**Higher-Order Thoughts, Neural Realization, and the Metaphysics of Consciousness**

**Rocco J. Gennaro, University of Southern Indiana**

[The final version is published in *Consciousness: Integrating Eastern and Western Perspectives*, Prem Satsangi, Stuart Hameroff, Vishal Sahni, and Pami Dua (eds.), New Age Publishers (India), 2016, pp. 83-102.]

The higher-order thought (HOT) theory of consciousness is a reductive representational theory of consciousness which says that what makes a mental state conscious is that there is a suitable HOT directed at that mental state. Although it seems that any neural realization of the theory must be somewhat widely distributed in the brain, it remains unclear just how widely distributed it needs to be. In section I, I provide some background and define some key terms. In section II, I argue *against* the view that HOT theory should treat first-order (i.e. world-directed) conscious states as requiring prefrontal cortical activity though it is reasonable to suppose that conscious states are realized in the brain. In section III, I then explore some of the key background metaphysical issues involved in understanding the nature of consciousness, such as the debate between realism and idealism as well as the prospects for solving the so-called “hard problem” of consciousness. Some of the differences in question often mirror the traditional differences between Western and Eastern perspectives on the nature of consciousness. Overall, I argue that some form of realism and physicalism is more plausible than the opposing views. I also argue that materialists (and especially HOT theorists) can offer plausible replies to the hard problem.

**I. Background and Terminology**

Perhaps the most fundamental and commonly used notion of the term ‘conscious’ is captured by Thomas Nagel’s famous “what it is like” sense (Nagel 1974). When we are in a conscious mental state, there is “something it is like” for us to be in that state from the subjective or first-person point of view. When we smell a rose or have a conscious visual experience, there is something it “seems” or “feels like” from our own perspectives. An organism such as a bat is conscious if it is able to experience the world through its echolocation senses. There is something it is like to be a conscious creature, whereas there is nothing it is like to be a table or tree. This is primarily the sense of “conscious state” that I use throughout this chapter. “What it’s like” basically means “how a conscious state is for the subject.”

 The HOT theory of consciousness says that what makes a mental state conscious is that there is a suitable HOT directed at the mental state (Rosenthal 1997, 2002, 2005, Gennaro 1996, 2012). Higher-order thoughts (HOTs) are meta-psychological states, that is, mental states directed at other mental states. HOT theory is primarily concerned with explaining how conscious mental states differ from unconscious mental states. For various reasons, HOT theorists hold that such meta-awareness is best understood as a “thought” composed of concepts (Gennaro 2012). Further, it seems reasonable to think that conscious mental states are states that we are “aware of” in some sense. Indeed, defense of HOT theory often tends to start with the highly intuitive claim that has come to be known as the Transitivity Principle (TP):

(TP) A conscious state is a state whose subject is, in some way, aware of being in it (such as in Rosenthal 2005).

Thus, when one has a conscious state, one is aware of being in that state. For example, if someone is having a conscious desire or pain, then that person is surely aware of having that desire or pain. HOT theory says that the HOT is of the form “I am in M now,” where M references a mental state. Conversely, the idea that someone could be having a conscious state while totally unaware of being in that state seems very odd (if not an outright contradiction). A mental state of which the subject is completely unaware is clearly an unconscious state. For example, we are not aware of having a subliminal perception, and thus it is an unconscious perception.

It is also crucial to mention here that according to HOT theory, when a conscious mental state is a first-order world-directed state, the HOT is *not* itself conscious; otherwise circularity and an infinite regress would follow. When the HOT is itself conscious, there is a yet higher-order (or third-order) thought directed at the second-order state. In this case, we have *introspection*, which involves a conscious HOT directed at an inner mental state. When one introspects, one’s attention is directed back into one’s mind. For example, what makes my desire to write a good chapter a conscious first-order desire is that an unconscious HOT is directed at the desire. In this case, our conscious focus is directed at the book and my computer screen, so we are not consciously aware of having a HOT from the first-person point of view. When one introspects that desire, however, one then has a conscious HOT (accompanied by a yet higher, third-order, HOT) directed at the desire itself (Rosenthal 1986, 1997).

[Insert figure 1 here]

**II. HOT Theory, Neural Realization, and The Prefrontal Cortex**

There are of course also several common objections to HOT theory to which HOT theorists have replied. For example, many have urged that HOT theory rules out infant and animal consciousness. Since a HOT is a kind of metacognitive or meta-psychological state of the form “I am in mental state M now,” it might seem that animals (or at least *most* animals) do not possess such sophisticated I-concepts and mental concepts. I have responded at length to this objection (such as Gennaro 2012, chapters 7 and 8) but it is not my main focus in this paper.

Still, HOT theory is a (mentalistic) reductionist theory in the sense that conscious states are explained in terms of unconscious mental states and awareness. Nonetheless, as Rosenthal urges, HOT theory is poised to explore a “second step” reduction to the physical level. The most natural candidate for physical realization is in the brain, but I have previously argued that HOTs need not occur in the prefrontal cortex (PFC) (Gennaro 2012, chapter 9). Here I summarize and briefly elaborate on this view. In this context also, some have urged that HOTs are likely to occur in the prefrontal cortex (PFC) and, since infants and most animals do not have much (if any) prefrontal cortical activity, they are incapable of having conscious states according to HOT theory. However, a strong case can be made that HOTs, or at least many unconscious HOTs, do not involve the PFC. Other areas of the brain are more likely to underlie HOTs, such as the medial and inferior parietal cortices, the temporoparietal cortex, the posterior cingulate cortex, and the anterior cingulate cortex (Newen and Vogeley 2003, Gennaro 2012, chapter 9). Even when considering the neural signatures of “theory of mind” and “mindreading” (i.e. thinking about others’ thoughts), Newen and Vogeley (2003) cite numerous experiments indicating that such metarepresentation is best located in the anterior cingulate cortex and that “the capacity for…[theory of mind] contexts showed differential activation in the right temporo-parietal junction and the medial aspects of the superior parietal lobe” (Newen and Vogeley 2003, p. 538). Some evidence for holding that the PFC is not required for conscious states is that basic conscious experience is not decreased entirely even when there is extensive bilateral PFC damage or lobotomies.

In addition, Malach and colleagues show that when subjects are engaged in a perceptual task or absorbed in watching a movie, there is widespread neural activation but little PFC activity (Grill-Spector and Malach 2004, Goldberg, Harel, and Malach 2006). Although some other studies do show PFC activation, this is arguably due to the fact that subjects are required to *report* their experiences. Thus, the PFC is likely to be activated when there is *introspection*, not merely when there are outer-directed conscious states (especially during demanding perceptual tasks). But these introspective states involve more sophisticated psychological capacities than merely having conscious states, and introspection is not necessary for having conscious states according to HOT theory. Further support for the above can be found, for example, in the work of Saxe and colleagues who have extensively studied the brain regions most associated with mindreading abilities (Saxe and Powell 2006, Saxe and Pelphrey 2009, Saxe 2010). They identify the temporo-parietal junction, the posterior cingulate, the medial precuneus, and parts of the temporal sulcus as the primary sites for this kind of cognition.

Kuzuch (2014) presents a very nice discussion of the PFC in relation to higher-order theories, arguing that the lack of dramatic deficits in visual consciousness even with PFC lesions presents a compelling case against higher-order theories. In some ways, I agree with much of Kozuch’s analysis especially with respect to the notion that some (visual) conscious states do not require PFC activity (sometimes focused more on the dorsolateral PFC, or dlPFC). For example, in addition to the studies I cited above, Kozuch references Alvarez and Emory (2006) as evidence for the view that

Lesions to the orbital, lateral, or medial PFC produce so-called executive dysfunction. Depending on the precise lesion location, subjects with damage to one of these areas have problems inhibiting inappropriate actions, switching efficiently from task to task, or retaining items in short-term memory. However, lesions to these areas appear not to produce notable deficits in visual consciousness: Tests of the perceptual abilities of subjects with lesions to the PFC proper reveal no such deficits; as well, PFC patients never report their visual experience to have changed in some remarkable way (Kozuch 2014, 729).

However, Kozuch rightly notes that my version of HOT theory is left undamaged, at least to some extent, since I do not require that the PFC is where HOTs are realized. We must again also keep in mind the distinction between unconscious HOTs and conscious HOTs (= introspection). Perhaps the latter require PFC activity given the more sophisticated executive functions associated with introspection but having first-order conscious states does not require introspection.

**III. The Metaphysics of Consciousness**

 Of course, the above discussion does not explicitly address the larger metaphysical issues lurking in the background (though see Gennaro 2005, Gennaro 2012, chapters 2 and 4, for much more along these lines). For example, dualists will of course deny that mental states are ultimately neural or physical processes at all. Others might urge that materialism suffers from some kind of insurmountable “explanatory gap” between the mental and physical or cannot answer the so-called “hard problem” of consciousness (Chalmers 1995, Levine 1983, 2001). Some might even reject the notion that there really are mind-independent material or physical things at all and so will embrace some kind of idealism such that consciousness (or conscious minds) is the only real and fundamental substance (or force) in nature. I now address these questions in turn with an eye toward some of the traditional differences between Eastern and Western views.

1. **Materialism vs. Dualism**

 There are two broad traditional and competing metaphysical views concerning the nature of the mind and conscious mental states: dualism and materialism. While there are many versions of each, the former generally holds that the conscious mind or a conscious mental state is non-physical in some sense. On the other hand, materialists hold that the mind is the brain, or, more accurately, that conscious mental activity is identical with neural activity. It is very important to recognize that by ‘non-physical,’ dualists do not merely mean ‘not visible to the naked eye.’ Many physical things fit this description, such as the atoms in the air in a typical room. For something to be non-physical, however, it must literally be outside the realm of physics, that is, not in space at all and so undetectable in principle by the instruments of physics. It is equally important to recognize that the category “physical” is broader than the category “material.” Materialists are called such because there is the tendency to view the brain, a material thing, as the most likely candidate to identify with the mind. However, something might be physical but not material in this sense, such as an electromagnetic or energy field. One might therefore instead be a “physicalist” about mind or consciousness in some broader sense and still not be a dualist in the above sense. Thus, to say that the mind is non-physical is to say something much stronger than that it is non-material. Dualists, then, tend to believe that conscious mental states or minds are radically different from anything in the physical world at all. This was especially the case for Descartes and his version of “substance dualism” but also seems to hold for some dualists today (e.g. Swinburne 1986) and is often tied to various theological positions.

 Nonetheless, there are very well-known and serious problems with substance dualism. For example, one is simply the issue of just how does (or could) such radically different substances causally interact. How does anything non-physical causally interact with something physical, such as the brain? No such explanation is forthcoming or even seems possible. Yet substance dualists do typically hold that mental states causally interact with the body, as in familiar cases such as my desire for something cold to drink causes my body to move toward the refrigerator for some water. Moreover, if causation involves some kind of transfer of energy from cause to effect, then how is this possible if the mind really is non-physical? This is in part why Ryle (1949) mockingly called the Cartesian view a belief in the “ghost in the machine.” There are, to be sure, other versions of dualism such as “property dualism” which is probably the most popular form of dualism today. Property dualism is, at least on the surface, a more modest version of dualism which says that there are mental *properties* (i.e. characteristics or aspects of things) that are neither identical with nor reducible to physical properties. Mental phenomena are non-physical properties of physical substances. There are actually several different kinds of property dualism but what they have in common is the idea that conscious *properties*, such as the color qualia involved in a conscious experience of a visual perception, cannot be explained in purely physical terms and, thus, are not themselves to be identified with any brain state or process. But, for one thing, it is not always clear just how property dualism should be interpreted in terms of its fundamental metaphysics, for example, isn’t it also committed to substance dualism? (Francescotti 2001). Alternatively, some also wonder how property dualism actually differs from some kind of non-reductive *materialism*.

On the other hand, epiphenomenalism holds that mental events are caused by brain events but those mental events are mere “epiphenomena” which do not, in turn, cause anything physical at all, despite appearances to the contrary. But problems for epiphenomenalism abound. First, it is incredibly counterintuitive. What could be more obvious than my pain makes me cry or that the visual experience of the charging lion makes me run away? Second, if conscious mental states really do and cause nothing, then there is no reason why they should have evolved. If conscious states clearly modify our behavior in certain ways, then they are very useful from an evolutionary perspective. (Nonetheless, see Robinson 2004 for a recent defense of epiphenomenalism.)

 There are also theoretical factors which seem to be on the side of materialism, such as adherence to the so-called “principle of simplicity” which says that if two theories can equally explain a given phenomenon, then we should accept the one which posits fewer objects or forces. In this case, even if dualism could equally explain consciousness (which would of course be disputed by materialists), materialism is clearly the simpler theory in so far as it does not posit any objects, processes, or properties over and above physical ones. Materialists will wonder why there is a need to believe in the existence of such mysterious non-physical entities or properties. Moreover, in the aftermath of the Darwinian revolution, it would seem that materialism is on even stronger ground provided that one accepts basic evolutionary theory and the notion that most animals are conscious. Given the similarities between the more primitive parts of the human brain and the brains of other animals, it seems most natural to conclude that, through evolution, increasing layers of brain areas correspond to increased mental abilities. It does not seem that many dualists think that animals have non-physical minds and yet animals surely have *some* similar conscious states to us, such as basic sensory perceptions and various feelings and emotions, such as pains and anger.

 Many will also point to the well-known fact that damage to specific areas of the brain causes very specific mental defects as evidence that materialism is true, or at least as evidence supporting what we might call the “dependence thesis,” that is, conscious experience *depends* upon proper brain function. Although the implications of this evidence have probably been appreciated for centuries (would you rather be shot in the head or foot!?), the level of detailed neuropsychological knowledge has increased dramatically in recent years. A substance dualist might of course respond that such phenomena do not absolutely refute her metaphysical position since it could be replied that damage to the brain simply *causes* corresponding damage to the non-physical mind. However, this raises a host of other equally difficult questions, such as: Why not opt for the simpler and more straightforward explanation that brain damage causes mental damage simply because mental processes *are* brain processes? How can a non-physical mind be “damaged” anyway when brain damage occurs?

 Moreover, the level of detail understood today is important because it makes clear that very specific mental changes occur when, and only when, very specific brain damage occurs. It is true that such a “correlation” is not the same as identity or cause, but the best explanation for the neuropsychological evidence is clearly that conscious mental activity depends for its existence upon the relevant brain activity. The details are exactly what one would expect if conscious mental activity so depended upon brain functioning. In humans, damage to particular brain regions, such as due to disease, trauma, or stroke, is associated with specific impairments of perception, memory, cognition, emotion, and decision-making. Drugs that alter brain activity produce corresponding changes in perception, memory, cognition, emotion, or personality, depending upon the neurotransmitter systems involved and particular brain regions affected. Loss of vision results from damage to areas of visual cortex. Loss of hearing and/or the ability to recognize sounds, including speech, results from damage to regions of auditory cortex within the temporal lobes. Prosopagnosia, the inability to recognize familiar faces, is typically caused by damage to the occipitotemporal cortex (fusiform gyrus). Damage to brain regions involved in emotional regulation, which include the limbic system, particularly the amygdala, commonly result in impaired processing of emotional stimuli. Furthermore, numerous neurological disorders and diseases, such as Alzheimer’s, bipolar disorder, amnesia, depression, schizophrenia, epilepsy, and mental retardation, which are all characterized by profound changes in cognitive function and awareness, are also all associated with specific biochemical, neurophysiological, or neuroanatomical changes in the brain. While some of the specifics are still not fully understood and the subject of vigorous ongoing scientific investigation, what is not much in dispute is that these mental disorders and diseases are direct consequences of aberrant brain function.

 Thus, it seems reasonable to hold that some form of materialism is true and that mental states are realized in neural processes.

1. **The Hard Problem**

One might of course still object that materialism faces serious problems of its own. To be sure, it would be silly to think that materialists have all the answers. One well-known objection can be found in Chalmers (1995) who presents the “hard problem” of consciousness as follows:

The really hard problem is the problem of experience....a subjective aspect....It is undeniable that some organisms are subjects of experience. But the question of how it is that [organisms] are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have a visual or auditory experience...? ...How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises (1995, p. 201).

Chalmers himself eventually opts for a form of property dualism. In a somewhat similar vein, Levine (1983, 2001) coined the expression “the explanatory gap” to express a difficulty for any materialistic attempt to explain consciousness. Although not himself concerned to reject the metaphysics of materialism, Levine gives eloquent expression to the idea that there is a key conceptual gap in our ability to explain the connection between conscious mental (or “phenomenal”) properties and brain properties. The basic problem is that it is, at least at present, very difficult for us to understand the relationship between brain properties and phenomenal properties in any explanatory satisfying way, especially given the fact that it seems possible for one to be present without the other. There is an odd kind of arbitrariness involved: *Why* or *how* does some particular brain process produce that particular taste or visual sensation? It is difficult to see any real *explanatory connection* between specific conscious states and brain states in a way that explains how or why the former are identical with the latter. There is therefore an explanatory gap between the physical and mental.

But, first of all, it may be that Chalmers and others expect too much, namely, grasping some “causal link” between the brain and consciousness. After all, if conscious mental states just *are* ultimately *identical* to brain states, then there may simply be a “brute fact” that does not need any further explaining. It may even be that some form of dualism is presupposed in Chalmers’ argument, especially to the extent that brain states are assumed to “give rise to” or “produce” (or cause) consciousness, as opposed to using the language of identity.

Secondly, and perhaps more important, is recognition of the fact that very different concepts can still pick out (or refer to) the same property or object in the world (Loar 1997). For example, out in the world there is only the one “stuff,” which we can conceptualize either as ‘water’ or as ‘H2O.’ The traditional distinction, made most notably by Frege in the late 19th century, between “meaning” (or “sense”) and “reference” is relevant here. Two or more concepts, which can have different meanings, can refer to the same property or object, much like ‘Venus’ and ‘The Morning Star.’ Materialists, then, explain that it is essential to distinguish between mental properties and our concepts of those properties. By analogy, there are so-called “phenomenal concepts” which use a phenomenal or first-person property to refer to a conscious mental state, such as a sensation of red (see Alter and Walter 2007 for much more on this theme). In contrast, we can also use various concepts couched in physical or neurophysiological terms to refer to that same mental state from the third-person point of view. There is thus but one conscious mental state which can be conceptualized in two different ways: either by employing first-person experiential phenomenal concepts or by employing third-person neurophysiological concepts. It may then just be a brute fact about the world that there are such identities and the appearance of arbitrariness between brain properties and mental properties is just that – an *apparent* problem leading many to wonder about the alleged explanatory gap. Qualia would still be identical to physical properties. Moreover, this response provides a diagnosis for why there even *seems* to be such a gap, namely, that we use very different concepts to pick out the same property. Science will be able, in principle, to close the gap and solve the hard problem of consciousness in an analogous way that we now have a very good understanding for why “water *is* H2O” or “heat *is* mean molecular kinetic energy” that was lacking centuries ago. Maybe the hard problem isn’t so hard after all – it will just take some more time. After all, the science of chemistry didn’t develop overnight and we are relatively early in the history of neurophysiology and in our understanding of phenomenal consciousness. It seems premature to give up now.

 Finally, it is crucial to emphasize for my purposes that HOT theory is reductionist only in the sense that state consciousness is to be explained in terms of unconscious mental states and awareness. HOT theory itself does not even attempt to explain consciousness in nonmentalistic or neural terms. Of course, as we have seen, HOT theorists tend to be materialists in the end but prefer to leave that empirical question for a separate second step reduction to be filled in later by brain science. On the other hand, other theories of consciousness are reductionist in the stronger sense that they attempt to explain consciousness directly in physical or neurophysiological terms (e.g. Crick and Koch 1990, Crick 1994).

I believe that HOT theory can provide necessary and sufficient conditions for what makes a mental state conscious but whatever realizes that theory in our brains is a separate empirical question. The quotation above from Chalmers does show, however, how statements of the hard problem can sometimes gloss over these differences. Nonetheless, the key common general question is: How exactly does x (where ‘x’ is invoked by some alleged theory of consciousness) explain how consciousness arises from the presence of x?

A HOT theory solution is that HOTs explain how consciousness arises because the concepts that figure into the HOTs are presupposed in conscious experience. Let us stick to first- order perceptual states. In very much a Kantian spirit, the idea is that we first passively receive information via our senses or what Kant (1781/1965) calls our “faculty of sensibility,” which we might think of as early perceptual processing. Some of this information will then rise to the level of unconscious mental states which can also cause our behavior in various ways. But such mental states do not become conscious until the “faculty of understanding” operates on them via the application of concepts. I contend that we should understand such concept application in terms of HOTs directed at the incoming information. Thus I consciously experience the green table as a green table partly because I apply the concepts GREEN and TABLE (in my HOTs) to the incoming information via my visual perceptual apparatus. More specifically, I have a HOT such as “I am seeing a green table now.” Kant (1781/1965) famously urged that it takes the cooperation of both the sensibility and understanding to produce conscious experience: “Objects are given to us by means of sensibility. . . . They are thought through the understanding, and from the understanding arise concepts” (A19/B33). It is crucial to remember that these HOTs are not normally themselves conscious and so they and their concepts are “presupposed” in conscious experience. We might say that the understanding unconsciously “synthesizes” the raw data of experience. Part of the motivation for HOT theory is to explain when and how an unconscious state becomes conscious, and the answer is that the subject becomes aware of the state, that is, a HOT is directed at it (recall also the Transitivity Principle). Thus the concepts in question must also be in the HOTs and are at least partly responsible for the “what it is like” nature of qualitative experience.

Notice that this solution is unlike reductionist accounts in nonmentalistic terms, and so it is immune from Chalmers’ criticism about the plausibility of those theories. For example, there is no problem about how some specific brain activity produces conscious experience. Chalmers’ criticism that functional explanations are inadequate because one can always ask “Why is the performance of these functions accompanied by experience?” (1995, p. 203) is equally beside the point. The HOT theory is not a functional explanation which merely addresses what Chalmers calls the “easy” problems of consciousness. In any case, HOT theory contends that a reductionist theory of consciousness can be provided in mentalistic terms in a way that can solve the hard problem.

1. **Realism and Idealism**

One might even go further and reject the very existence of mind-independent physical substances (including bodies and brains) and embrace some kind of idealism such that consciousness (or conscious minds) is the only real and fundamental substance or thing in nature or the universe. Indeed, this at least seems to be the view of some other authors in this volume and at the 2013 Toward a Science of Consciousness conference in India (e.g. Chopra and Kafatos, this volume). Idealism holds that there are only immaterial mental substances, a view more common in some Eastern philosophies (such as in the Vedanta tradition). The most prominent Western proponent of idealism was the eighteenth century empiricist George Berkeley (1710/1975). The idealist agrees with the substance dualist, however, that minds are non-physical but rejects the notion that there are really mind-independent “material” or (even) “physical” things at all. This is sometimes today framed in terms of some kind of all-encompassing fundamental “consciousness” in the universe which somehow “permeates” or “creates” perceptible reality (Chopra and Kafatos, this volume). Indeed, some suppose that contemporary quantum physics actually leads us naturally toward this sort of view and away from realism. Realism is simply the view that there are mind-independent physical objects and we can know at least something about what they are really like. Importantly, however, there is also a long-time common distinction made between *direct* realists who think that we are directly aware of external mind-independent objects and *representative* (or *indirect*) realists, such as Locke (1689/1975), who hold that we directly aware of our own sensory or perceptual states but indirectly aware of outer objects (which cause our sensory states). Both realist views claim that we can know, at least to some extent, what mind-independent reality is really like.

 One important side issue first: For those who use quantum physics as a reason to embrace idealism, there is significant potential for confusion by conflating what is ‘non-material’ with what is ‘non-physical.’ As we saw earlier, this is a very important distinction. One might reject the existence of *material* objects based on current quantum theory. In a similar fashion, perhaps there is only “really” energy of some kind which can form into material things under the right conditions. But none of this leads automatically to the conclusion that there are no mind-independent *physical* things (or forces). After all, isn’t quantum physics still *physics*? Isn’t there still something *physical* in the universe in that case? If so, then we are not really talking about idealism as it is typically construed in the history of philosophy.

 In any case, why would someone hold an idealist view and what problems does it face? After all, in order to judge one theory as better than another we must compare its advantages and disadvantages to other theories.

Unlike some realists, I am willing to concede here (at least for the sake of argument) that we are only *directly* aware of our own conscious or sensory states and so we should reject direct realism.

However, one central question remains: What, outside of us, *causes* these sensory states? After all, even Berkeley conceded that something (or someone) *else* must cause our conscious states; we certainly don’t seem to be constantly causing such states in ourselves, i.e. conjuring up in imagination, say, our visual perceptions as we walk down the street. Thus, his answer was that God is constantly causing sensory ideas in each of us. (Some in the Eastern tradition might call “God” something more like “pure consciousness” instead.) So Berkeley had to invoke a very busy “hands on” God to answer such a question about what causes our sensory perceptions. Well, according to Bishop Berkeley, God is omnipotent and omniscient and so is obviously up to the task! But we should note that Berkeley is making no less of an *inference* than the indirect realist who instead infers that mind-independent physical objects likely cause our sensory states. However, in order for an idealist even just to explain why we all see *the same* table in a room, God has to cause many angle-different spatial visual perceptions in each of us. And that’s assuming that it even makes sense to say that it is the *same* table according to idealism (since that sounds mind-independent and enduring in some sense). The realist, however, has the far simpler and ready answer that the single mind-independent table is simply causing each of our visual perceptions. Realists argue that it is a much more reasonable to suppose that it is the existence of the single mind-independent table which causes our visual perception of the table. This is of course *not* to say that tables and any such objects really are just the way they appear. Perhaps we cannot even know what they are “really” like, but this would at most result in a view like Kant’s Transcendental Idealism, that is, there are mind-independent objects or things but we cannot know what they are like “in themselves.”

But what is crucial here is that even if we agree that “we are only directly aware of own consciousness or sensory ideas,” this does not automatically lead to an idealist conclusion. It may be true that some realists uncritically assume something more like “direct” or “naïve” realism, but surely there are other reasons why one might reasonably infer that there is some kind of mind-independent reality, even if we disagree about how much we can know about it and even agree that different people (and animals for that matter) experience that reality in different ways. That is, we must be careful not to fall prey to a false dichotomy between idealism and direct realism. This would be to make an elementary mistake. Furthermore, Locke centuries ago also distinguished between “secondary qualities” (e.g. color, sound), which are those features of perceived objects which are mind-dependent (at least partially), and “primary qualities,” which are arguably mind-independent (e.g. shape, size). Now he may have been wrong about some of his specific examples, but my point here is that this further viable option of indirect or representative realism has also been around for a long time (not to mention the additional option of Kant's transcendental idealism).

We might also therefore concede that *the way* humans experience the world (tables, streets, food, etc.) are very likely different from other species. Interestingly though, the main evidence for this today comes from understanding the differences in our physical brains and perceptual apparatus. But due to that same brain evidence, one might also argue that some of our experiences are not entirely different from all other animals. It probably depends on the animal or kind of experience. Surely there can be some experiential overlap among species depending on brain similarities. But it is the comparative brain evidence that actually helps to explain differences in our experiences. So, for example, I would expect a bat to experience the world more differently than humans than a chimpanzee would. An insect with a hundred eyes or a dolphin must experience a different perceptual reality since they have different sensors and different brains. But none of this can be used as an argument for idealism as far as I can see. Could one really use differences between animal brains – material objects themselves! – to show that there are no mind-independent material objects at all? This would be curious indeed. If nothing physical (or even material) at all exists mind-independently, then the same should go for our bodies including our brains themselves (including neurons, microtubules, neurotransmitters, etc.). Surely a reasonable person might have a hard time swallowing such a consequence. Berkeley might respond that “the brain exists only in the mind” as he once said. Maybe this is true somehow, but surely this will sound very strange to a great many intelligent and reasonable people. Further, if we each experience *reality* differently, doesn’t even this alone presuppose that there is some kind of mind-independent reality in the first place? It could simply be that there is the mind-independent food, tree, table, street, planet, etc.., *and* that we each experience those same mind-independent things differently to some degree. A realist can again acknowledge that physical objects are not “really” the way they “appear” to us. Just because humans experience the world differently from other species, it does not follow that there is *no* mind-independent reality (even if we can’t know what it is like). We must be careful not to fall for a false dilemma between direct realism and idealism.

 In addition, there are significant standard problems for any idealist view:

First, when ten of us see a table that we are all sitting around, is there one table or ten? If ten, then how can any science be done at all and why do our experiences cohere so well? If just one, then isn’t it mind-independent in some sense?

Second, when a completely new astronomical observation or discovery is made, such as the discovery of a very distant galaxy, didn’t the galaxy have to exist first and mind-independently? If so, how can that be if idealism is true? Wasn’t there a *time lag* from when the light left the galaxy and our visual conscious perception of the galaxy? How could that be if there is only consciousness and sensory ideas?

Third, if we do not allow for any inferences from each of our own minds to other things, then it also seems that we shouldn’t claim to know that there are other *minds* either. This, I think, leads to an untenable solipsist view such that each person can only know, or even reasonably believe, that he or she exists.

Fourth, if idealism is true, doesn’t this result in a kind of “reverse hard problem,” namely, how does “mind” or “consciousness” create matter or even just the appearance of matter? This is notoriously difficult to answer. Attempts to address this sort of problem by the likes of Berkeley, Leibniz, and some Indian scholars are fraught with difficulties.

Perhaps we should even agree that we cannot “prove” (in some very strong sense of the term) *either way* what mind-independent reality is really like. Nonetheless, I still think that there is good reason to believe *that* there is something mind-independent because it is the best way to explain what causes our perceptions and the coherent stream of our sensory states in the first place, among other things. So this would be more like an “inference to the best explanation” or perhaps a Kantian-style transcendental argument. For one thing, the coherence of our “waking life” sensory states is of course so much greater that our dream states so I do not think that any analogy between idealism to conscious dreams works well. Much the same is the case for hallucinatory experiences. Further, those kinds of experiences are also increasingly being investigated as physical processes in the brain.

Here, just as an example, is a Kantian-style argument in line with the above argument (see, for example, based on Bardon 2006):

(1) I make judgments about the temporal order of my own mental states.

(2) I could not make judgments about the temporal order of my own mental states without having experienced enduring substances independent of me undergoing alteration.

*Therefore*, (3) Mind independent enduring substances exist.

As Bardon (2006, sec. 1) explains, Kant at least tries to establish “a claim to knowledge of the existence of enduring, independent objects by showing that the skeptic is committed to something (in this case, consciousness of one’s own perceptions as ordered in time) that is impossible without the existence of such objects. The skeptic thus is either committed to the existence of such things by virtue of accepting the obvious fact that we are conscious of our own perceptions as ordered in time, or presumes the existence of such things in the very attempt to raise doubt about it.” In short, we couldn’t even distinguish the “subjective succession” of our perceptions and the more “objective succession” of experienced objects (or events) unless there was at least something enduring independently of our minds.

It may be impossible to prove that there is a mind-independent reality if we mean something like “prove with absolute certainty and using empirical premises only,” but then it is surely also impossible to prove that there is *no* mind-independent reality either. On the other hand, if we are permitted to use some *a priori* reasoning (or premises) or Kantian-style transcendental arguments (such as the one above), then it may very well be possible to prove *that* there is a mind-independent reality of some kind (even if we can’t *know* exactly what it’s like).

 **D. A Brief Look at the Indian Buddhist Tradition**

Interestingly, although idealism is embraced more in the Eastern (especially Indian) tradition, this is not always the case and many of the above issues are raised in the relevant literature (see e.g. Coseru 2012a, 2012b). I will briefly describe one line of thought.

Coseru (2012a) first explains that “the term that most often translates as consciousness (vijñāna) is synonymously used for designating ‘mind’ or the ‘life principle’ in the most generic sense of those terms” (p. 10). It does seem true that at least the Yogācāra School holds to the idealist view that consciousness acts as a “causal condition” of the appearance of phenomena, though not all commentators agree that this is an outright rejection of external mind-independent objects. Coseru (2012a) points out that the Yogācāra school, associated with the works of the half-brothers Asaṅga and Vasubandhu, is often identified as the “Mind-only” or “Cognition-only” School (Cittamātra or Vijñaptimātra) which adheres to “a metaphysical picture of cognitive awareness or consciousness as the only ultimately existing reality” (p. 35). Moreover, Vasubandhu does offer a detailed defense of idealism. But, for example, one of his arguments is that since non-existing entities can be consciously apprehended, consciousness must have ontological priority over material objects. This does not of course establish a thoroughgoing idealism. At best, it shows that our conscious minds are capable of conjuring up or imagining some non-existent objects. It does not show that our conscious minds imagine or create *all* such appearances. Indeed, as Berkeley recognized and as we saw earlier, we certainly do not seem to cause all of our own sensory perceptions. Quite the contrary -- most perceptions seem “given” to us or caused by something (or someone) else. Vasubandhu also appeals to a “dream argument,” namely, that “objects in a dream appear as having determined spatio-temporal coordinates without there being any correspondence to external realities” (p. 35). But the objection to this argument raised above applies equally here, and we also saw earlier that appeals to dreaming cannot account for the coherent stream of experiences in normal waking life. Further, many of the things we dream about can be traced back to earlier perceptions in our waking life.

Coseru (2012a) notes that May (1971) and Schmithausen (1973) are, for example, among the strongest proponents of the idealist interpretation the Yogācāra ontology of mind. However, Coseru also explains that “Wayman (1965) is perhaps the first scholar to have challenged the idealist interpretation of the vijñapti-mātra (i.e. ‘cognition only’ or ‘nothing but cognition’)” (p. 38). Wayman argues that Yogācārins like Vasubandhu do not really deny the reality of external objects, but merely more plausibly argue that cognitive activity can arise in the absence of an external object. Numerous scholars since Wayman have argued against the idealist interpretation of the Yogācāra metaphysical world view (such as Kochumutton 1982, Kalupahana 1987, Lusthaus 2002, Dunne 2004).

Other authors, such as Dignāga and Dharmakīrti, do explicitly take perceptual objects to have causal efficacy. What is called ‘perception’ involves an object giving rise to a perceptual image (ākāra) which represents the specific properties of the object. “The perceptual image is causally determined by the object, but the *manner* of its appearance is determined by factors that are intrinsic to cognition itself” (Coseru 2012a, p. 43, emphasis added). This position, for example, fits comfortably with the view of representational realists like Locke. The way, or ‘manner’, of an object’s appearance is, at least in part, determined by the nature of our perceptual apparatus, such as with the secondary qualities. The same would also be true with respect to the ways that different species might experience objects, as we discussed earlier.

So there is not even general consensus as to how best to think of the metaphysics of the Yogācāras. But one alternative view would simply be a typical “phenomenological” view, that is, physical entities are mind-independent, but we have no way of knowing how they are in themselves apart from the way we perceive and conceive them. It is crucial to distinguish the *metaphysical* claim of the *existence* of a mind-independent reality of some kind with the *epistemological* claim that we can *know* what such reality is like “in itself.” This is reminiscent of the more Kantian position outlined above. It is Transcendental Idealism, not Berkeleyan Idealism. Nor is it any form of idealism which rejects the existence of all physical mind-independent reality. Indeed, Kant himself was very careful to distinguish his view from the likes of Descartes, Berkeley, and Leibniz.

**IV. Conclusion**

 In any case, although it seems that any eventual neural realization of HOT theory must be somewhat widely distributed in the brain, it remains unclear how widely distributed it needs to be. I have argued against the view that HOT theory should treat first-order (i.e. world-directed) conscious states as requiring prefrontal cortical activity. Nonetheless, even though a HOT theorist is not automatically committed to a materialist theory of mind, it is reasonable to suppose that conscious states are realized in the brain. I then explored some of the key background metaphysical views involved in understanding the nature of consciousness, such as the debate between realism and idealism as well as the prospects for solving the so-called “hard problem” of consciousness. Some of the differences in question often mirror traditional differences between Western and Eastern perspectives on the nature of consciousness. Overall, I have argued that some form of realism and physicalism is more plausible than the opposing views and that materialists (and especially HOT theorists) can offer plausible replies to the hard problem.

**References**

Alter, T. and Walter, S. eds. 2007. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. New York: Oxford University Press.

Alvarez, J. and Emory, E. 2006. Executive Function and the Frontal Lobes: A Meta-Analytic Review. *Neuropsychology Review* 16: 17-42.

Bardon, A. 2006. Transcendental Arguments. *Internet Encyclopedia of Philosophy*. Available at: <http://www.iep.utm.edu/trans-ar/>

Berkeley, G. 1710/1975. *A Treatise Concerning the Principle of Human Knowledge*, in George Berkeley *Philosophical Works*, edited by Michael R. Ayers, everyman edition. London: J. M. Dent, 1975.

Chalmers, D. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2: 200-219.

Chopra, D. and Kafatos, M. From Quanta to Qualia. This volume.

Coseru, Christian, 2012a “Mind in Indian Buddhist Philosophy”, *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/mind-indian-buddhism/>.

Coseru, C. 2012b. *Perceiving Reality: Consciousness, Intentionality, and Cognition in Buddhist Philosophy*. New York: Oxford University Press.

Crick, F. and Koch, C. 1990. Toward a Neurobiological Theory of Consciousness. *Seminars in Neuroscience* 2: 263-75.

Crick, F. 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Scribners.

Dunne, J. 2004. *Foundations of Dharmakīrti’s Philosophy*. Sommerville, MA: Wisdom Publications.

Francescotti, R. 2001. Property Dualism without Substance Dualism? *Philosophical Papers* 30: 93-116.

Gennaro, R. 1996. *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness*. Amsterdam and Philadelphia: John Benjamins.

Gennaro, R. 2005. Consciousness. *Internet Encyclopedia of Philosophy*. Available at: <http://www.iep.utm.edu/consciou/>

Gennaro, R. 2012. *The Consciousness Paradox*. Cambridge, MA: The MIT Press.

Goldberg, I., Harel, M., and Malach, R. 2006. When the Brain Loses Its Self: Prefrontal Inactivation during Sensorimotor Processing. *Neuron* 50: 329-339.

Grill-Spector, K. and Malach, R. 2004. The Human Visual Cortex. *Annual Review of Neuroscience* 7: 649-677.

Kalupahana, D. J. 1987. *The Principles of Buddhist Psychology*. Albany: State University of New York Press.

Kant, I. 1781/1965. *Critique of Pure Reason*. Translated by N. Kemp Smith. New York: MacMillan.

Kochumuttom, T. A. 1982. *A Buddhist Doctrine of Experience: A New Translation and Interpretation of the works of Vasubandhu the Yogācārin*. Delhi: Motilal Banarsidass.

Kozuch, B. 2014. Prefrontal Lesion Evidence against Higher-Order Theories of Consciousness. *Philosophical Studies* 167: 721-746.

Levine, J. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354-361.

Levine, J. 2001. *Purple Haze: The Puzzle of Conscious Experience*. Cambridge, MA: MIT Press.

Loar, B. 1997. Phenomenal states. In Block, N., Flanagan, O., and Güzeledere, G. eds. *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Locke, J. 1689/1975. *An Essay Concerning Human Understanding*. P. Nidditch (ed.). Oxford: Clarendon.

Lusthaus, D. 2002. *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Cheng wei-shih lun*. London: Routledge Curzon.

May, J. 1971. La philosophie bouddhique idéaliste. *Asiatische Studien* 25: 265–323.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review* 83: 435-456.

Newen, A. and Vogeley, K. 2003. Self-representation: searching for a neural signature of self-consciousness. *Consciousness and Cognition* 12: 529-543.

Robinson, W. 2004. *Understanding Phenomenal Consciousness*. New York: Cambridge University Press.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson and Company.

Rosenthal, D.M. 1997. A Theory of Consciousness. In Block, Flanagan, and Güzeledere eds. *The Nature of Consciousness*. Cambridge, MA: The MIT Press.

Rosenthal, D.M. 2000. Consciousness and Metacognition. In D. Sperber ed. *Metarepresentation*. Oxford: Oxford University Press.

Rosenthal, D.M. 2002. Explaining Consciousness. In D. Chalmers ed. *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press.

Rosenthal, D.M. 2005. *Consciousness and Mind*. New York: Oxford University Press.

Saxe, R. 2010. The right temporo-parietal junction: a specific brain region for thinking about thoughts. In A. Leslie and T. German eds. *Handbook of Theory of Mind*. Psychology Press.

Saxe, R., and Pelphrey, K. 2009. Introduction to a special section of developmental social cognitive neuroscience. *Child Development* 80: 946-951.

Saxe, R., and Powell, L. 2006. It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science* 17: 692-699.

Schmithausen, L. 1973. Spirituelle Praxis und Philosophical Theorie im Buddhismus. In *Zeitschrift für Missionswissenschaft und Religionswissenschaft* 57: 161–86.

Swinburne, R. 1986. *The Evolution of the Soul*. New York: Oxford University Press.

Wayman, A. 1965. The Yogācāra Idealism. *Philosophy East and West* 15: 65–73.