Mandik, P. (in press). The Neurophilosophy of Subjectivity. In John Bickle (ed.), *Oxford Handbook of Philosophy and Neuroscience*. New York: Oxford University Press.

The Neurophilosophy of Subjectivity

PETE MANDIK, Department of Philosophy, William Paterson University of New Jersey.

ABSTRACT: The so-called subjectivity of conscious experience is central to much recent work in the philosophy of mind. Subjectivity is the alleged property of consciousness whereby one can know what it is like to have certain conscious states only if one has undergone such states oneself. I review neurophilosophical work on consciousness and concepts pertinent to this claim and argue that subjectivity eliminativism is at least as well supported, if not more supported, than subjectivity reductionism.

§ 0. Introduction

Conscious experience, according to many philosophers, is subjective. Claims about the so-called subjectivity of consciousness are offered as apparently obvious facts about consciousness. Further, in much philosophical literature, these supposedly obvious claims are exploited as the bases for some not-at-all-obvious conclusions, like, for instance, the conclusion that no neuroscientific theory could possibly explain consciousness. If such claims are correct, then the neurophilosophy of consciousness is somewhere between impossible and ridiculous.

In this article, I will present a case that claims of the subjectivity of consciousness are far from obvious or innocent and are instead very strong empirical claims about the structure, acquisition, and content of concepts. As such, it is entirely appropriate to bring empirical considerations to bear on the question of whether experience is essentially subjective. I describe various neurophilosophical accounts of the relevant notions (concepts, consciousness, sensation, introspection, etc.) and build a case against the subjectivity of consciousness. Along the way I discuss the prospects of neurophilosophical accounts of subjectivity and argue for the superiority of subjectivity eliminativism over subjectivity reductionism.

My plan is as follows. First, I conduct a quick review of the notion of subjectivity as it figures in some classic discussions in the philosophy of mind, especially those surrounding the work of Nagel and Jackson. I develop the idea that the account of subjectivity is one-way knowability. Next, I turn to discuss neurophilosophical perspectives on the topics of consciousness, phenomenal character, and "knowing what it is like". Finally, I bring the insights developed in the previous sections to bear on the twin questions of whether (1) in perception, we perceive properties that may be known in no other way and (2) in introspection we introspect properties that may be known in no other way. My conclusions will be that both questions merit negative answers.

§ 1. Subjectivity and the Philosophy of Mind

What, in the context of the philosophy of mind, is subjectivity? Subjectivity has something to do with consciousness, but it is not consciousness itself. Subjectivity has something to do with the so-called phenomenal character of conscious states, but it is not identical to phenomenal character. Subjectivity is an alleged property of phenomenal character, namely, the property of being one-way knowable. More specifically, the claim that phenomenal character is subjective is the claim that the only way to know some phenomenal character is by *having* a conscious experience that has that character. (This is a first pass and will be refined further later.) Whatever the relevant sense of "know" is here, it is the sense relevant to "knowing what it is like" to have some conscious experience. Before we further refine these characterizations, some brief historical remarks are in order. Much contemporary discussion of the related notions of subjectivity and "what it is like" stem from the work of Thomas Nagel (1974). Nagel got philosophers worried about the question "what is it like to be a bat?" and urged that since bat experience must be so very different from our own, we could never know. Nagel further suggested that no amount of knowledge about bat behavior or bat physiology could bridge the gap. While Nagel didn't draw dualistic conclusions from these considerations, other philosophers did. In particular, Jackson (1982) and subsequent commentators developed an argument against physicalism based on the special ways in which what it is like to have conscious states with certain phenomenal characters must be known. The most famous version of Jackson's ensuing "knowledge argument" centered on a thought experiment about a woman, Mary, who somehow knows all of the physical facts, especially those about the neural basis of human color vision, without ever herself having seen red. Upon seeing red, Mary allegedly must necessarily be surprised and thus learns only then what it is like to see red. Since she knew all of the physical facts before coming to know what it is like to see red, knowing what it is like must be knowing something non-physical. It is at this point that we can attempt some further refinement of the subjectivity claim. At a first pass, we might try to characterize the key idea behind Nagel's and Jackson's arguments as

> (K): For all types of phenomenal character, in order to know what it is like to have a conscious experience with a phenomenal character of a type, one must have, at that or some prior time, a conscious experience with a phenomenal character of the same type.

One appealing feature of (K) is that it would justify the claims that humans can't know what it is like to have bat experiences and people who have only seen black and white can not know what it is like to see red. However, even fans of Nagel and Jackson are likely to reject (K) on the grounds that there are many types of phenomenal characters for which (K) is highly implausible. Suppose that there was some shade of gray or some polygon that Mary had never seen before. Few philosophers are likely to suppose that Mary would be surprised on seeing a 65.5% gray or a 17-sided polygon for the first time. Perhaps, then, the idea behind subjectivity considerations is better put by modifying (K) by replacing "the same" with "a relevantly similar" and replacing "all" with "at least one," resulting in the following.

(K+): For at least one type of phenomenal character, in order to know what it is like to have a conscious experience with a phenomenal character of a type, one must have, at that or some prior time, a conscious experience with a phenomenal character of a relevantly similar type.

Thus we have, in K+, an explication of what it means to equate the subjectivity of phenomenal character with the one-way-knowability of phenomenal character.

More remains to be said on what the sense of knowledge is that is most relevant here. Proponents of the knowledge argument for dualism have interpreted knowing what it is like as a kind of propositional knowledge. Some physicalists have granted this assumption while others have questioned it, offering that a different kind of knowledge, such as know-how, constitutes knowing what it is like (Nemirow 1980, 1990; Lewis 1983, 1988). In this article I will simply grant, without further discussion, the propositional assumption, but for a longer discussion of this point see Mandik (2001). It is thus assumed that knowing what it is like to have a conscious experience with some phenomenal character *C* is the same as knowing some proposition *P*. It is further assumed that *P* is the proposition that having a conscious experience with some phenomenal character *C* is like such-and-such. There are two general possible reasons why *C* would be one-way knowable, one having to do with belief and the other having to do with warrant. The belief reason says that one cannot know *P* without having *C* (I intend "having *C*" as shorthand for "having a conscious experience with phenomenal character *C*") because one is incapable of believing *P* without having *C*. The warrant reason grants that one can believe *P* without having *C* but claims that one's belief that *P* cannot have sufficient warrant to count as knowing *P* unless one has had *C*. Neurophilosophical work pertinent to subjectivity has mostly concerned the belief claim and in this chapter my attention will be similarly restricted. We must turn now to ask why it would be the case that believing *P* requires having *C*. A natural kind of answer to such a question is one that appeals to concepts. What one can or cannot believe depends on what concepts one does or does not have. If one has no concept of, e.g. dark matter, then one cannot have beliefs about dark matter. Several authors have proposed that we have special concepts—so-called *phenomenal concepts*—by which we think of the phenomenal characters of our conscious states. Part of what is supposed to be special about phenomenal concepts is that one can acquire a phenomenal concept of *C* only by having conscious experiences with phenomenal character *C*. While there are various ways of construing phenomenal concepts, I shall focus here only on those construals whereby the assertion of the existence of phenomenal concepts is inconsistent with the falsity of K+. (For examples of such construals, see Papineau, 1999; and Tye, 2000) If there are such things as phenomenal concepts, then they can explain the one-way knowability of *C* that is constitutive of the subjectivity of *C*: *C*'s being subjective is due to *C*'s being one-way conceptualizable. The phenomenal concept of *C* is the unique concept that picks out *C* as such.

I will argue that when viewed from a neurophilosophical point of view, the phenomenal concepts proposal looks very strange indeed. For a brief sketch of how one might attempt to ground phenomenal concepts in neuroscience, we can turn to a recent proposal by Beaton (2005):

We would be wrong to think that, when Mary is released, just gaining the correctly configured sensory apparatus is sufficient for her to know what it is like

> to see red. Just picking out red—in V4 say—without any connection to the brain regions responsible for additional, more conceptual abilities, would be the equivalent of blindsight. In order for Mary to come to consciously see red, she has to be able to *report* that she is seeing red now, to be able to choose to remember and imagine it at will, and to act on her imagination as appropriate, after having seen it. It seems highly likely that these more complex, conceptual abilities are functions of the associative and frontal regions of the human brain…, regions which are functionally distinct from lower level sensory cortex in the important sense that sensory cortex can continue to effectively carry out the vast majority of its tasks without the presence of higher brain regions, whilst the converse is not true….
>
> On the present view then, on exposure to colour, Mary gains a new configuration in her sensory cortex (specifically in the region dedicated to the processing of colour), but she additionally gains a new neural configuration in her associative and/or frontal cortical regions. This additional configuration corresponds to Mary's having gained a new concept, a concept which I will gloss as '*red_as_experienced*'….[S]he now posseses a new concept, of red as experienced, grounded iin the very sensory apparatus which enables her to detect and respond to red stimuli (pp. 13-14).

Before further developing neurophilosophical considerations appropriate for evaluating the phenomenal concepts proposal, it will be useful to sketch some neurophilosophical accounts of consciousness, character, and knowing what it is like.

§ 2. The Neurophilosophy of Consciousness, Character, and Knowing What it is Like

There is no shortage of philosophical theories of consciousness these days and the several merit the label "neurophilosophical". Neurophilosophy is oft distinguished from philosophy of neuroscience in that the former involves the application of neuroscientific concepts to philosophical problems and the latter is the philosophical examination of neuroscience (Bickle, Mandik, & Landreth, 2006). Neurophilosophical theories of consciousness bring neuroscientific concepts to bear on philosophical problems concerning consciousness. In Mandik (2006) I review neurophilosophical accounts of consciousness due to Churchland, Prinz and Tye and in Mandik (2005; in press) I present one of my own. Here I attempt a sketch of the main outlines of the latter theory, though the differences between them are negligible with respect to the issues of significance in the current chapter.

There are three main questions that philosophical theories of consciousness are concerned to answer. The first concerns the question of what the difference between conscious states and unconscious states amounts to. States of the retina carry information about the distribution of visible light arrayed in front of the organism, but retinal states are alone insufficient for consciousness. Higher-level states, such as one's abstract knowledge that dogs are mammals, are likewise insufficient for consciousness—you have known for some time that dogs are mammals but were unlikely to be consciously thinking that dogs were mammals a few moments ago. The second question concerns what it is that one is conscious of when one has conscious states. I have just now become conscious of the noise that my air conditioner makes, although I am quite sure it has been

making that noise before I became conscious of it. In what does this aspect of consciousness consist? Finally, the third question of consciousness is the one that most preoccupies philosophers working on consciousness: the question of what phenomenal character consists in.

To convey the outlines of the neuroscience relevant to consciousness it will be useful to focus on just one kind of consciousness, namely, visual consciousness and the basic relevant neuroscience concerning vision. Visual information is processed through various levels in a visual processing hierarchy. The lowest levels are at the sites of transduction in the retina. Progressively higher levels of processing are located in, in order, the sub-cortical structures of the lateral geniculate nucleus (LGN), primary visual area (V1), various cortical areas in the temporal and parietal lobes, and, at the highest levels, areas in the frontal cortex and hippocampus. Two main features of the hierarchy are especially worth noting. The first is that distinctive kinds of perceptual information are represented at distinct levels. The second is that information flows not only from lower to higher levels, but also back down from higher to lower. This first point, concerning distinctive kinds of information, is especially useful for identifying which states of neural activation are most relevant to various conscious states. So, for example, one element of conscious visual perception involves subjective contour illusions such as Kaniza triangles, however, neural activity in V1 does not reflect the presence of subjective contours and activity in V2 does (von der heydt, Peterhans, & Baumgartner, 1984). One general remark that can be made about the difference in the information represented at the different levels is that at the lowest levels the information is very specific and is increasingly abstract at higher levels. For example, color constancy is registered at levels higher than V1 (Zeki 1983). Very low levels are sensitive to the specific colors, viewer-relative locations, and orientations of specific stimuli. At higher levels, neural activity indicates less of a preference for such specificities. These various facts about the representational contents of conscious experiences and the kinds of information present at the various levels of the hierarchy help to locate conscious states at relatively intermediate levels of the hierarchy. However, simply being an intermediate level state of neural activation does not suffice for being a conscious state, and this is where the second crucial fact about the processing hierarchy comes into play. Conscious perceptual states are those involved in neural activations that involve the bottom-up activation of intermediate levels which are also undergoing modulation by various top-down influences. According to the neurophilosophical theory of consciousness spelled out in Mandik (2005; in press), a conscious state is a hybrid composed of a pair reciprocally interacting states, one of which is higher in the hierarchy than the other. Evidence for the need for such reciprocal interaction includes results of transcranial magnetic stimulation experiments by Pascual-Leone and Walsh (2001) whereby V5 activity sufficed for conscious motion perception only if information was allowed to feed back down from V5 to V1. Another line of evidence for the reciprocity hypothesis comes from Lamme et al. (1998) wherein anesthetized animals show brain activity in response to stimuli, but only of a feed-forward type that involved no feedback from higher to lower levels. Regarding the three questions of consciousness, the question of what makes states conscious has already been addressed: states are conscious when they are intermediate level hybrids parts of which are reciprocally interacting representations. Moving on to the question of what it is that we are conscious of when we have conscious states, the natural

suggestion is that we are conscious of the representational contents of the reciprocally interacting states. The answer to this question is closely related to the question of phenomenal character: what constitutes what it is like to be in a conscious state is what the conscious state represents, that is, the character of a conscious state is one and the same as the way the states represents things to be. (This is not necessarily to adopt a first-order representationalist position such as Dretske (1995) or Tye (1995; 200), for it is open that some of the things that are represented are other mental states and thus sometimes, though not always, the conscious states will involve higher-order representations.)

Visual consciousness comes in two modes that are relevant to the present discussion: perception and introspection. The following account is adapted from Mandik (2006a; in press), which builds on Churchland (1979). First, we can characterize perception in a way that is neutral with respect to whether the perception in question is conscious perception. Thus, the account of perception is that it is the automatic conceptual exploitation of the information sensations carry about events in the external world (in exteroception) and in the body (in interoception). Thus, when one sees a coffee mug, one has a sensation that carries information about the coffee mug and this sensation in turn elicits the automatic (non-inferential) application of certain concepts, such as the concept of a mug. Introspection is the automatic (non-inferential) conceptual exploitation of the information that mental states carry about themselves.

In terms of the visual processing hierarchy, we can identify the concepts in question with representational states that are relatively high in the hierarchy and the sensations with states relatively low in the hierarchy. Unconscious perception takes place when a sensation elicits the application of a concept without feedback from the conceptual level back down to the sensational level. A conscious perception occurs when the feedback occurs in addition to the processes that alone would suffice for unconscious perception.

The account of introspection is modeled on the account of perception. Where perception of external events is the automatic conceptual exploitation of information that sensations carry about those events, the introspection of sensation is the automatic conceptual exploitation of information that sensations carry about themselves. When the sensations are in appropriately reciprocal interactions with the elicited concepts, the introspection involved is the kind relevant to discussions of knowing what it is like.

Perception and introspection are ways of getting knowledge. But the open question is whether the knowledge thereby gotten can only be gotten in those ways. To address the question, it will be helpful to give a characterization of what the knowledge consists in that is relatively neutral on the question at hand. Leaving aside knowledge of what it is like for a moment, let us consider some general remarks about knowledge and some prototypical instances of knowledge. Consider, for example, your knowledge of how tall you are. There was some time at which you acquired that knowledge and it was likely a time in which you were undergoing a conscious occurrent mental state the content of which is that your height is such and such. This information was then stored to be available for later retrieval. You may very well be retrieving it right now and thus undergoing another conscious occurrent mental state, this time the conscious thought that your height is such and such. However, in the expanse of time between acquisition and retrieval you did know what your height was even though you were not undergoing an

occurrent conscious mental state with that content. Your stored knowledge over the intervening period may or may not count as an occurrent mental state, but it is clear that it does not count as a conscious mental state. As such, you can have that knowledge even at times that you are subject to no conscious states at all (as in, perhaps, when you are in a deep sleep). Putting this in terms of concepts, one can have a concept at a time in at least some cases in which one is not at the same time making use of that concept in an occurrent conscious mental state. Another point that this example allows us to highlight is that even though the knowledge of your height was acquired by having a conscious experience, no particular kind of conscious experience was required to learn that your height was such and such. You could have equally well learned that fact even if you were color blind or even totally blind. Putting this in terms of concepts, the concepts involved in knowing that your height is such-and-such can be acquired regardless of which particular type of conscious experience concerned your height being such-and-such.

Let us return now to the question of knowing what it is like. The intuition, K+, lying behind the subjectivity claim entails that there is at least one kind of knowledge that can be had only if one has had or is having an experience of a certain type. The phenomenal concepts strategy under discussion is making a claim about the relevant concepts in question, namely that the concepts constitutive of knowing what it is like cannot be acquired without undergoing, now or previously, an experience with the phenomenal character in question. This is a claim made by physicalists as well as anti-physicalists, so it is fair to ask whether a neurophilosophical defense of the claim can be given. There are two general strategies one might pursue in defending the claim about concepts. The first strategy defends it in terms of the structure of the concepts themselves. The second strategy defends it in terms of the semantic facts about the representational contents of the concepts. I turn now to address these strategies in sections 3 and 4 respectively.

§ 3. The Structure of Concepts Defense of Subjectivity

There are various things that the postulation of concepts is supposed to explain, but for current purposes we can focus on just two: categorization and inference. When I categorize diverse visual stimuli, say those from a seeing a china pug and those from seeing a Doberman, as both indicating examples of dogs, it is my concept of dogs that allows me to do this. When I draw an inference, for example, that this china pug is a mammal, on the bases of my prior belief that all dogs are mammals, the inference is enabled by my possessing, among other things, a concept of mammals. A neurophilosophical account of concepts must, at a minimum, provide for an account of what concepts are such that they can play these roles in categorization and inference.

For a relatively simple neural model of how concepts figure in categorization, we can look to Paul Churchland's (1989) account of concepts in terms of feed-forward neural networks. In the neural models in question, the networks consist in a set of input neurons, a set of output neurons, and a set of "hidden" neurons intervening between the inputs and outputs. Neurons are connected by various connections that can be “weighted,” meaning that the connections maybe assigned values that determine how much influence the activation of one neuron can have on the activation of another. Information flows through the network along the connections. A network is a strictly feed-forward network if information flows only from input neurons to hidden neurons and from hidden neurons to

output neurons. In a massively connected feed-forward network, each input neuron is connected to every hidden neuron and each hidden neuron is connected to every output neuron. Learning takes place by making adjustments to the connection weights. To consider an example of the learning of a categorization task, consider a network trained to visually discriminate cats from dogs. The inputs are an array of retinal cells upon which black and white bitmapped photographs of dogs and cats may be projected. The output units are two units, labeled "dog" and "cat" respectively. Initially, the connection weights are set to random values and the network is expected to be at chance at correctly identifying a given photograph as of a dog or of a cat. But by making gradual adjustments to the weights, (via the application of a learning rule based on a measurement of the difference between the error and the correct response) the network can come to learn to make the correct classification. Churchland proposes to identify concepts with attractors in hidden unit activation space. For each unit in the hidden layer, we can represent its level of activation along a dimension. The multiple units thus define a multidimensional state space. The presentation of a photograph to the network will result in a pattern of activation across the units that can be defined as a single point in activation space. The points corresponding to presentations of dogs will cluster closer together in activation space than the points corresponding to presentations of cats. The network's concept of dog is a region of activation space, the center of which determines a dog that would be perceived as the prototypical dog.

Churchland's account of concepts seems to supply a case for a kind of concept empiricism relevant to addressing the subjectivity claim. Of course, we need to put aside the very difficult question of how concepts would be concepts of the sensations instead of concepts of the same distal stimuli that the sensations are sensations are of. Additionally, we need to assume that the concepts involved in knowing what it is like to see red are learned. Whatever concepts are acquired, are acquired only when or after the sensations are had. And this is in keeping with K+. However, these simple networks have serious shortcomings as models. Feed-forward networks, lacking lateral and recurrent connections, are poor models of consciousness, for they lack recurrent connections and for similar reasons are poor models of concepts, for they cannot account for inference. Of course, this suggests that we should consider slightly more complicated neural models of concepts, ones that have lateral and feedback connections in addition to feed-forward connections. One such model is the one that figures centrally in what Damasio and Damasio (1994) call "convergence zones". According to Damasio and Damasio, a convergence zone is a neuronal grouping in which multiple feedback and feed-forward connections from separate cortical regions converge. A convergence zone generates patterns of activity across widely separated neuronal groupings by sending signals back down to multiple cortical regions. Damasio and Damasio also postulate that convergence zones form hierarchical organizations in which higher-level convergence zones are connected to other, lower, convergence zones. The lowest levels of the hierarchy would be pools of neurons that are not themselves convergence zones but supply information that would get bound in convergence zones. Convergence zones account for inference insofar as the lateral and top down connections allow for the endogenous triggering of conceptual representations: one thinks of mammals because one was previously thinking of dogs as opposed to thinking of mammals because an actual mammal triggered the perceptual application of the concept.

Convergence zones do not only provide models of inference, but a more flexible kind of recognition one might associate with genuinely conceptual systems. Cohen and Eichenbaum (1993), hypothesize that hippocampus is a locus of convergence zones. The flexible kind of recognition is illustrated as follows. Eichenbaum et al (1989) showed that rats with hippocampal lesions can be trained to prefer certain stimuli, but the preferences will not be exhibited in contexts that differ significantly from the initial training conditions. The rats demonstrate their conditioned preference by moving toward the target odor. A hippocampus-damaged rat, when presented with two odor sources, *A* and *B*, can be trained to prefer *A* to *B*. This rat will demonstrate this preference even after a sustained delay period. However, if the rat is presented with the preferred odor *A*, along with some odor *N*, that differs from the non-preferred odor *B* that accompanied *A* in the learning trials, the rat will demonstrate no preference whatsoever. Healthy rats do not exhibit such a lack of preference—their preference is demonstrated even in novel situations such as the presentation of *A* with *N*.

Given the convergence zone model, we can turn to casting doubt on the phenomenal concepts proposal: why couldn't top down or lateral connections suffice for concept acquisition? Whatever set of changes in connections in a network suffice for the acquisition of a concept, in particular, the concept involved in knowing what it is like to see red things, why couldn't those changes be brought about by means other than the perceptual presentation of red things or even activations of the neural areas constitutive of having an experiences as if one were seeing a red thing?

Whatever knowing what is like consists in, it is the down-stream effects of having certain experiences. Why couldn't the resulting structures be installed without having such causes? Such a situation is at least possible in theory. Dennett (2005) illustrates the possibility in terms of the fanciful thought experiment of "Swamp Mary":

> Just as standard Mary is about to be released from prison, still virginal with regard to colors…, a bolt of lighting rearranges her brain, putting it by Cosmic Coincidence into exactly the brain state she was just about to go into *after* first seeing a red rose. (She is left otherwise unharmed of course; this is a thought experiment.) So when, a few seconds later, she is released, and sees for the first time, a colored thing (that red rose), she says just what she would say on seeing her *second* or *nth* red rose. "Oh yeah, right, a red rose. Been there, done that" (p. 120).

What I am suggesting here is less fanciful than Swamp Mary. What I am suggesting is the possibility that concepts in low-level convergence zones can be instilled by having the requisite connection weights modified via the influence of lateral influence from other low-level convergence zones and top-down influence from high-level convergence zones. If there were a technique whereby the non-feed-forward installation of a concept can be achieved, then physically omniscient Mary would *know* that technique. It of course is open whether she would thereby be able to employ the technique. The phenomenal concepts proposal constitutes a claim to the effect that such a thing would be impossible. That such a thing would not be possible is a bold *empirical* conjecture about the way humans are wired, at least insofar as we are looking to the structure of concepts to supply a possible explanation of the alleged subjectivity of phenomenal character. But perhaps a different kind of case can be made, namely, if we look to the conditions that determine the representational contents of concepts, we will

see what makes phenomenal concepts the only concepts capable of representing phenomenal character. In the next section I turn to consider some content-based considerations regarding whether phenomenal character is subjective.

§ 4. The Content of Concepts Defense of Subjectivity

There are two ways in attempting to spell out how an account of representational content can secure the subjectivity of phenomenal character. The first involves the contents of perception. The second involves the contents of introspection.

In Mandik (2001) I attempted to give a neurophilosophical account of subjectivity whereby the reason certain phenomenal characters are one-way knowable is because those phenomenal characters are representational contents of perception such that what is represented can only be represented by those perceptual states. Assuming that perceptual representations have their contents due to certain kinds of causal interactions that allow them to function as detectors of (typically) environmental properties, I suggested that there may be a certain environmental property such that there is only one "detection supporting causal interaction" that the property can enter into, and further, the unique detector is the perceptual representation itself. I discussed how such a view of uniquely representable properties might be based in an understanding of egocentric representations. Egocentric representations include those mentioned previously in connection with the lowest levels of the visual processing hierarchy, such as LGN and V1 neural activations that represent stimuli in retinocentric spatial locations. I argued that the notion of egocentric representation generalizes to non-spatial examples, and developed, in particular, an account of the egocentricity of temperature representations in thermoperception. (2001, pp 194-196). As Akins (2001) has pointed out, the outputs of thermoreceptors to not simply reflect the triggering temperatures, but also what part of the body the temperature is being applied to and what temperature was previously at that portion of the body. While Akins argues that such states are not representations, I argued that they are representations, just not representations of subject-independent properties. As such, what is detected is not a given temperature but instead

> …whether the given temperature is, for example, too hot, too cold, or just right. The property of being too hot cannot be defined independently of answering the question "too hot for whom?"…(Mandik 2001, p. 195)

What it means for all of these representations to be egocentric is that they don't simply represent things; they represent those things in relation to the representing subject. I argued that such representations have contents that are one-way representable. I attempted to illustrate this feature in terms of imagistic or pictorial representations, wherein part of what is represented is what something looks like from some literal point of view.

> Consider a pictorial representation, such as a photograph of a complex object like the Statue of Liberty. Given the point of view from which the photograph was taken, only a fraction of the surface of the statue is explicitly represented in the photograph. Certain regions seen from one angle would be obscured from another angle. Consider the set of regions that are captured by the photograph: the set of all and only the regions explicitly represented in the photograph.
> […]

> [I]n specifying the set comprised of all and only the spatial regions captured in the image, one does not carve nature at the joints, but instead carves nature into a gerrymandered collection of items that would be of no interest apart from their involvement in a particular representation. That much of neural representation is concerned with such gerrymandered properties should not come as an enormous surprise. For instance, it makes sense that an animal's chemoreceptors would be less interested in carving nature into the periodic table of elements and more interested in carving nature into the nutrients and the poisons—categories that make no sense apart from the needs of a particular type of organism. (p. 197)

Another way I illustrated the idea of one-way representable properties was by reference to an example of Dennett's concerning the torn halves of a Jell-O box a pair of spies used as unique un-forgeable identification. As I described the case, "the only *detection supporting* causal interactions that one piece of the Jell-O box enters into are causal interactions involving the other piece of the box" (Mandik 2001, p. 198). As Dennett (1997) describes the case,

> [t]he only readily available way of saying what property M is [the property that makes one piece of the Jell-O box a match for the other] is just to point to our M detector and say that M is the shape property detected by this thing here. (p. 637)

This completes the sketch of the content-based defense of the subjectivity of phenomenal character based on egocentric representations. However, I've grown to have misgivings about whether egocentric representations can serve as a basis for genuine subjectivity. To see this, let us focus on the Jell-O box example and what an introspective analog would be. If I had a torn Jell-O box half, one way I can identify it as such is to match it to its torn partner. But subsequent to that I can stick it in my pocket and travel far away. Later in my journey I reach into my pocket, fish around for it, and easily find it. I can do so even if it is not the only thing in my pocket. There may be a coin and a tube of Chapstick in there as well. But I am able to identify the torn Jell-O box half without literally retracing my steps and tracking down its partner. It is strictly false, then, that the only way I can identify that torn Jell-O box is by the matching process. There are multiple ways in which I can represent that half. I can represent it as the thing that matches the thing in my pocket. One might think that there is an essentially demonstrative element involved here, that the box half is essentially represented in virtue of there being some point at which I was able to point to it and say it is that thing detected by this thing. But this is not obviously correct. It seems open for some one who never was in a position to demonstratively refer to the Jell-O box half to refer to it by description, like "the largest thing that Mandik had in his pocket last Wednesday."

So, even if there were environmental properties that only entered into detection-supporting causal interactions with certain sensory states, it seems dubious that those sensory states constitute the *only* representations of those properties. Suppose Jones has such sensory states and Smith does not. Even if Jones' sensory states are the unique detectors of those environmental properties, Smith can still represent those properties. Smith can represent them under a description such as "the environmental properties uniquely detected by Jones' sensory states".

The availability of this kind of move does not seem to depend on any kind of story about representational content. Schier (ms) has criticized the Mandik (2001) account of subjectivity and offered as superior a substitute that differs primarily in

relying on an isomorphism-based instead of causal-covariance-based theory of representational content. It's not clear, though, that replacing the account of representational content is going to protect the account from the charge that Smith can represent the content of Jones' sensory states as "the environmental properties uniquely detected by Jones' sensory states."

Is this kind of move cheating? If a picture is indeed worth a thousand words, then isn't it cheating to say that whatever a picture *P* represents, the same thing can be represented by the description "whatever *P* represents"? Even if the "whatever *P* represents" move is somehow disqualified, the following move remains open: just add words to the description. Why couldn't a sufficiently long description represent all of the same things without itself being a picture? Analogously, why couldn't a conceptual representation of what it is like to have an experience represent all of the things that the experience does without itself being the experience or the consequence of having had the experience? I want to continue to address such questions but by shifting focus slightly. I want to turn from the question of whether perception uniquely represents environmental properties to whether introspection uniquely represents certain mental properties.

At the heart of peoples' intuitions about the Mary thought experiment seems to be an intuition that with respect to at least some phenomenal characters, especially those connected to the visual perception of colors, having is knowing. I think that this intuition doesn't stand up under pressure and it is worth considering some of the neurophilosophical pressure that can be applied to it. First let us begin with a neurophilosophical proposal from Paul Churchland concerning what qualia are. Beginning with a focus on color qualia, Churchland identifies them with certain properties in a neural network modeled in accordance with Land's theory of color vision. According to Land, human color vision is reflectance discrimination, which is accomplished via the reception of three kinds of electromagnetic wavelengths by three different kinds of cones in the retina. In terms of a neural network and state-space representations, Churchland identifies color qualia with points of a three-dimensional state-space wherein the three dimensions are defined by the three kinds of cells and their preferred ranges of electro-magnetic wavelengths. Churchland identifies color sensations with neural representations of colors, that is, with neural representations of spectral reflectance. Sensations are points in three-dimensional color space and perceived similarity between colors is mirrored by proximity in this neural activation space. Churchland generalizes the account to include qualia from other sensory modalities. Emphasizing how the account allows for a representation of qualia aside from simply having them, he 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 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).

In later work, Churchland emphasizes how such an identity theory allows one to predict falsifiable data about consciousness. The data in question is not just third-person accessible data. Paul Churchland makes an excellent case for this in his recent "Chimerical Colors: Some Novel Predictions from Cognitive Neuroscience" (2005) in which very odd color experiences are predicted by a neural model of chromatic information processing. In brief, the differential fatiguing and recovery of opponent processing cells gives rise to afterimages with subjective hues and saturations that would never be seen on the reflective surfaces of objects. Such "chimerical colors" include shades of yellow exactly as dark as pitch-black and "hyperbolic orange, an orange that is more 'ostentatiously orange' than any (non-self-luminous) orange you have ever seen, or ever will see, as the objective color of a physical object" (p. 328). Such odd experiences are predicted by the model that identifies color experiences with states of neural activation in a chromatic processing network. Of course, it's always open to a dualist to make an *ad hoc* addition of such experiences to their theory, but no dualistic theory ever predicted them. Further, the sorts of considerations typically relied on to support dualism—appeals to intuitive plausibility and a priori possibility—would have, you'd expect, ruled them out.

Who would have, prior to familiarity with the neural theory, predicted experiences of a yellow as dark as black? One person who would not have thought there was such an experience as pitch-dark yellow is Ludwig Wittgenstein, who once asked "[W]hy is there no such thing as blackish yellow?" (1978, p. 106). I think it safe to say Wittgenstein would be surprised by Churchland's chimerical colors. At least, I know I was, and I literally grew up reading Churchland. However, to be certain that we have an example of someone who is surprised—for I would like to conduct a thought experiment about them—let us consider someone, call him "Larry", who has seen yellow and black and in general all the typical colors a normally sighted adult has seen. Suppose that Larry has never had a chimerically colored afterimage such as hyperbolic orange or pitch-dark yellow. Suppose further that Larry is aware of none of the neuroscience that predicts the existence of such experiences. Now, let us compare Larry to Hyperbolic Mary. Like Jackson's Mary, Hyperbolic Mary knows all of the physical facts about how human color vision works, including the predictions of chimerically colored afterimages. Suppose also, that like Mary toward the end of Jackson's story, Hyperbolic Mary has been let out of the black and white room and has seen tomatoes, lemons, grass, cloudless skies, and the like. In short, she has had the average run of basic color experiences. Let us stipulate that she has had all the types of color experiences that Larry has had. The crucial similarity between Mary and Larry is that not only have they seen all of the same colors, neither has had chimerically colored afterimages. Neither has experienced hyperbolic orange or pitch-dark yellow. The crucial difference between Larry and Hyperbolic Mary is that only Hyperbolic Mary is in possession of a theory that predicts the existence of hyperbolic orange and pitch-dark yellow. And here's the crucial question: who will be more surprised upon experiencing chimerical colors for the first time, Larry or

Hyperbolic Mary? I think it's obvious that Larry will be more surprised. I also think this has pretty significant implications for what we are to think the knowledge of what it is like consists in. One thing that knowing what it is like consists in is something that will determine whether one is surprised or not. Fans of Jackson's Mary must grant this, for they are fond of explicating Jackson's Mary's ignorance of what it is like in terms of her alleged surprise at seeing red for the first time. Well, Hyperbolic Mary is *less* surprised than Larry on seeing chimerical colors for the first time. This shows that she must have more phenomenal knowledge—more knowledge of what it is like to have certain experiences—than did Larry. Mary was able to represent, in introspection, more properties of her experiences than Larry. And her introspective capacity was augmented by her neuroscientific concepts.

Does this example suffice to show that *all* knowledge of what it is like can be had without the requisite experiences? No, it does not. But it does help to show—especially in concert with the other considerations concerning the structure and content of concepts—just how beholden to empirical considerations the subjectivity intuition is. In order for the subjectivity intuition to be true, quite a bit that is actually up in the air concerning the neural bases of concepts would have to be true. If, contrary to the considerations I've offered here, that stuff does turn out to be true, no one should believe it until further arguments are given. We should be skeptical of claims that the subjectivity of phenomenal character is known by *intuition*. If it is to be known at all, it is to be known via neurophilosophy, and the neurophilosophical considerations weigh more heavily against subjectivity than in favor of it.

Another point that the tale of Hyperbolic Mary and Larry helps to bring out is my concluding point: If it is unreasonable to expect Larry to predict the possibility of hyperbolic orange, pitch-dark yellow, and the like, then it seems unreasonable to predict, on introspection and intuition alone, the *impossibility* of pre-experiential knowledge of what it is like to see red. It is unreasonable, then, to think introspection and intuition suffices for establishing the subjectivity of consciousness.

Acknowledgements
For helpful discussion of this and related material I thank Torin Alter, John Bickle, David Chalmers, Nick Treanor, Chase Wrenn, Tanasije Gjorgoski, Eric Schwitzgebel, and Tad Zawidzki.

References

Akins, K. (2001). Of sensory systems and the 'aboutness' of mental states. In W. Bechtel, P. Mandik, J. Mundale, & R. Stufflebeam (Eds.) *Philosophy and the Neurosciences: A Reader*. Oxford: Blackwell.

Beaton, M. J. S. (2005). What RoboDennett Still Doesn't Know. *Journal of Consciousness Studies*, 12(12), pp.3-25.

Bickle, John, Mandik, Peter, Landreth, Anthony, (2006). The Philosophy of Neuroscience. The *Stanford Encyclopedia of Philosophy* (Spring 2006 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2006/entries/neuroscience/>.

Churchland, P. (1979). *Scientific Realism and the Plasticity of Mind*, Cambridge University Press, Cambridge.

Churchland, P. (1989). *A Neurocomputational Perspective*. Cambridge, MA: MIT Press.
Churchland, Paul. 2005. Chimerical Colors: Some Novel Predictions from Cognitive Neuroscience. In: Brook, Andrew and Akins, Kathleen (eds.) *Cognition and the Brain: The Philosophy and Neuroscience Movement*. Cambridge: Cambridge University Press.
Cohen, N. and Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT press.
Dennett, D. (1997). Quining qualia. In N. Block, O. Flanagan, & G. Güzeldere (Eds.) *The Nature of Consciousness.* Cambridge, MA: MIT Press.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA, MIT Press.
Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly,* 32, 127–136.
Lamme, V. A. F., et al. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8, 529 – 535.
Lewis, D. (1983): Postscript to 'Mad Pain and Martian Pain', in his *Philosophical Papers*, vol. 1, pp. 13–32. New York: Oxford University Press.
Lewis, D. (1990): 'What Experience Teaches', in *Proceedings of the Russellian Society*. Sydney: University of Sydney, 1988. Rpt. in W. Lycan (ed.) *Mind and Cognition*, pp. 499–518. Cambridge: Basil Blackwell.
Mandik, P. (2001). Mental representation and the Subjectivity of Consciousness. *Philosophical Psychology* 14(2): 179-202.
Mandik, P. (2005). Phenomenal Consciousness and the Allocentric-Egocentric Interface in R. Buccheri et al. (eds.); *Endophysics, Time, Quantum and the Subjective*. World Scientific Publishing Co.
Mandik, P. (2006a). The Introspectability of Brain States as Such. In Brian Keeley, (ed.) *Paul M. Churchland: Contemporary Philosophy in Focus*. Cambridge: Cambridge University Press.
Mandik, P. (2006b). The Neurophilosophy of Consciousness. In Max Velmans and Susan Schneider (eds.) *The Blackwell Companion to Consciousness*. Oxford: Basil Blackwell.
Mandik, P. (in press). An Epistemological Theory of Consciousness?. In Alessio Plebe, ed. *Philosophy in the Neuroscience Era, Special issue of the Journal of the Department of Cognitive Science, Univ. of Messina*.
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435–450.
Nemirow, L. (1980): Review of Mortal Questions, by Thomas Nagel. *Philosophical Review* 89, 473–477.
Nemirow, L. (1990): 'Physicalism and the Cognitive Role of Acquaintance', in W. Lycan (ed.) *Mind and Cognition*, pp. 490–499. Cambridge: Basil Blackwell.
Papineau, D. (1999). "Mind the Gap" *Philosophical Perspectives*
Pascual-Leone, A. and Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness, *Science*, 292, 510–512.
Schier, E. (ms). Representation and the knowledge intuition.
Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.
Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA: MIT Press.
von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science,* 224, 1260–1262.

Wittgenstein, L. 1978. Some Remarks on color, ed. G.E.M. Anscombe, Oxford: Blackwell.
Zeki, S. (1983). Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelengths and colour. *Neuroscience*, 9, 741–756.