Available online at www.sciencedirect.com
ScienceDirect
Consciousness and Cognition 16 (2007) 897–912
Consciousness and Cognition
www.elsevier.com/locate/concog

A cross-order integration hypothesis for the neural correlate of consciousness

Uriah Kriegel

Department of Philosophy, University of Sydney, Sydney, NSW 2006, Australia
Department of Philosophy and Center for Consciousness Studies, University of Arizona, Tucson, AZ 85721, USA

Received 27 June 2006
Available online 16 May 2007

Abstract

One major problem many hypotheses regarding the neural correlate of consciousness (NCC), face is what we might call "the why question": *why* would this particular neural feature, rather than another, correlate with consciousness? The purpose of the present paper is to develop an NCC hypothesis that answers this question. The proposed hypothesis is inspired by the cross-order integration (COI) theory of consciousness, according to which consciousness arises from the functional integration of a first-order representation of an external stimulus and a second-order representation of that first-order representation. The proposal comes in two steps. The first step concerns the "general shape" of the NCC and can be directly derived from COI theory. The second step is a concrete hypothesis that can be arrived at by combining the general shape with empirical considerations.

Keywords: Consciousness; NCC; Representation; Self-representation; Cross-order integration; Metacognition; Synchrony

1. Introduction

One major problem many hypotheses regarding the neural correlate of consciousness face is what we might call "the why question": *why* would this particular neural feature, rather than another, correlate with consciousness? Consider, just by way of illustration, the early hypothesis that 40 Hz oscillations in the cerebral cortex are the neural correlate of consciousness (Crick & Koch, 1990). One might wonder *why* cortical 40 Hz oscillations would be particularly fit to implement conscious awareness. Presumably 40 Hz oscillations do not inherently possess any magical feature that might spark conscious awareness. So an explanation is called for that would shed light on 40 Hz's alleged correlation with consciousness.

One attitude toward the why question is to say that it has no answer. Thus, a proponent of the 40 Hz hypothesis could say that it is simply a brute fact that 40 Hz oscillations are the neural correlate of consciousness (NCC)—there is no deep reason why they should be. To my mind, this attitude is somewhat unsatisfying,

E-mail address: kriegel@email.arizona.edu

1053-8100/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2007.02.001

though perhaps not entirely unreasonable. In any case, it would surely be *better* if we could not only identify the NCC, but also answer the why question about it.[1] Thus we might just distinguish two kinds of NCC hypotheses, which we might label, for the sake of convenience, as *explanatory* and *descriptive*. An explanatory NCC hypothesis offers a specific neural structure or process as the NCC, and says *why* it is this structure or process, rather than another, that is the NCC. A descriptive NCC hypothesis offers an NCC, but does not say why.[2]

Arguably, the only way to offer an explanatory hypothesis about the NCC is to (i) identify the essential cognitive and/or phenomenological characteristics of consciousness and (ii) show that the hypothesized NCC underlies those characteristics. This would require a theory of consciousness at the cognitive and/or phenomenological level. The present claim is that only when conjoined with such a theory can a hypothesis about the NCC be explanatory rather than merely descriptive, that is, answer the why question.

More subtly, we might distinguish two directions of research in the search for the NCC. One proceeds *bottom-up* and seeks concomitant variations between, on the one hand, certain neural structures and processes, and on the other hand, the presence and absence of conscious awareness. The other is *top-down* and seeks the NCC by first theorizing about the key ''macro-level'' (cognitive and/or phenomenological) features of consciousness and then looking for the neural structures and processes that underlie those features. These directions of research are certainly not exclusive, and a healthy interplay between bottom-up and top-down considerations is possible and desirable—and in fact characterizes virtually all scientific work on the NCC. But what I want to stress is that only an approach with a strong top-down component can answer the why question and produce explanatory NCC hypotheses.[3]

One prominent hypothesis that certainly does put a strong emphasis on the top-down explanatory direction in a sustained manner is the hypothesis that widespread activity in a fronto-parietal network is the NCC (or at least is a major component thereof) (Baars, 2002; Baars, Ramsøy, & Laureys, 2003; Dehaene & Naccache, 2001; Dehaene, Sergent, & Changeux, 2003). What makes the hypothesis explanatory rather than descriptive is that its proponents have an answer to the why question. The answer is that (a) consciousness is the global workspace of cognition and (b) the widespread activity in the fronto-parietal network appears to underlie precisely such a global workspace. Claim (a) is contributed by a theory of consciousness at the cognitive level, the Global Workspace theory,[4] while claim (b) is contributed by neuroscientific research. This illustrates the way a ''macro-level,'' cognitive theory of consciousness can be deployed to answer the why question and make an NCC hypothesis explanatory.

The purpose of the present paper is to develop a different explanatory NCC hypothesis, one that is inspired by a different theory of consciousness at the cognitive and phenomenological level, a theory I call the *Cross-Order Integration* theory.[5] The proposal comes in two steps. The first step concerns the ''general shape'' of the NCC and can be directly derived from cross-order integration (COI) theory. The second step is a concrete hypothesis that can be arrived at by combining the general shape with empirical considerations.

[1] For one thing, in a state of imperfect knowledge it encourages skepticism about the hypothesis. Even if 40 Hz oscillations correlate with consciousness in a range of observed cases, in the absence of an explanation for this correlation, there is no reason to believe that the correlation will hold also in the far vaster set of as yet unstudied circumstances. (Conversely, if a hypothesis about the NCC identifies a structure or process that correlates significantly, but not perfectly, with consciousness, but also offers a good explanation of *why* the two should correlate, then it is reasonable—at least more reasonable than when the hypothesis is offered as a brute fact—to suspect that the evidence against the hypothesis is accounted for by some limitations or blindspots in the relevant experimental studies.)

[2] These labels are intended just as such—as labels. It is no part of this labeling practice to imply a judgment on the usefulness of the two types of hypothesis.

[3] A purely bottom-up approach is likely to produce only descriptive NCC hypotheses. This claim is not intended to be uncontroversial. Conceivably a case could be made for a purely bottom-up approach. In some respects, the debate here would be a methodological parallel of an ongoing philosophical debate about the nature of reduction. Whereas some philosophers maintain that all reduction requires an explanatory dimension (Chalmers & Jackson, 2001), others hold that it does not (Block & Stalnaker, 1999). Thus, it is debated whether the reduction of water to H_2O required an explanation of why H_2O should be water, by showing that H_2O underlay the observable macrophysical phenomena associated with water, or could proceed simply by stating that, since the presence and absence of water correlated perfectly with the presence and absence of H_2O, the best explanation for this is that they are one and the same.

[4] For the *locus classicus*, see Baars (1997).

[5] The theory is introduced, under this name, in Kriegel (2005) and further developed (under a different name) in Kriegel (2006). But the ideas underlying it are already present in Kriegel (2003b).

The paper divides in four. Section 2 summarizes the main features of the COI theory of consciousness. Section 3 derives the general shape of the NCC from the theory. Section 4 proceeds to articulate a more concrete proposal. Section 5 discusses the hypothesis' testability.

2. The cross-order integration theory of consciousness

Many psychological events and states are representational. A perception of a blue patch *represents* the blue patch. Some perceptual representations of blue patches are conscious; they are perceptual experiences that represent blue patches. But there are also non-conscious representations of blue patches—in subliminal perception, blindsight, dorsal-stream visual processing, etc. So representing a blue patch is not sufficient for consciousness. The same consideration applies to the representation of any external stimulus. We can always readily envisage both a conscious and a non-conscious representation of the same stimulus.

According to one family of views, this consideration does not entirely sever the connection between consciousness and representation. Although representing something is *not* the mark of conscious states, it may well be that *being represented* (in the right way, at least) *is*. This family of views is motivated by the basic idea that conscious states are states we are aware of (Rosenthal, 1990, 2002). To be sure, we are aware of them in some special, phenomenologically immediate way that requires elucidation. But we are certainly aware of them. The thought is that since being aware of something involves representing it, if conscious states are states we are aware of (in an appropriately unmediated way), then conscious states are states that are represented by us (in the kind of way that would make us immediately aware of them).[6]

What is this special way in which all and only conscious states are represented? Here appear first divisions within the family. There are at least four distinct views on the matter one can find in the relevant literature. I will refer to these views as higher-order thought (HOT) theory, higher-order monitoring (HOM) theory,[7] self-representationalism, and the Cross-Order Integration (COI) theory.

According to HOT theory, conscious states are states we have *thoughts* about—thoughts that were not formed through any kind of conscious inference (Rosenthal, 1993, 2002). Having the thought makes us aware of the conscious state the thought is about, and the fact that the thought was not arrived at by a conscious inference makes that awareness experientially immediate. Thus when a subject has such a non-inferential thought about one of her psychological states, the state in question becomes conscious.

According to HOM theory, by contrast, conscious states are states that are represented by a higher-order monitoring mechanism whose operation resembles the operation of the mechanisms of sensory perception (Armstrong, 1968; Lycan, 1996). The monitoring mechanism's representation of the conscious state makes us aware of the state, and the fact that the monitoring mechanism works analogously to sensory mechanisms ensures that the awareness in question is immediate in the appropriate sense.

According to self-representationalism, conscious states are those that are represented, not by higher-order thoughts or perception-like states, but simply by themselves (Smith, 1986; Caston, 2002; Kriegel, 2003a). Their being represented *at all* makes us aware of them, and their being represented *by themselves*, rather than by separate higher-order states, makes the awareness unmediated in a very straightforward sense. Thus when a subject is in a psychological state which, whatever else it represents, also represents itself, the state in question is conscious.

According to COI theory, conscious states arise from the integration, or unification, of what are initially two distinct representations, a first-order representation of an external stimulus and a higher-order representation of that first-order representation; once the two representations are unified, they form a single representational state with two parts, one directed at the other and the other directed at the stimulus. The fact that this unified representational state has a higher-order component makes us aware of the state in question, and the

[6] For a straightforward presentation of this reasoning, see Lycan, 2001. The view that conscious states are states we are aware of having has been the target of some debate in the philosophical literature. There is no need to reproduce the main moves in this debate here. For some arguments against the view, see Dretske (1993) and Block (1995). For some arguments (other than Lycan's) in favor of the view, see Rosenthal (1986), Chalmers (2003) and Levine (2001) Ch. 4.

[7] Often also referred to as higher-order perception theory, on the grounds that it construes the representation of a conscious state as quasi-perceptual, in the sense of being strongly analogous to floor-level perception.

fact that the two representations are cognitively fused into one makes the awareness in question phenomenologically immediate. Thus when a subject has a higher-order representation that is unified with the first-order representation it represents, their representational unity constitutes a conscious state.[8]

This paper adopts the COI theory. I will not present detailed argumentation in favor of the theory.[9] The purpose of the paper is rather to show that, once we adopt the theory, we obtain an attractive NCC hypothesis that is motivated primarily by top-down considerations. Such an NCC hypothesis has not been produced to-date by the proponents of HOT, HOM or the self-representational theory. So in a way the paper as a whole produces a roundabout argument for COI theory; namely, that unlike its competitors,[10] COI theory delivers an NCC hypothesis.

Nonetheless, let me motivate the theory minimally with some elementary phenomenological considerations. Let us consider a mundane conscious experience. Suppose that, as I stare at the blue sky, I undergo the following experience. First of all, I have a visual experience of a blue expanse. But I also at the same time hear the roar of car engines passing by; can feel the seat under me, as well as the soles of my shoes; am faintly aware, in an unpleasantly anxious sort of way, that I have yet to pay last month's telephone bill; and am even more faintly aware of myself having this very bluish-roarish-anxious conscious experience.

Some of the items within my overall experience are more phenomenally pronounced than others, but each contributes *something* to my experience's overall phenomenology. If the feeling of the soles of my shoes were to suddenly disappear (say, because my foot nerves were incapacitated), I should notice a change in my overall phenomenology. We can distinguish, in fact, at least six separate individual items in the overall phenomenology of my experience: the visual perception of blue, the auditory perception of roar, the tactile sensations of the seat and of the soles, the emotional feeling of anxiety, and the unmediated awareness of myself as undergoing the above. These six items are then (synchronically) unified, in that they do not feel like just so many unconnected items occurring side by side, but rather like a single cohesive experiential field.

This phenomenological description is not meant to be uncontroversial, but I do hope that it resonates with the reader. What I want to stress is that there are three components of different kinds in this overall phenomenology. First, there are the five first-order sensory and emotional items; secondly, there is the second-order awareness of them; and thirdly, there is the unity among all six items. COI captures these three kinds of component very straightforwardly: as a matter of the cognitive unification of five first-order representations with a single higher-order representation.

The same cannot be said of the HOT, HOM, and self-representational theories. HOT and HOM theories construe our awareness of our conscious experiences as external to that experience (as implemented in a separate mental state), and therefore cannot find place for that awareness among the phenomenal items making up the experience.[11] Self-representationalism cannot account for the fact that our awareness of our conscious experience is only *one part* of the experience; for it construes the experience as, in some sense, wholly representing its-whole-self. The internal phenomenological structure of the experience as involving three types of component is lost.

It appears, then, that COI theory is phenomenologically adequate in a way that its competitors are not. In any case, it is clear that it is phenomenologically *different* from its competitors, and in that sense presents a distinct account of consciousness at the phenomenological "macro" level. We can also safely say that it presents a distinct account of consciousness at the *cognitive* "macro" level, for it portrays consciousness in terms of a representational structure unlike that employed by its competitors. Inasmuch as the notion of representation is central to our understanding of cognition (and after all, cognition is commonly thought of as essentially the manipulation of representations), COI theory draws an intimate connection between consciousness and cognition, though quite a special and unusual one—which arguably is what one would expect.

[8] For versions of this view, though not under this label, see Van Gulick (2001, 2006) and Kriegel (2003b, 2005, 2006) (only Kriegel, 2005 uses the label "Cross-Order Integration").

[9] For such argumentation, see Kriegel (2005) and Kriegel (2006).

[10] At least this is the case regarding its competitors within the family of views inspired by the idea that conscious states are states we are aware of (in a phenomenologically immediate way).

[11] Since the higher-order representation is external to the conscious state, it is itself unconscious and does not make part of the subject's overall phenomenology.

Much more can be said about COI theory and its philosophical motivation and underpinning. However, the focus of the present paper is not so much on the independent reasons on the basis of which one might adopt COI theory as on the theoretical serviceability of the theory: the fact that it delivers a potentially concrete and testable hypothesis regarding the NCC. Thus the preliminary sketch of COI theory provided in this section should suffice to derive a proposal regarding the general shape of the NCC in the next section.

3. The "General Shape" of the neural correlate

According to COI theory, a conscious state arises from the cognitive integration of a first-order representation and a higher-order representation of that first-order representation. To generate a testable hypothesis about the NCC from the COI theory, we would need to match a neural element to three items: (i) a floor-level representation,[12] (ii) a higher-order representation of that floor-level representation, and (iii) the functional integration of these two representations. There is thus a neural triad of correlates we would need to pin down before we can specify the neural correlate of a given conscious state.

The first element, however, falls outside the purview of the theory of consciousness proper. Depending on which kind of conscious state we are dealing with, different parts of cognitive neuropsychology would have to be "fielded" to match the appropriate neural correlate of the first element. For a conscious visual experience of blue, the floor-level representation would probably consist in neural activation in some subpopulation of neurons in V4.[13] But for an auditory conscious experience it would have to be activation in auditory cortex, for a conscious experience of recognizing a friend's face it would be activation in the fusiform face area (see Kanwisher (2000)), etc. These are essential components of the neural correlates of specific conscious states, but none of them correlates with the presence or absence of consciousness in general, since each has to do with a particular kind of conscious state. We can think of these floor-level representations as determining the specific *contents* of consciousness once consciousness is present, rather than ensuring the presence of consciousness in the first place.

It is worth pausing to stress the distinction between the contents of consciousness and consciousness as such. The contents of consciousness vary across different conscious experiences: a visual experience of blue, a visual experience of a friend's face, and an auditory experience of a bagpipe have different contents. But these three experiences also share something—they are all conscious experiences! There is something that is invariable across all conscious experiences, and which we may therefore call *consciousness-as-such*.

It is a virtue of COI theory that it implies a neat separation of the neural correlate of consciousness-as-such from the neural correlates of the contents of consciousness. The neural correlate of a conscious state is given, according to COI theory, by a functionally integrated floor-level representation of a stimulus and higher-order representation of that floor-level representation. The first element in the neural triad mentioned above, the floor-level representation of the stimulus, constitutes the neural correlate of the conscious content; the second and third elements of the triad, the functional integration with the higher-order representation, constitute the correlate of consciousness as such.

The fact that COI theory separates out the contents of consciousness and consciousness as such is a virtue, because the search for the NCC is a search for the neural correlate of consciousness as such. Ultimately, what is sought is the neural *signature* of consciousness: some neural feature that is present in all conscious experiences and absent in all non-conscious mental activity. Many theories of consciousness face significant conceptual difficulties trying to separate out that neural element which is invariable across all conscious experiences from what corresponds to the variable contents of changing conscious experiences. They might identify a neural correlate of, say, some kind of color consciousness, and perhaps a correlate of, say, emotional consciousness, but face difficulty identifying the *neural commonality* among the two. It is possible to maintain that there is *no* commonality, of course, but that would be to maintain that there is simply no neural correlate of

[12] In most real-life cases, there will be many floor-level representations, not just one. But that complexity should not affect the basic structure of the conscious experience, so for the sake of simplicity I will work here with a simplified model in which there is only one floor-level representation.

[13] See Zeki et al. (1991) for seminal work on the neural substrate of color representation.

consciousness as such. It is a virtue of COI theory that it pinpoints what would constitute the neural correlate of consciousness as such, namely, the second and third elements in the above neural triad.[14]

The second element is the neural correlate of higher-order representation. Presumably, it would have to be seated in an area of the brain that is associated with quite sophisticated cognition, since the relevant representations are sophisticated ones: they are representations not of simple sensory stimuli, after all, but of psychological states. Having such representations requires the ability to direct the cognitive system onto itself and engage in what has come to be known as *metacognition* (cognition of cognition). A good preliminary guess is that the neural substrate of such metacognitive activity is to be found in the prefrontal cortex, which is associated with executive function and other forms of sophisticated cognition.[15]

To the extent that the COI approach to the NCC requires sophisticated cognitive activity in the prefrontal cortex, it is at odds with some other, more ''minimalist'' approaches. A brain system that can detect and process information about psychological states, and ascribe these states to oneself as opposed to others, is bound to be an evolutionary late arrival. If such abilities are necessary for consciousness, some of the lower animals we would intuitively like to ascribe consciousness to may not be conscious after all.[16]

The third element in our triad is the functional integration of the two representations. By ''functional integration,'' I mean the process of unifying disparate bits of information into a single representation in a functionally significant way, that is, in such a way that the functional role of the single representation is in some sense more than the sum of the functional roles of the different bits of information making it up. Imagine a creature that engages in some activity only when it's cold and muggy; then the process by which a representation that it's cold and a representation that it's muggy are integrated into a single representation of it being cold and muggy is a process of functional integration.

There are many forms of functional integration in the cognitive system. The neural mechanisms underlying them may inspire different analogies for the functional integration of a conscious experience and one's awareness of that experience. By far the most extensively studied of those, at least in recent years, is the kind of functional integration involved in feature binding. When a person perceives a blue patch moving from left to right, the patch's blueness is represented in V4, but its left-to-right motion is represented in MT. The brain must find a way to represent that the blueness and the motion belong together as two features of one and the same stimulus.[17] A natural place to start looking for the third element in our triad is in the neural mechanism that subserves binding in general, in the hope that it might also subserve the binding of a floor-level representation and its higher-order representation.[18]

[14] Chalmers (2000) argues that the ultimate goal of the search for the NCC is to identify the neural correlate of the contents of consciousness: ''The ideal is to find a neural system from whose activity we might determine the precise contents of a visual [or other] experience.'' But the fact that the contents of consciousness vary across different conscious experiences, whereas the NCC should be that element which is invariable across all conscious experiences, suggests that this is the wrong methodological priority. The search for the NCC should be a search for the neural signature of consciousness, that is to say, the correlate of consciousness-as-such, or consciousness *per se*.

[15] It is important not to confuse the notion of ''high level'' cognition with that of higher-order cognition. The former denotes any kind of sophisticated cognitive activity. The latter denotes metacognition. If there was a form of unsophisticated metacognition, it would be higher-order but not high level. And sophisticated cognition that is not directed at cognition is high level but not higher-order.

[16] It is important to note, however, that some behaviors that are performed consciously by conscious creatures may be performed unconsciously by non-conscious creatures. If so, while consciousness is *sufficient* for the performance of these behaviors, and perhaps even *facilitates* it, it is not quite *necessary* for it. There may thus be an element of sympathetic projection involved in some of our pre-theoretic ascriptions of consciousness to animals. (See Carruthers (2004) for an argument to that effect.) When constructing a theory of consciousness, the key principle to keep in mind is that conscious states are states the subject is aware of. This principle means that consciousness implicates metacognition. The metacognitive abilities required may be quite elementary, however, so unless we take a very pessimistic view of animal metacognition (Povinelli, 2004; Povinelli & Vonk, 2003), we may be able to ascribe consciousness (justifiably) to quite a few animals. But it is certain that some of the animals we are intuitively inclined to ascribe consciousness to cannot be conscious if we are to also hold on to the (equally intuitive) principle that conscious states are states the subject is aware of having.

[17] This is the so-called binding problem (see Treisman & Schmidt, 1982; Treisman, 1996).

[18] This third element introduces, in all probability, a non-anatomical aspect to the NCC. Although one model of integrating two pieces of information might involve sending both to some special place in the brain where ''it all comes together,'' a more psychologically plausible model involves only a special relationship between the two pieces of information—for example, a relationship created by the binding mechanism. If so, there is no *neuroanatomical* correlate of consciousness. Activity (over baseline) in the brain areas responsible for the floor-level and higher-order representation is not sufficient for consciousness. A further element is the relationship between them that guarantees their functional integration.

What makes binding a form of functional integration is that its upshot is functionally significant. There is very little we know about the exact functional significance of binding, but there is evidence that it enhances task performance for some tasks.[19] Presumably, the functional significance of binding would have to do with downstream consumability: when two representations are bound, they are treated "as one" by downstream consumer systems, such as executive function systems, action guidance systems, verbal report systems, etc. Thus, when a subject is presented with a blue patch moving from left to right, once the representations of the blueness and the left-to-right motion are bound, downstream operations on one are thereby also operations on the other.

In conclusion, COI theory has a relatively specific proposal to make about the general shape of the NCC. A conscious state arises from the functional integration of a floor-level representation of some external stimulus and a higher-order representation of that floor-level representation. For example, a conscious perceptual experience of a blue patch involves the functional integration, perhaps through binding, of above-baseline activity in the visual cortex and above-baseline activity in the part, or one of the parts, of the brain subserving metacognition (perhaps in prefrontal cortex). Only when something like this happens does there emerge a mental act that folds within it both an awareness of a stimulus and an awareness of that awareness. Such fusion of world-awareness and self-awareness is the phenomenological mark of consciousness, according to COI theory, and is therefore what we need to seek in the brain if we are to identify the NCC.

4. A concrete hypothesis

As noted above, COI theory separates out the neural correlate of consciousness as such from the correlates of specific contents of consciousness. The correlate of consciousness as such involves only the second and third elements in the triad: higher-order representation and functional integration. Thus to generate a concrete hypothesis about the NCC, what is needed is evidence about the concrete brain structures and mechanisms involved in higher-order representation and functional integration. It was also noted that the most natural places to look for these are in studies of the neural mechanisms underlying, respectively, metacognition and binding: metacognition creates higher-order representations, binding produces functional integration.

4.1. Higher-order representation

Imaging studies of metacognition have found, with remarkable consistency, that activity in the prefrontal cortex, especially in the right lobe, is associated with cognition of one's psychological states and character traits. More specifically, there is a remarkable convergence on the medial prefrontal cortex (mPFC) (Lane, Fink, Chau, & Dolan, 1997; Craik et al., 1999; Gusnard, Akbudak, Shulman, & Raichle, 2001; Kelley et al., 2002; Johnson et al., 2002) with relevant activity in the anterior cingulate cortex (ACC) as well (Kelley et al., 2002; Johnson et al., 2002). Thus it might be thought that the natural hypothesis for the COI theorist to make is that mPFC activity corresponds to the second element in the triad.

There is, however, a problem with these studies that may undermine their relevance to our present project. The problem is that many of the areas associated with metacognition appear to be associated with other tasks as well, especially those involving social cognition. Both metacognition and social cognition involve cognition of psychological states; the difference is that metacognition is directed at psychological states of *oneself*, whereas social cognition is directed at psychological state of *another*. If the mPFC is activated both in metacognition and social cognition, the natural inference is that the mPFC subserves the cognition of psychological states in general, rather than more specifically cognition of one's *own* psychological states. Thus it may be that the mPFC subserves Theory of Mind (ToM)-related activity rather than metacognition proper. Metacognition is one kind of application of ToM—namely, its application to oneself—but social cognition is another.[20]

A recent study that targeted precisely the difference between cognition of psychological features of oneself and those of a significant other found that while in both tasks the mPFC was activated above baseline, it was

[19] See, e.g., Stopfer, Bhagavan, Smith, and Laurent (1997) for odor discrimination.

[20] That is, the involvement of the mPFC in both metacognition and social cognition suggests that the mPFC does not subserve metacognition as such, but whatever is common between metacognition and social cognition—which is to say cognition of psychological states. See Frith and Frith, 1999 for a review of the connection between mPFC and ToM.

also activated equally, hence non-differentially, in both. What was activated differentially was rather the dorsolateral prefrontal cortex (dlPFC), which was significantly more active during cognition of one's own psychological features than during cognition of the significant other's features (Schmitz, Rowley, Kawahara, & Johnson, 2006). If so, it may in the end be the dlPFC that is the better candidate for capturing the second element in our triad.

Clinical studies of subjects with deficits in metacognitive assessment of their own features and abilities provides convergent evidence. For starters, it has been found that schizophrenics unaware of their condition tend to have reduced gray matter volume in the dlPFC relative to schizophrenics with a more accurate understanding of their condition (Flashman et al., 2001). Similarly, patients who have a distorted assessment of their abilities after traumatic brain injury have been shown to exhibit lower activity in the medial and right dorsolateral PFC relative to traumatic brain injury patients with a more accurate assessment of their abilities (Schmitz et al., 2006).

A further problem, however, threatens to undermine the relevance of all these results. Many of the non-clinical studies in this area are based on a paradigm in which subjects are asked to state whether some trait adjectives describe them accurately. For example, the subject is presented with the adjective "honest," or "polite," and is asked to state whether s/he is indeed honest, or polite. The metacognitive judgment the subject is asked to render is not a judgment about an occurrent, local event in her mental life, but rather about a standing feature of her overall personality. Let us call the former kind of judgment *state metacognition* and the latter *trait metacognition*. The problem is that these studies tell us rather about the neural mechanisms underlying trait metacognition, whereas the mechanisms relevant to consciousness are those underlying state metacognition. Likewise for the clinical studies, which are focused on metacognition of standing abilities as opposed to occurrent states. It is possible that the mechanisms underlying metacognition of occurrent states are the same as those underlying standing features and abilities, but there is no a priori necessity that this be so. In fact, there is no a priori reason to expect that the structures and processes that subserve trait metacognition are the same as those that subserve state metacognition.

What are needed, then, are studies that target state metacognition while making sure that they dissociate it from what we might call *state social cognition*: the cognition of occurrent psychological states of others. Progress in this area is sure to be made over the next few years, and we may then learn more about what exactly subserves state metacognition. But it is reasonable to expect the relevant area to be in the prefrontal cortex, most probably medial or dorsolateral, perhaps with assistance from the anterior cingulate.

To get another angle on state metacognition, we might consider taking into account research on cognitive control (or "attentional control"). Large parts of this literature are concerned with error detection and conflict resolution in the cognitive system. When the cognitive system's activity produces what appear to be erroneous results, or conflicting results, some correction of the error, or resolution of the conflict, is in order. The system thus needs to have mechanisms for (a) detecting such problems and (b) fixing them. There are two assumptions, of a theoretical order, that would make these mechanisms relevant to our present inquiry. The first is that the mechanism in charge of detecting the problems is engaged in ongoing monitoring of activity in the system. The second is that a mechanism engaged in ongoing monitoring produces representations of floor-level representations, i.e., higher-order representations. Against the background of these two assumptions, it seems that the mechanism in charge of detecting error and conflict in the system produces higher-order representations.

The most straightforward experimental paradigm here employs the Stroop task (Stroop, 1935) in which subjects who are asked to state the color in which a color word is inked are presented sometimes with words whose meaning and color are congruent (e.g., "yellow" written in yellow) and sometimes with words whose meaning and color are incongruent (e.g., "red" written in yellow). Response times are consistently longer in the incongruent cases (this is the "Stroop effect"). This suggests that there is some sort of interference from the semantic processing of the word meanings in the performance of the actual task, which concerns only the color in which the word is displayed. This sort of interference is known as "cognitive conflict," and is thought to be the main trigger for deployment of cognitive control, whose job it is to resolve the conflict by assigning greater cognitive resources to the task at hand (this is the exercise of "attentional control") (Botvinik, Braver, Carter, Barch, & Cohen, 2001).

Interestingly, research on cognitive control has homed in on two of the three brain areas we have already met: the ACC and the dlPFC. The leading model, due to Cohen, Carter, and collaborators, is fairly straightforward. In a nutshell, the model says that the ACC is in charge of detecting conflict and error, and when it does, it sends the information regarding the problem to the dlPFC, which proceeds to fix the problem (see Botvinik et al., 2001; Milham et al., 2001; Milham, Banich, Claus, & Cohen, 2003). The model draws a neat division of labor between the ACC and the dlPFC: the ACC is in charge of *detecting* problems, the dlPFC in charge of *fixing* them (MacDonald, Cohen, Stenger, & Carter, 2000). If we embrace this model, then this research would suggest that the ACC is the monitoring mechanism that produces higher-order representations.

Others have argued against the one-directional model of Cohen et al., suggesting that both the ACC and the dlPFC are involved in both the detecting and the fixing of problems (Gehring & Knight, 2000). It is not very important for our purposes which model is favored by the evidence. If the Cohen model is correct, then the evidence from research on attentional control is that only the ACC produces higher-order representations, whereas the dlPFC does not.[21] If the alternative model is correct, then the evidence from research on attentional control is that both the ACC and the dlPFC are involved in producing higher-order representations. What is interesting is that this evidence is in line with evidence from metacognitive evaluation research. Thus we may conclude with some confidence that it is highly probable that higher-order representations are produced by the ACC, mPFC, and/or dlPFC.[22]

This evidence offers the COI theorist seven possible hypotheses about the neural correlates of higher-order representations: three hypotheses that identify a single neuroanatomical seat (ACC, dlPFC, or mPFC); three that identify two neuroanatomical seats (ACC + dlPFC, ACC + mPFC, and dlPFC + mPFC); and one that identifies three seats (ACC + dlPFC + mPFC). At this relatively early stage of inquiry, there is no particular reason for the COI theorist to be attached to one of these hypotheses rather than another. Whichever of them finds most independent (bottom-up) support as hypothesis about a component of the NCC is the one the COI theorist should embrace.

An argument could be made, however, that such independent evidence favors the dlPFC. There are two particular studies that have found clear differential activation in the dlPFC between a conscious condition and a non-conscious condition, though there is also a piece of evidence that bears against the dlPFC.

The first study focused on blindsight. Saharaie et al. (1997) used fMRI to compare overall brain activation during stimulus detection by blindsight patient GY in (i) the blind part of his visual field and (ii) the normal part of his visual field. The most significant difference between the two conditions, and by most measures the only significant one, was activation of the dlPFC, which was activated in the conscious cognition but not in the non-conscious condition.[23]

The second study is more recent and focused on metacontrast masking (Lau & Passingham, 2006). Subjects were presented with a shape stimulus—a square or a diamond—that was masked by a later stimulus at varying intervals. The subjects were asked, first, to judge whether the stimulus was a diamond or a square, and second, to report whether their shape judgment was based on what they saw or was just a guess. Lau and Passingham looked for a pair of different intervals separating the target stimulus from the masking stimulus, such that the first judgments were correct at more or less the same level in both, but the second reports differed, so that in one interval subjects reported guessing significantly more often than in the other. And indeed they found two such intervals: 33 ms and 100 ms. Judgments about the shape were correct 68% of the time when the interval was 33 ms and 70% when it was 100 ms (so the difference was statistically insignificant). But subjects reported just guessing 51% of the time in the 33 ms interval and only 40% of the time in the 100 ms interval (so the difference was statistically much more significant). On the assumptions that (a) above-chance correct judgments about the stimulus indicate that the subject did *perceive* the stimulus, and (b) reports of just guessing indicate that the subject did not *consciously* perceive it, these results suggest that, at these two intervals, there is no significant

[21] There may still be evidence from *other* research that the dlPFC produces higher-order representations.

[22] At the same time, this does not rule out the possibility that higher-order representations are produced independently in other areas.

[23] I have been told by more than one expert that this study is less technically sound than would be ideal. This suggests to me that, on its own, it would present little evidence for an NCC hypothesis. But in a supporting role, treated as subsidiary evidence to back more central evidence, it need not be disregarded.

difference in the quality of subjects' perception of the stimulus (or at least in the quality of their task performance), but there is a significant difference in the presence of consciousness (with greater presence in the 100 ms interval). Lau and Passingham compared brain activity in subjects performing the task in the two intervals and found only one brain area that was significantly more active in the 100 ms one—the dlPFC.[24]

These studies support the concrete hypothesis that the dlPFC is the sole anatomical seat of higher-order representations of one's own occurrent mental states. There is evidence against the dlPFC as well, however. It comes from sleep research, where it has been known for a while that the dlPFC is strongly deactivated during REM sleep, hence during dream experiences. On the assumption that dreams are conscious experiences, this constitutes counter-evidence against the hypothesis that dlPFC is a component of the NCC.[25]

The most straightforward way the COI theorist can accommodate this piece of evidence is to adopt one of the other seven possible hypotheses about the neural correlates of higher-order representations, since those do not give the dlPFC the exclusive function of producing higher-order representations. Probably the most plausible of these other hypotheses would be the hypothesis that higher-order representations are produced by the dlPFC and ACC. Perhaps the claim could be that both brain areas can produce higher-order representations, and while the dlPFC produces most of them, especially in standard circumstances (in some suitable sense of "standard"), the ACC produces some of them, especially in less standard circumstances.

The response, then, is that although dream research provides evidence against the supposition that the dlPFC is the sole producer of higher-order representations, the COI hypothesis as such is not wedded to the view that it is. The option of giving a central, though perhaps not exclusive, place to dlPFC activity as the neural correlate of higher-order representations of one's own occurrent mental states, as a component of the NCC, is still open to the COI theorist, and is well supported by currently available evidence. As noted above, further research on state metacognition, as opposed to both feature metacognition and (what we called) state social cognition, should shed further light on the second element in our neural triad. Let us now turn to discussion of the third element.

4.2. Functional integration

Let us turn now to the third element in our neural triad—the correlate of cognitive unification, or functional integration. It was noted above that studies of the binding mechanism are a good place to start looking for it. In a series of seminal studies, Treisman has shown that some patients may exhibit "misbinding" (or "illusory conjunctions") (see, e.g., Treisman & Schmidt, 1982). When presented with a green square and red circle, they may perceive a green circle and red square. Their perceptual system represents all the right features in the environments, but puts them together in the wrong way. Such misbinding shows that a distinctive mechanism must be involved in the "putting together" of stimuli.

The "binding problem" is the problem of how the brain executes such feature binding (see Treisman, 1996). A representation of a red circle involves three representational elements: (a) a representation of the redness, (b) a representation of the circularity, and (c) a representation of their togetherness as two aspects of a single object. The first two elements involve straightforward activation of subpopulations of neurons in dedicated parts of the brain, say V4 and V2, respectively. When the system detects redness or circularity, these neurons fire their electrical impulse at an increased rate. Thus the increased rate of firing in the V4 and V2 subpopulations represent the redness and the circularity. But what represents their togetherness? It cannot be a similar event of increased firing rate in a different part of the brain, for that event would then need to be itself bound with the other two. Some "cleverer" mechanism must be in place for representation of togetherness.[26]

[24] This result can be co-opted by the COI theorist, then, according to whom a perceptual state becomes conscious only when it is represented by a higher-order representation with which it is functionally integrated. If this higher-order representation is implemented in the dlPFC, this would explain why two perceptual states can (no doubt, in unusual circumstances) be just as perceptually reliable while only one state is conscious.

[25] Thanks to J. Allan Hobson and Antti Revonsuo for pressing this difficulty on me.

[26] The problem is especially evident when the system is presented with two objects, say a red circle and a green square. There may be two representations of togetherness in the brain, but something must make sure that what is represented as going together are the redness and circularity, on the one hand, and the greenness and squareness, on the other, rather than some other combinations.

The leading (though not uncontested)[27] model of the neurophysiological process underlying feature binding is the neural synchrony model. The idea was first proposed by von der Malsburg (1981). In the example above, each of the two subpopulations in V4 and V2 fires its electrical impulse at an increased rate. But the rates might be different. Von der Malsburg's idea was that the brain could represent the togetherness of two features by *synchronizing* the firing rates: when two features belong together, the brain activities that represent each involve firing rates that are not only increased relative to baseline, but are also more or less equal to one another. The actual occurrence of synchronization in the millisecond range was later confirmed, mainly in experiments on binocular rivalry in cats and monkeys (see Singer, 1994; Engel, Fries, Konig, Brecht, & Singer, 1999; Engel & Singer, 2001; Revonsuo, 1999 for a more theoretical overview). The mechanism that brings about this alignment is the binding mechanism, and the alignment is the binding.

Even when the two subpopulations fire at the same *rate*, they might not fire at the same *time*. However, there may also be a mechanism that brings two subpopulations to fire not only at the same rate, but also at the same time. The difference between synchronizing rate and synchronizing time may give the system the tools to represent a hierarchy of binding. For example, an audiovisual experience of a trumpet may involve representations of the trumpet's shape, color, and sound, so that the firings that encode the visual features of shape and color are time-synchronized with each other but only rate-synchronized with the firing that encodes the auditory features pertaining to the trumpet's sound. In this case, the intra-modal unity of shape and color is stronger or tighter than the inter-model unity of these visual features with sound, and the tighter unity is reflected in the tightness of synchrony.[28]

In essence, the firing rate is the brain's medium for representation of individual features, but the brain may employ other media for representing certain central relationships *among* features, such as compresence (''togetherness''). Synchrony of rate and time are two such media, but there may well be others. The more media the brain can avail itself of for representing relationships among stimuli, the more complex the resulting representations will be, possibly including varying degrees of unity among the stimuli. Such a hierarchy of binding is called for by the existence of large-scale binding in the brain (Varela, Lachaux, Rodriguez, & Martinerie, 2006). Large-scale binding is defined roughly as binding of information across parts of the brain between which transmission of information would take at least eight milliseconds. It has even been speculated that a hierarchy of binding extends across the entire cognitive system, with a weak level of unity imposed at the highest level, that of our standing worldmodel (including, as part, our self-model) (Metzinger, 1995). Given that binding mechanisms range from the quite local (e.g., binding of information represented in adjacent areas of visual cortex) to such large scale (including cross-hemispheric binding (Bressler, 1995)), one would expect the brain to have the capacity to recognize different levels of unity and organize them hierarchically.

Synchronization takes place not only within visual cortex, but also across brain areas. To effect cross-modal feature binding, there needs to be synchronization between brain areas dedicated to different kinds of perceptual information. There is even evidence of sensorimotor synchrony, which presumably effects binding of perceptual representations with motor representations (Roelfsema, Engel, Konig, & Singer, 1997). There is no reason why there would not similarly be some kind of binding of perceptual representations with self-related representations. This might involve synchronization of representations in, say, visual cortex and prefrontal cortex.

An interesting fact regarding synchrony is that although it connects what are at the sub-personal level separate representations, it results in what is at the personal level a unified single representation. When representations of a trumpet's color, shape, and sound in V4, V2, and A1 are synchronized, we do not experience ourselves to have three closely related representations. Rather, we experience ourselves to have a single representation of a colorful, shapely, melodious trumpet.[29]

[27] See Shadlen and Movshon (1999) for a critique.

[28] This is of course a simple example, but much more can be said about the relationship between different levels of synchrony. Indeed, the relationship between rate coding and temporal coding via synchrony is an issue that is continuously addressed in computational neuroscience. Our understanding of the issues involved is still very incomplete, but further research over the next decade or so should shed much more light on the possibilities raised in the main text.

[29] One of the fascinating questions in this area is *how* sub-personal synchronization underlies or realizes personal-level unity. This question ought to receive much more attention than it has to-date. Here I do not offer any account of this phenomenon, but ultimately an account ought to be provided.

4.3. Cross-order integration

The COI hypothesis for the NCC is based on the idea that the same sort of synchronization may unify a representation of a stimulus with a representation of that very representation. If the brain harbored two synchronized representations, one in V4 representing redness and another in (say) the dlPFC representing increased firing rate in V4, at the personal level we would experience ourselves to have a single representation that folds within it both an awareness of red and an awareness of that awareness. That is the sort of cognitive character that, according to COI theory, distinguishes conscious states from non-conscious ones.

This, then, is the COI hypothesis about the neural correlate of color consciousness: it is constituted by synchronized representations in V4 and (say) the dlPFC.[30] As for other kinds of visual consciousness, their neural correlates are given by synchronized representations in the relevant parts of the visual cortex and the dlPFC. The neural correlates of *auditory* consciousness are given by synchronized representations in the relevant parts of the auditory cortex and the dlPFC. And so on and so forth. In each case, the neural correlate of consciousness *as such* is given by the synchrony with representations in the dlPFC, or whatever area turns out to subserve higher-order representation, while the neural correlate of the *contents* of consciousness is given by whatever representations are thus synchronized with the relevant higher-order representation.

Presumably, there is some functional significance to such cross-order synchronization, and it ought to be of a piece with the functional significance of regular binding. This suggests that, when a mental state is conscious, its representation of the stimulus, and the representation of that representation, are treated "as one" by downstream consumer systems. This might explain why all and only conscious mental states seem to be verbally reportable and have the sort of immediate impact on rational deliberation and belief formation that they do. Just how this functional difference between bound cross-order states and unbound ones plays out is something worth investigating in future work.

It is plausible to maintain that, in complicated cases involving multiple sub-personal first-order representations, what enters the subject's overall phenomenology at any one time are the synchronized first-order and higher-order representations and any other first-order representations that may be independently bound with the first-order representation that is bound with the higher-order one. Thus, if the dlPFC representation represents a V4 representation, the two are synchronized, and the V4 representation is also bound with a V2 representation, then the content of the V2 representation will appear in the subject's phenomenology as well, even though V2 is not directly represented by the dlPFC representation. The rationale for this is that once the V4 and V2 representations are bound, at the personal level they form a single representation, and so by representing V4, the dlPFC representation represents just one part of a larger state all of which is made conscious by synchronization with the higher-order representation.

The hypothesis turns in determinate verdicts about specific cases. Let us consider several scenarios involving five sub-personally distinct representations: in V4 of a trumpet's color, in V2 of its shape, in A1 of its sound, in somatosensory cortex of a tickle in one's left arm, and in the dlPFC of the increased firing rate in V4. The first scenario is one in which the dlPFC representation is synchronized with the representations in V4, V2, and A1, but not with that in somatosensory cortex. The hypothesis predicts that the subject undergoes a conscious experience of the colored, shaped, heard trumpet, but that the tickle remains non-conscious. In a second scenario, the representation in dlPFC is synchronized with the representations in V2 and V4, but not with those in A1 and somatosensory cortex. The hypothesis predicts that the subject undergoes an experience of a colored and shaped trumpet, but that the sound and the tickle are non-conscious. In a third scenario, the dlPFC representation is synchronized with the somatosensory representation, but not with the V2, V4, and A1 representations. Here the hypothesis predicts that the subject experiences neither the trumpet (visually or auditorily) nor the tickle: not the trumpet, because the dlPFC representation is not synchronized with the trumpet-directed audiovisual representations, and not the

[30] Keep in mind, however, that the dlPFC is only one of the options the COI theorist may advert to—though perhaps the most appealing option at this point.

tickle, because the dlPFC representation does not represent the somatosensory representation with which it is synchronized.[31]

The COI hypothesis for the NCC is an *explanatory* NCC hypothesis. For visual consciousness, for instance, it not only designates synchronized activity in visual cortex and prefrontal cortex as the neural correlate. It also says *why* this should be so. It should be so because the phenomenological and cognitive mark of consciousness is a certain fusion of world-awareness and self-awareness. The synchrony of visual cortex and prefrontal cortex activity is simply the micro-level correlate of this macro-level description. The visual cortex activity underlies the world-awareness; the prefrontal activity underlies the self-awareness; and the synchrony underlies their fusion. Thus synchronized activities in the visual and prefrontal cortices would make for a fusion of visual world-awareness and self-awareness—which is why, according the COI hypothesis, it is the neural correlate of visual consciousness.

4.4. Comparison to similar hypotheses

As already stressed, the COI hypothesis just laid out is inspired by a strong top-down approach to the NCC. Interestingly, there are NCC hypotheses in circulation that already share the same upshot as the COI hypothesis either at the top level or at the bottom level, though (to my knowledge) no single hypothesis that shares both.

The NCC hypothesis that shares COI's top-level cognitive and phenomenological conception of consciousness (as involving the sort of peculiar fusion of world-awareness and self-awareness stressed above) is Hans Flohr's. Flohr's hypothesis is very unlike the COI one at the bottom level, as it designates the binding of neural assemblies by NMDA as the NCC. But interestingly, it is inspired by a very similar macro-level view of consciousness. Of a system deploying NMDA for the binding of neural assemblies, Flohr writes: "The system can bind diverse first- and higher-order representations, embed first-order representations in a model of itself and thereby represent itself as an actually representing system" (Flohr, 1995; 160). In other words, the reason *why* NMDA-bound neural assemblies should underlie consciousness, according to Flohr, is that such assemblies will recover the distinguishing mark of consciousness, which Flohr evidently takes to be the ability of a system to represent while representing itself to represent—precisely the COI-theoretical view on the matter.

The NCC hypothesis that shares (more or less) COI's bottom-level view of the NCC is the hypothesis that the correlate of visual consciousness (or at least a major component thereof) is synchronized activity in the visual cortex with projections to the prefrontal cortex. This hypothesis was put forward both by some proponents of Global Workspace theory (Dehaene et al., 2003), and on occasion by Crick and Koch (1995, 2003). It features all the same ingredients as the COI hypothesis for visual consciousness: visual cortex, prefrontal cortex, synchrony. There are nonetheless two important related differences between this hypothesis and the COI hypothesis. The first is that this hypothesis refers to our visual cortex, prefrontal cortex, and synchrony as only *part* of the story about the NCC. The second is that only the COI hypothesis offers the particular answer to the why question that it does. The two differences are related: given that the COI hypothesis appeals to prefrontal cortex activity and synchronization in order to recover what it takes to be the cognitive and phenomenological marks of consciousness, once these marks are recovered there is no need to postulate any further elements in the NCC.

5. Tests and testability

The concrete hypothesis that color consciousness involves synchronized V4 and dlPFC activity is straightforwardly testable. If an experimental condition could be created in which subjects report on a color experience despite lack of either (i) V4 activity, (ii) dlPFC activity, or (iii) synchronization of the two, then the hypothesis will have been falsified.

[31] Note, however, that if there is a separate higher-order representation of the audiovisual or somatosensory representations, or a separate higher-order representation of floor-level representations that are synchronized with the audiovisual and/or somatosensory representations, then (according to the present hypothesis) these audiovisual and/or somatosensory representations would be conscious after all.

In search of such falsification, a natural place to start is lesion studies. If color consciousness survives a lesion in V4 or the dlPFC, that would be evidence against the hypothesis. In fact, such lesion evidence exists for the parallel hypothesis in which the ACC is hypothesized to subserve higher-order representation. For subjects who have undergone cingulatomy are still conscious.

There is a methodological danger in using lesion studies exclusively, however (Chalmers, 2000). It is well known that some functions that are subserved by one brain area can be subserved by another after lesion. Thus a possible interpretation of post-cingulatomy consciousness may well be that, although the ACC performs monitoring functions in healthy subjects, the functions are recovered by another brain area if the ACC is incapacitated.

A safer method for testing the present hypothesis may be using transcranial magnetic stimulation (TMS) to inhibit activity in the dlPFC (or the ACC). If consciousness is no longer reported when dlPFC activity is sufficiently inhibited, that would constitute evidence in favor of the concrete hypothesis.

It is important to keep in mind, however, the distinction between the "general shape" of the NCC according to COI theory and the concrete hypothesis proposed in the previous section. If it is shown that consciousness remains unaltered despite inhibition of dlPFC activity, this would falsify the concrete hypothesis, but not the "general shape" claim just yet. One interpretation of this result that would be open to the COI theorist would be that s/he was wrong to designate the dlPFC as the seat of higher-order representation, but that the general idea that the NCC is constituted by the binding of floor-level representations with higher-order representations about them is correct. Another interpretation open to the COI theorist would be that the dlPFC is not the *only* seat of higher-order representations, and when it is incapacitated, another brain area can still produce such representations.

These considerations illustrate the way in which COI theory is more loosely connected to the empirical evidence than the concrete hypothesis offered in the previous section. Evidence that might falsify the concrete hypothesis would not automatically falsify COI theory, since COI theory as such is compatible with a number of different concrete hypotheses regarding the NCC. We have focused on the one that takes higher-order representation to be grounded in the dlPFC and functional integration to be effected by neural synchronization, but other views about the neural correlates of higher-order representation and functional integration would result in other versions of COI theory. For example, it was noted that there are at least seven different hypotheses about the neural correlate of higher-order representation that the COI theorist could advert to. Thus falsification of the previous section's concrete hypothesis would be falsification of just one version of COI theory, not falsification of the theory itself.

Nonetheless, COI theory as such is falsifiable as well. COI theory has a number of different versions, but a relatively limited number. If all these versions are falsified, then COI theory is thereby falsified as well. To falsify all versions would be to show that consciousness survives lack of activity in any area plausibly construed as responsible for higher-order representation or that it survives lack of any processes plausibly construed as functional integration. For example, if all seven hypotheses resulting from the seven different views regarding the neural correlate of higher-order representation were falsified, COI theory itself would be thereby falsified. Thus COI theory, despite being inspired by top-down considerations of cognitive and phenomenological features of consciousness, is nonetheless testable at the neuroscientific level.

Interestingly, COI theory can also be empirically distinguished from the HOT and HOM views discussed in Section 3. On these views, although conscious states are targeted by higher-order representations, they are not integrated with them. Thus HOT and HOM theories predict that there is no synchronization of the floor-level and higher-order representations involved in a conscious experience, whereas COI theory predicts that there is. This is an empirical prediction that distinguishes COI theory from some of the other theories inspired by the principle that conscious states are states we are aware of having.[32]

[32] How COI theory might be empirically distinguished from self-representationalism is a trickier issue that I bypass here.

6. Conclusion

This paper sketched a novel hypothesis regarding the neural correlate of consciousness. According to the Cross-Order Integration hypothesis, the neural correlate of a visual experience is synchronized activity in the visual cortex and (probably) the dorsolateral prefrontal cortex. The hypothesis resembles, in some respects, other hypotheses already in circulation (especially those put forward by Dehaene's group), and in any case appeals to elements that should be familiar from discussions of the NCC (synchrony, prefrontal cortex). But unlike other hypotheses, the COI hypothesis is motivated by top-down considerations. It starts out from consideration of the cognitive and phenomenological marks of consciousness, and then seeks the neural structures and processes that bear those marks. Such a top-down approach might be worth pursuing independently of the specific merits of the COI hypothesis.[33]

References

Armstrong, D. M. (1968). *A materialist theory of the mind*. New York: Humanities Press.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford and New York: Oxford University Press.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Science, 6*, 47–52.

Baars, B. J., Ramsøy, T. Z., & Laureys, S. (2003). Brain, conscious experience and the observing self. *Trends in Neurosciences, 26*, 671–675.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227–247.

Block, N. J., Flanagan, O., & Guzeldere, G. (Eds.). (1997). *The nature of consciousness: Philosophical debates*. Cambridge, MA: MIT Press.

Block, N. J., & Stalnaker, R. (1999). Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review, 108*, 1–46.

Botvinik, M. M., Braver, T. S., Carter, C. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Bressler, S. L. (1995). Large-scale cortical networks and cognition. *Brain Research, 20*, 288–304.

Carruthers, P. (2004). Suffering without subjectivity. *Philosophical Studies, 121*, 99–125.

Caston, V. (2002). Aristotle on consciousness. *Mind, 111*, 751–815.

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions*. Cambridge, MA: MIT Press.

Chalmers, D. J. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical perspectives* (pp. 220–272). Oxford and New York: Oxford University Press.

Chalmers, D. J., & Jackson, F. (2001). Conceptual analysis and reductive explanation. *Philosophical Review, 110*, 315–361.

Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., et al. (1999). In search of the self: A positron emission tomography study. *Psychological Science, 10*, 26–34.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2*, 263–275, Reprinted in Block, Flanagan, and Guzeldere (1997).

Crick, F., & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature, 375*, 121–123.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*, 119–126.

Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Science USA, 100*, 8520–8525.

Dretske, F. I. (1993). Conscious experience. *Mind, 102*, 263–283.

Engel, A. K., Fries, P., Konig, P., Brecht, M., & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition, 8*, 128–151.

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlate of sensory awareness. *Trends in Cognitive Science, 5*, 16–25.

Flashman, L. A., McAlister, T. W., Johnson, S. C., Rick, J. H., Green, R. L., & Saykin, A. J. (2001). Specific frontal lobe subregions correlated with unawareness of schizophrenia: A preliminary study. *Journal of Neuropsychiatry Clinical Neurosciences, 13*, 255–257.

Flohr, H. (1995). Sensations and brain processes. *Behavioral Brain Research, 71*, 157–161.

Frith, C. D., & Frith, U. (1999). Interacting minds–a biological basis. *Science, 286*, 1692–1695.

Gehring, W. J., & Knight, R. T. (2000). Prefrontal–cingulate interactions in action monitoring. *Nature Neuroscience, 3*, 516–520.

[33] For comments on a previous draft of this paper, I would like to thank Tim Bayne, David Chalmers, Ilya Farber, Rocco Gennaro, Antti Revonsuo, and two anonymous referees for *Consciousness & Cognition*. In working on this paper, I have benefited from financial support by the University of Arizona's Cognitive Science Program and the University of Sydney's SESQUI fellowship scheme. I have also benefited from presenting drafts of this paper on two occasions: at a conference titled *Neurophilosophy: The State of the Art* and at the University of Arizona's Center for Consciousness Studies discussion forum. I would like to thank the audiences there, in particular Farid Masrour, Jesse Prinz, and Logan Trujillo. There are many more people could thank for useful discussion of material central to this paper, but standing out are Frances Balcomb, Marie Banich, J. Alan Hobson, and Hakwan Lau.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Science USA, 98*, 4259–4264.

Johnson, S. C., Baxter, L., Wilder, L., Heiserman, J., Pipe, J., & Prigatano, G. (2002). Neural correlates of self-reflection. *Brain, 125*, 1808–1814.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*, 759–763.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Calgar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*, 785–794.

Kriegel, U. (2003a). Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy, 33*, 103–132.

Kriegel, U. (2003b). Consciousness, higher-order content, and the individuation of vehicles. *Syntheses, 134*, 477–504.

Kriegel, U. (2005). Naturalizing subjective character. *Philosophy and Phenomenological Research, 71*, 23–57.

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In Kriegel and Williford 2006 (pp. 143–170).

Kriegel, U., & Williford, K. W. (2006). *Self-representational approaches to consciousness*. Cambridge, MA: MIT Press.

Lane, R. D., Fink, G. R., Chau, P. M., & Dolan, R. J. (1997). Neural activation during selective attention to subjective emotional responses. *Neuro Report, 8*, 3969–3972.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Science USA, 103*, 18763–18768.

Levine, J. (2001). *Purple haze*. Oxford, New York: Oxford University Press.

Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

Lycan, W. G. (2001). A simple argument for a higher-order representation theory of consciousness. *Analysis, 61*, 3–4.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*, 1835–1838.

Metzinger, T. (1995). Faster than thought: Holism, homogeneity, and temporal coding. In T. Metzinger (Ed.), *Conscious experience*. Thorverton: Imprint Academic.

Milham, M. P., Banich, M. T., Claus, E. D., & Cohen, N. J. (2003). Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. *Neuro Image, 18*, 483–493.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., et al. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research, 12*, 467–473.

Povinelli, D. J. (2004). Behind the ape's appearance: Escaping anthropomorhism in the study of other minds. *Daedalus: Journal of the American Academy of Arts and Sciences, Winter*, 29–41.

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *Trends in Cognitive Science, 7*, 157–160.

Revonsuo, A. (1999). Binding and the phenomenal unity of consciousness. *Consciousness and Cognition, 8*, 173–185.

Roelfsema, P. R., Engel, A. K., Konig, P., & Singer, W. (1997). Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature, 385*, 157–161.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 94*, 329–359.

Rosenthal, D. M. (1990). A theory of consciousness. *ZiF technical report 40*, Bielfield, Germany. Reprinted in Block et al. (1997).

Rosenthal, D. M. (1993). Thinking that one thinks. In M. Davies & G. W. Humphreys (Eds.), *Consciousness: Psychological and philosophical essays*. Oxford: Blackwell.

Rosenthal, D. M. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind*. Oxford, New York: Oxford University Press.

Saharaie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C. R., & Brammer, M. J. (1997). Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proceedings of the National Academy of Science USA, 94*, 9406–9411.

Schmitz, T. W., Rowley, H. A., Kawahara, T. N., & Johnson, S. C. (2006). Neural correlates of self-evaluative accuracy after traumatic brain injury. *Neuropsychologia, 44*, 762–773.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: A critical evaluation of the temporal binding hypothesis. *Neuron, 24*, 67–77.

Singer, W. (1994). The organization of sensory motor representations in the neocortex: A hypothesis based on temporal coding. In C. Umlita & M. Moscovitch (Eds.), *Attention and performnace XV: Conscious and nonconscious information processing*. Cambridge, MA: MIT Press.

Smith, D. W. (1986). The structure of (self-)consciousness. *Topoi, 5*, 149–156.

Stopfer, M., Bhagavan, S., Smith, B. H., & Laurent, G. (1997). Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. *Nature, 390*, 70–74.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Treisman, A. M. (1996). The binding problem. *Current Opinions in Neurobiology, 6*, 171–178.

Treisman, A. M., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*, 107–141.

Van Gulick, R. (2001). Inward and upward—Reflection, introspection, and self-awareness. *Philosophical Topics, 28*, 275–305.

Van Gulick, R. (2006). Mirror, mirror—Is that all? In Kriegel and Williford 2006 (pp.11–39).

Varela, F., Lachaux, J.-P., Rodriguez, E., & Martinerie, J. (2006). Phase-synchronization and large-scale integration. *Nature Reviews Neuroscience, 2*, 229–239.

von der Malsburg, C. (1981). The correlation theory of brain function. *Technical report 81-2*, Max-Planck-Institute for Biophysical Chemistry, Gottingen.

Zeki, S., Watson, J. D. G., Lueck, C. J., Friston, K. J., Kennard, C., & Frackowiak, R. S. J. (1991). A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience, 11*, 641–649.