: Conscious machines: The ai perspective. In: AAAI Fall Symposium Series, North America, September (2014)

25. Schechtman, M.: The narrative self. In: Gallagher, S. (ed.) The Oxford handbook of the self. Oxford University Press (2011)
26. Searle, J.R.: Minds, brains, and programs. Behavioral and brain sciences **3**(03), 417–424 (1980)
27. Siewert, C.: Consciousness and intentionality. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2017 edn. (2017)
28. Sopa, G.L.: Cutting Through Appearances: Practice and Theory of Tibetan Buddhism. Shambhala (1989)
29. Strawson, G.: The minimal subject. In: Gallagher, S. (ed.) The Oxford handbook of the self. Oxford University Press (2011)
30. Varela, F., Thompson, E., Rosch, E.: The Embodied Mind. MIT press (1991)
31. Zahavi, D.: Subjectivity and selfhood: Investigating the first-person perspective. MIT press (2008)