

Ch 4. The Neurophilosophy of Consciousness

0. Introduction

THE STORY SO FAR: An account of consciousness needs, to get rolling, a credible answer to the question, “what makes this account an account of *consciousness*?” and appeals to (Deflated) Transitivity, (Deflated) Transparency, and WIL seem to best get us in the ball park. A physicalist account of consciousness is going to need to be a *reductive* physicalist account of consciousness. And if the reductive-physicalist account in question is going to make any kind of use of representation, it better do so in ways that don’t run afoul of unicorns and their inexistent brethren. It increasingly looks like we need a physicalistic representational account of consciousness that is internalistic. What internal things matter most? My bet is on brains. Time to start making good on the bet.

In this chapter I now turn to examine sample neurophilosophical theories of consciousness. I will raise problems for them to be solved in subsequent chapters where I develop my own neurophilosophical account.

In keeping with the remarks in chapter zero on the definition of neurophilosophy as well as the three questions of consciousness (the question of state consciousness, the question of transitive consciousness, and the question of phenomenal character), the discussion of this chapter will be centered on philosophical accounts of state consciousness, transitive consciousness, and phenomenal character that make heavy use of contemporary neuroscientific research in the premises of their arguments.

There are three philosophers whose work on the Neurophilosophy of

consciousness I find especially illuminating to examine in concert: Paul Churchland, Jesse Prinz, and Michael Tye. Sections 1,2, and 3 will be devoted to them, respectively. Section 4 is devoted to initial contrasts and comparisons of the three thinkers. Section 5 is dedicated specifically to contrasts and comparisons regarding phenomenal character and section 6 discusses problems to be solved in subsequent chapters.

1. Churchland

Paul Churchland articulates what he calls the "dynamical profile approach" to understanding consciousness (2002). According to the approach, a conscious state is any cognitive representation that is involved in (1) a moveable attention that can focus on different aspects of perceptual inputs, (2) the application of various conceptual interpretations of those inputs, (3) holding the results of attended and conceptually interpreted inputs in a short-term memory that (4) allows for the representation of temporal sequences.

Note that these four conditions primarily answer the question of what makes a state a conscious one. Regarding the question of what we are conscious of, Churchland writes that "a conscious representation could have any content or subject matter at all" (p. 72) and he is especially critical of theories of consciousness that impose restrictions on the contents of conscious representations along the line of requiring them to be self-representational or meta-representational (pp. 72 – 74).

Much of Churchland's discussion of the dynamical profile account of consciousness concerns how all of the four conditions may be implemented in recurrent neural networks. A recurrent neural network may be best understood in terms of contrast

with feed-forward neural networks. The neural networks discussed in chapter 1 were feed-forward networks. In feed-forward networks, the flow of information is strictly from input to output (via interneurons if any are present). In recurrent networks there are feedback (or "recurrent") connections as well as feed-forward connections. (For further discussion of artificial neural networks, see Garson 2002).

Let us turn now to Churchland's account of how the four elements of the dynamical profile of conscious states might be realized in recursive neural networks. It helps to begin with Churchland's notion of the conceptual interpretation of sensory inputs and we do well to begin with what Churchland thinks a concept is. Consider a connectionist network with one or more hidden layers that is trained to categorize input types. Suppose that its inputs are a retinal array to which we present grayscale images of human faces. Suppose that its outputs are two units, one indicating that the face is a male and the other indicating that the face is female. After training the configuration of weights will be such that diverse patterns of activation in the input layer provoke the correct response of "male" to the diversity of male faces and "female" for female faces. For each unit in the hidden layer, we can represent its state of activation along one of several dimensions that define activation space. A pattern of hidden layer activation will be represented as a single point in this space. This space will have two regions: one for males and one for females. Regions in the center of each of the two spaces will constitute "attractors" that define what, for the network, constitutes prototypical female faces and prototypical male faces, respectively.

The addition of recurrent connections allows for information from higher layers to influence the responses of lower layers. As Churchland puts the point:

This information can and does serve to 'prime' or 'prejudice' that neuronal population's collective activity in the direction of one or other of its learned perceptual categories. The network's cognitive 'attention' is now preferentially focused on one of its learned categories at the expense of the others. (p.75)

Churchland is not explicit about what this might mean in terms of the example of a face categorization network, but I suppose what this might mean is that if the previous face was a prototypical female, then the network might be more likely to classify an ambiguous stimulus as female. We can construe this as exogenous cueing of attention. Churchland goes on to further describe shifts of attention in recurrent networks that we might regard as endogenous. "Such a network has an ongoing *control* of its topical selections from, and its conceptual interpretations of, its unfolding perceptual inputs." (p.76).

Recurrent connections allow for both a kind of short term memory and the representation of events spread out over time. In a feedforward network, a single stimulus event gives rise to a single hidden layer response then a single output response. With recurrence however, even after the stimulus event has faded, activity in lower layers can be sustained by information coming back down from higher layers and that renewed lower-level activity can itself reactivate higher layers. Also, what response a given stimulus yields depends in part on what the previous stimuli were. Thus do recurrent connections implement a kind of memory. (The rate of decay may be modulated by modulating connection weights, so that, for instance, greater weights result in longer-term memory.) The ability to hold on to information over time allows for the representation of

events spread out over time, according to Churchland, and the representation in question will not be a single point in activation space but a trajectory through it.

The discussion so far has been rather abstract, in the sense of abstracting away from neural details, and, unfortunately, Churchland (2002) does not go into much neuroanatomical or neurophysiological detail. Of course, he does advert, though tentatively, to the account in Churchland (1995) wherein he endorses Llinas' view whereby consciousness involves recurrent connections between the thalamus (a bilateral structure at the rostral tip of the brainstem) and cortex. Part of the appeal of localizing consciousness in these structures presumably involves the role hypothesized for recurrence as well as the ideas that consciousness involves systems responsible for wakefulness and arousal (thalamus), diverse “higher” functions (the various portions of the cortex), and a system that can act as a relay between the various “higher” functions (the thalamus again).

I will have more to say about this later, but for now I will briefly summarize Churchland's dynamic profile account with respect to the three questions of consciousness as follows. With respect to the question of state consciousness, according to Churchland, conscious states are neural representations that have a particular dynamic profile. With respect to the question of transitive consciousness, Churchland's account imposes no limitations on what one can be conscious of: one could be conscious of just about anything according to Churchland. With respect to the question of phenomenal character, ‘what it is like’ to have a conscious state is going to be determined by the representational content of that state. More will be said about these points after we've had opportunity to examine some other neurophilosophical theories of consciousness.

2. Prinz

The neurophilosophical account of consciousness by Prinz (2000, 2004) is relatively similar to Churchland's and fills in a lot of neuroanatomy and neurophysiology that Churchland leaves out. Prinz characterizes the processing hierarchy I discussed earlier (In ch. 1, sec 6, "Interlude: The Neuroscience of Visual Consciousness") and then notes that the contents of consciousness seem to match it with representations at the intermediate level of processing (areas v2-v5). This means that the contents of conscious states do not abstract entirely from points of view as do the highest level of the processing hierarchy but neither are they the same as the representations at the lowest level. However, Prinz argues that intermediate representations are alone insufficient for consciousness. They must additionally be targeted by attention. Prinz thinks attention is required because of considerations having to do with the pathology of attention known as "neglect." Prinz cites Bisiach's (1992) study of neglect patients who were able to demonstrate certain kinds of unconscious recognition. Prinz infers from such results that not only did high-level stuff areas in the visual hierarchy get activated (they are necessary for the kinds of recognition in question) but also that intermediate levels had to have been activated. (Prinz seems to be assuming that information can only get to higher levels of cortical processing by way of intermediate level but one wonders if perhaps the intermediate level of was bypassed via a sub-cortical route.)

Given the large role that Prinz assigns to attention in his theory of consciousness, the question naturally arises as to what Prinz thinks attention is and what it does. Prinz endorses the account of attention by Olshausen, Anderson, and van Essen (1994) whereby attention involves the modulation of the flow of information between different

parts of the brain. Further, Prinz endorses the speculation that the attention crucial in making intermediate level representations conscious involves a mechanism whereby information flows from intermediate areas, through high-level visual-areas (infero temporal cortex) to working memory areas in lateral prefrontal cortex. According to Prinz, pieces of information in working memory “allow the brain to recreate an intermediate-level representation by sending information back from working memory areas into the intermediate areas.” (2004, p. 210). Prinz (2000) summarizes, emphasizing attention’s role, as follows:

When we see a visual stimulus, it is propagated unconsciously through the levels of our visual system. When signals arrive at the high level, interpretation is attempted. If the high level arrives at an interpretation, it sends an efferent signal back into the intermediate level with the aid of attention. Aspects of the intermediate-level representation that are most relevant to interpretation are neurally marked in some way, while others are either unmarked or suppressed. When no interpretation is achieved (as with fragmented images or cases of agnosia), attentional mechanisms might be deployed somewhat differently. They might “search” or “scan” the intermediate level, attempting to find groupings that will lead to an interpretation. Both the interpretation-driven enhancement process and the interpretation-seeking search process might bring the attended portions of the intermediate level into awareness. This proposal can be summarized by saying that visual awareness derives from Attended Intermediate-level Representations (AIRs). (p. 249)

Prinz's account of attention's role in consciousness seems a lot like Churchland's conceptual interpretation, short term memory, and of course, attention requirements on consciousness. That is, on Prinz's picture, and perhaps on Churchland's as well, conceptual interpretation and short-term memory aren't making contributions to consciousness separate from the contributions attention makes. Their contributions are, instead, constitutive of the contributions of attention to consciousness.

3. Tye

Tye raises objections to the sort of view advocated by Churchland and Prinz. Tye is critical of accounts of consciousness that build in constitutive roles for attention. Tye's claim is based on intuitive grounds (1995, p. 6). It is supposed to be intuitive one might have a pain for a length of time but not be attending it the entire time. Tye insists that there is still something it's like to have an unattended pain. It makes sense that someone committed to Transparency wouldn't think that attention is a requirement for consciousness insofar as attention is interpreted as a faculty by which one becomes conscious of conscious states, since Transparency involves the denial of the possibility of being conscious of conscious states. However, it is not clear that Tye's remarks about unattended pains are supposed to depend on a prior acceptance of Transparency. I merely mean to point out here that it makes sense that the two sorts of considerations would go together.

(It should also be noted that Prinz does not intend attention to be metarepresentational, and explicitly rejects a version of the AIR theory he calls HOT AIR (Prinz xxxx).)

Tye infers from these sorts of considerations that the neural correlate of visual consciousness is lower in the processing hierarchy than an attention-based theory would locate it. Tye thus locates the neural correlates of conscious states in “the grouped array” located in the occipital lobe and, regarding the phenomenon of blindsight, rejects “the hypothesis that blindsight is due to an impairment in the linkage between the spatial-attention system and the grouped array” (Tye 1995 p. 215-216) Tye accounts for the retained visual abilities of blind-sight subjects (p. 217) in terms of a sub-cortical pathway, the “tectal-pulvinar pathway”, from retina to superior colliculus that continues through the pulvinar to various parts of the cortex, including both the parietal lobe and area v4. Thus Tye seems to think consciousness is in V1. Prinz 2000 argues against this, citing evidence against locating consciousness in V1 (see Crick & Koch, 1995, and Koch & Braun, 1996, for reviews). Prinz writes:

As Crick and Koch emphasize, V1 also seems to lack information that is available to consciousness. First, our experience of colors can remain constant across dramatic changes in wavelengths (Land, 1964). Zeki (1983) has shown that such color constancy is not registered V1. Second, V1 does not seem responsive to illusory contours across gaps in a visual array (von der Heydt, Peterhans, & Baumgartner, 1984). If V1 were the locale of consciousness, we would not experience the lines in a Kanizsa triangle.” (pp. 245-246).

4. All Together Now

Turning from disagreements to agreements, we may note that Churchland, Prinz, and Tye agree that conscious states are representational states. They also agree that what

will differentiate a conscious representation from an unconscious representation will involve relations that the representation bears to representations higher in the processing hierarchy. For both Churchland and Prinz, this will involve actual interactions, and further these interactions will constitute relations that involve representations in processes of attention, conceptual interpretation and short term memory. Tye disagrees on the necessity of actually interacting with concepts or attention. His account is dispositional meaning that the representations need only be poised for uptake by higher levels of the hierarchy.

Turning to the question of transitive consciousness, we see both agreements and disagreements between the three authors. For Churchland, Tye, and Prinz, they all agree that what one is conscious of is the representational content of conscious states. In all cases what the subject is conscious of is what the representational contents of the conscious states are. However, these theorists differ somewhat in what they think the contents can be. Churchland has the least restrictive view: any content can be the content of a conscious state. Prinz's is more restrictive: the contents are not going to include high-level and invariant contents.

Tye's is the most restrictive: the contents will only be first-order and non-conceptual. Tye thinks that they are non-conceptual since he thinks that creatures without concepts—perhaps non-human animals and human infants—can have states for which there is something it is like to have them even though they possess no concepts. Tye says little about what concepts are and for this, among other reasons, it is difficult to evaluate his view. The reason Tye thinks the contents of consciousness are first-order is largely

because he believes in the pre-theoretic obviousness of Transparency. As I've spelled out in chapters two and three, I'm not terribly impressed with appeals to Transparency.

I do think, however, that the points upon which Tye, Prinz, and Churchland largely agree concerning transitive and state consciousness are points upon which I largely agree. I will develop this further in chapters five and six.

5. What About What It's Like?

I turn now to what neurophilosophical accounts have to say about phenomenal character. I focus, in particular, on the suggestion that phenomenal character is to be identified with the representational content of conscious states and will discuss this in terms of Churchland's suggestion of how qualia should be understood in terms of neural state spaces.

Our experience of color provides the most often discussed example of phenomenal character by philosophers and Churchland is no exception. While Churchland's view of the neural reduction of qualia has already been mentioned in chapters 1 and 2, it's worth delving into further detail here.

When Churchland discusses color qualia, he articulates a reductive account of them in terms of Land's theory that human perceptual discrimination of reflectance is due to the sensory reception of three kinds of electromagnetic wavelengths by three different

kinds of cones in the retina.³⁴ In keeping with the kinds of state-space interpretations of neural activity that Churchland is fond of, he explicates color qualia in terms of points in three dimensional spaces, the three dimensions of which correspond to the three kinds of cells responsive to electro-magnetic wavelengths. Each color sensation is identical to a neural representation of a color (a neural representation of a spectral reflectance). Each sensation can thus be construed as a point in this three dimensional activation space and the perceived similarity between colors and the subjective similarities between corresponding color qualia are definable in terms of proximity between points within the three-dimensional activation space.

“Evidently, we can reconceive [sic] the cube [depicting the three dimensions of coding frequencies for reflectance in color state space] as an internal ‘qualia cube’”(1989, p. 105). Churchland thinks this approach generalizes to other sensory qualia, such as gustatory, olfactory, and auditory qualia (ibid, p. 105-106). Bringing this view in line with the thesis of the direct introspection of brain states, Churchland writes:

The “ineffable” pink of one’s current visual sensation may be richly and precisely expressible as a 95Hz/80Hz/80Hz “chord” in the relevant triune cortical system.

The “unconveyable” taste sensation produced by the fabled Australian health tonic Vegamite [sic.] might be quite poignantly conveyed as a 85/80/90/15

“chord” in one’s four-channeled gustatory system (a dark corner of taste-space

³⁴ In later work, Churchland’s favorite color theory changes, resulting in a difference in what the three dimensions of color space are. But the basic story of a three dimensional color space remains the same.

that is best avoided). And the “indescribable” olfactory sensation produced by a newly opened rose might be quite accurately described as a 95/35/10/80/60/55 “chord” in some six dimensional system within one’s olfactory bulb.

This more penetrating conceptual framework might even displace the commonsense framework as the vehicle of intersubjective description and spontaneous introspection. Just as a musician can learn to recognize the constitution of heard musical chords, after internalizing the general theory of their internal structure, so may we learn to recognize, introspectively, the n -dimensional constitution of our subjective sensory qualia, after having internalized the general theory of *their* internal structure (ibid, p. 106).

Three particular and related features of Churchland’s view of qualia are of special note. The first is that qualia are construed in representational terms. The second follows from the first, namely, that qualia so construed are not intrinsic properties of sensations and thus overturns a relatively traditional view of qualia. The third is that it allows for intersubjective apprehensions of qualia. I turn now to discuss these three points in further detail.

To construe qualia in terms of representational content is to construe them as extrinsic properties of the states of which they are properties, since typical accounts will spell out representational content in terms of either (1) causal relations sensory states bear to states of the external world, or (2) causal relations that they bear to other inner states, or (3) some combination of the two sorts of relations. In neural terms, a pattern of activation in a neural network is the bearer of representational content in virtue of either (1) the distal or proximal stimuli that elicit the activation, or (2) other patterns of

activation that influence it via, e.g., recurrent connections or (3) some combination of the two. For reasons largely related to the concerns raised in chapter three, I'm not particularly enthusiastic about (1). I will sketch an account more in keeping with (2) in chapter seven. Assuming that something along the lines of (2) is viable, then while qualia turn out not to be intrinsic properties of individual states, they may nonetheless turn out to be intrinsic to nervous systems, and thus fully compatible with the version of neural reductionism it is my main aim in this book to defend.

6. Problems, Problems, Problems, or; Seriously, What About What It's Like?

The third issue raised by Churchland's account of qualia is how it opens the possibility of qualia being intersubjectively apprehended. I assume that insofar as they are intersubjective, then they are not truly subjective. I want to question, however, whether Churchland supplies a complete case for the intersubjectivity of qualia.

The above quoted passage contains Churchland's view that properties of neural states previously inexpressible could, if one acquired the relevant neuroscientific concepts and the skill to apply them introspectively, become expressible. However, this view seems to be in tension with the earlier mentioned view, (chapter 2) that concepts influence phenomenal character. The phenomenal character of an experience prior to the acquisition and introspective application of a concept will not, then, be the same as the phenomenal character of an experience after the acquisition and introspective application of that concept. Compare this to Nagel's (19xx) remarks on attempting to know what it's like to be a bat by donning wings and sonar mechanisms: such machinations would only

give knowledge of what it would be like for a human to emulate a bat, not knowledge of what it's like for bats. Thus, it doesn't yet look like Churchland has provided an escape from subjectivity, since there may remain certain representational contents of neural states that are directly and fully knowable only by the subject who has them. Insofar as such properties are unknowable from a neural point of view, doubts are raised about the adequacy of neural reductionism. I will return to the problem of the alleged subjectivity of conscious states in chapter nine.

The problem of the subjectivity of conscious states pales in comparison to a version of the hard problem concerning why it is like anything at all to be in the various neural states described in this chapter. Couldn't the various neural states described in the current chapter be had by zombies? The goal of chapter five is to address the problem of zombies in a way that will set the stage for the neurophilosophical theory of conscious to be articulated in chapter six.