Correlates, causes and the neurobiology of consciousness

2nd Consciousness Online Conference

Joe Neisser Grinnell College 2/2010

For the past ten years it has been widely accepted among philosophers that neuroscientists are conducting a search for the *neural correlates of consciousness* (NCC).[1] David Chalmers (2000) articulated a definition of the neural correlate concept which was meant to capture the observed relation between the contents of experience and activity in specific neural populations. But this now-standard definition is inadequate both to the phenomena of consciousness and to the actual research it is meant to illuminate. In what follows I first introduce the idea of an NCC as presented by Chalmers, and rehearse a criticism of it advanced by Noe & Thompson (2004). I then suggest that the neurobiological research in view should be understood as an attempt to localize the working parts of a complex causal mechanism rather than as an effort to identify the correlates of some discrete conscious content. Thus, the first 'C' in 'NCC' should stand for 'causes' rather than 'correlates.' I conclude by attempting a revised definition of the NCC.

1. The standard approach to the NCC

David Chalmers (2000) describes a research program in neuroscience dedicated to the search for the *neural correlates of consciousness* or *NCC.* Chalmers defined an NCC as follows:

> *An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness.* (Chalmers, 2000, p.31)

Chalmers argues that this conception of the NCC is appropriate given what we know about consciousness and the brain, and that the goal of identifying NCCs can be attained in the foreseeable future (Chalmers, 2000, p.38).

Chalmers focuses on a particular kind of research aimed at a particular kind of consciousness, which he calls *content consciousness*. Content consciousness is primarily distinguished from *creature consciousness*, which is a property of a whole animal – that of *being* conscious, as opposed to being unconscious, as in a coma. The search for the NCC for content consciousness involves finding brain differences that correlate with consciousness of particular intentional objects, such as visual stimuli. This research compares states of the brain in which the subject is unconscious of the stimulus with states of the brain in which the stimulus is consciously experienced. The paradigmatic example of research that aims at discovering the NCC for content consciousness is a set of experiments conducted by Leopold & Logothetis (1996) and Sheinberg & Logothetis (1997).[2] I pause to describe this research.

1.i The research paradigm

Logothetis' primate laboratory was the site of a series of experiments that exploit the phenomenon of binocular rivalry. In binocular rivalry, competing stimuli are presented, one to each eye. The result in consciousness is an oscillation between perception of one stimulus and perception of the other. The subject does not perceive a hybrid composed of both stimuli, but instead experiences a flipping back and forth between the two. As the name 'rivalry' suggests, there is a sort of competition between neural processes, in which the 'winner takes all'. Thus, as one or the other of the visual stimuli

'dominates' neural activity, we become conscious of it. With the use of electroencephalographic recordings, Sheinberg & Logothetis (1997) first identified neurons in the visual system of a rhesus macaque subject that respond to particular visual stimuli, such as a horizontal grate or a sunburst. They then trained the macaque subject to report its visual experience by pushing a button upon seeing the preferred stimulus pattern. The stimuli were then presented in a binocular rivalry situation, so that the monkey would respond when the preferred stimulus became dominant. Thus, the experimenters had a way to compare neural activation across very specific states of content consciousness - one in which the preferred stimulus is not conscious because suppressed by the rivalry, and one in which it is the 'winner.'

The elegance of the design lies in its ability to isolate the consciousness of a stimulus independently of the presentation of the stimulus. Logothetis et al were able to measure the difference in brain state that co-varies with conscious perception of the stimulus even while stimulus presentation is held constant. With this experimental design, these researchers have found a way to work directly in the field that Fechner called *inner psychophysics*. Inner psychophysics is the empirical study of the relation between the conscious mind and neurophysiology. Fechner distinguished inner psychophysics from *outer psychophysics*, which is the study of the relation between the mind and external stimuli. Historically, inner psychophysics had to be studied indirectly through outer psychophysics, because neuropsychologists could not precisely manipulate and measure neural activity independently from external stimulation (Hilgard, 1987, Boring, 1950). The binocular rivalry paradigm, together with improved microelectrical recording techniques, allows limited but quite direct experimental manipulation of the inner psychophysical relation. The results indicate that there is significant relation between the 'winning' visual perception and activation of particular neural populations in the inferior temporal cortex (IT) and the superior temporal sulcus (STS). Whenever the monkey reported seeing the preferred stimulus, 90% of the pre-identified neurons recorded in IT and STS were strongly active.

However, this activity was almost extinguished when the monkey saw the ineffective stimulus (Sheinberg & Logothetis, 1997, p.3413). Thus, these seem to be the neurons that make the difference for consciousness of that visual content. Chalmers identified this work as being paradigmatic of the NCC research program, and attempted to derive his general definition of the neural correlate from it.

1.ii The neural correlate of what?

According to Chalmers' definition of the NCC, the neural subsystem N will be sufficient for the occurrence of the conscious state, under conditions C. The NCC will be activation in the smallest ('minimal') neural system such that, when you have the neural activity you have the content consciousness. Or again, the NCC will be the neural state that is 'minimally sufficient' for the corresponding conscious content (Chalmers, 2000, pp 24-5). Call this aspect of the definition of the NCC the *sufficiency requirement.* Since there might be more than one neural subsystem that correlates with a given type of conscious content, and since these multiple neural states might not correlate with one another, Chalmers holds that the notion of a neural correlate should be understood to refer to a neural state that suffices for a given conscious state but is not necessary for it (Chalmers, 2000, p.24).[3]

But the sufficiency requirement for neural correlates seems misplaced when considered from a phenomenological point of view. The ceteris paribus clause ("under conditions C") indicates that the minimal system is sufficient for the conscious content only on condition that everything else is in place *including the further neural conditions for consciousness itself* - creature consciousness, background consciousness, and other conscious contents to which it is related. Whenever these conditions obtain, N is held to be sufficient for 'an experience', the content of which is specified by the activation profiles of N. But phenomenologically, this is an abstraction. 'The' experience individuated in this way does not exist as such, but only as part of a global consciousness that clearly does *not* correlate with N.

Something is missing from the NCC if it doesn't actually correlate with any experience. Intuitively, one might construe the identified neural activity simply as a contributor in the genesis of a global conscious experience ultimately having as one component the particular visual feature to which these cells are attuned (say, horizontal orientation of the grate). But in that case these momentary activations in local neural populations in visual system, which the definition of the NCC is carefully constructed to pick out, are not sufficient for any conscious experience. What is missing? Consciousness of a particular visual content is always consciousness for some subject. This first-person point of view is part of the experience itself but is not part of the NCC as defined by Chalmers. In the next section I consider this problem in more detail.

2 Noe & Thompson's phenomenological critique of the NCC concept

I next consider a critique of Chalmers' NCC program, offered by Alva Noe and Evan Thompson (2004). Noe & Thompson argue that the contents of visual consciousness have certain global features that cannot be pinned to the firing rates of neurons in specific areas in the visual system. That is, they hold that the experience and the putative neural correlates do not have one and the same content (Noe & Thompson, 2004, p.13). Noe & Thompson list three features of experiential content not 'matched' in neuronal receptive field representations. Visual perception is: *structurally coherent, intrinsically experiential, and active* (Noe & Thompson, 2004, p.14). The authors conclude that the NCC program is a flawed approach to content consciousness.

2.i Intrinsic experientiality

I focus on Noe & Thompson's second objection to the notion of an NCC, that consciousness is intrinsically experiential.[4] What Noe & Thompson have in mind here is roughly what various philosophers have called *subjectivity* or the *first-person point of view*. They write: "Perceptual content is intrinsically *experiential*, in the sense that the content of an experience is always the content as represented from a point of view" (Noe & Thompson, 2004, p.16). The analysis of experience as essentially involving a subjective point of view has been a constant theme in the literature.[5] Whenever there is consciousness, there is something it-is-like, and this is always *for* some subject (Nagel, 1974).[6] Similarly, the subjective point of view may be said to be an internal structure of consciousness, in the sense that what-it-is-like is always experienced 'from the inside.' This means that the first-person perspective is partly constitutive of content consciousness itself. Thus, we cannot just stipulate, as Chalmers does, that we are studying content consciousness rather than creature consciousness since the latter is an internal structure of the former. Noe & Thompson clarify this point by explaining that their notion of 'experientiality' refers to the way in which experience is *perspectivally self-conscious*. This means that "the perceiver and the world enter into the content of experience as background conditions" (Noe & Thompson, 2004b, pp.90, 91fn). So the content of experience is always perspectival, and part of the experiential content is *that* it is perspectival. This perspectival content concerns various possibilities for action from this vantage point, as well as expectations about changes in the point of view that will result. In this way, subjective, first-person experience is apprehended as 'for me,' in a way such that 'I' am specified by the structure of the experience itself. So, when Noe & Thompson criticize the NCC program for not taking into account the 'intrinsically experiential' nature of conscious content, they are referring to the well-established phenomenological analysis of the subjectivity of first-person experience.

2.ii Inner psychophysics

Note, however, that Noe & Thompson also emphatically and repeatedly state that that their argument is not directed against neurobiological research on consciousness in general. They only object to Chalmers' assumption that the global contents of consciousness will match or correlate directly with particular receptive field contents in isolated neural populations (Noe & Thompson, 2004b, p.87). They call this faulty assumption the 'matching content' assumption. The matching content assumption is expressed in the definition of the NCC as the requirement that the correlate be *sufficient* for the contents of visual experience.[7] This sufficiency requirement was discussed in part one above.

Yet while Noe & Thompson's critique of Chalmers' definition of the NCC is warranted, the actual research remains utterly unscathed. The experimental design employed by Logothetis et al ingeniously identifies neural activity that co-varies with particular features of experience such as perceptual consciousness of the spatial orientation of a bar. As I pointed out above, the binocular rivalry design is essentially a classical experiment in psychophysics. But the sufficiency requirement (matching content doctrine), stipulated by Chalmers, is unnecessary. A more old-fashioned aim for research in 'inner psychophysics' will still do perfectly well: to discover neural activation patterns that *predict* specific *changes* in consciousness. Neuroscientist Pierre Gloor has argued that today's neuropsychological research methods are essentially continuous with those first devised by Hans Berger in the 1920's and '30s: "[The] experimental condition differs from the baseline condition only with respect to a well defined psychological function. In this way, those changes specifically related to that function can be isolated from those it shares with the baseline condition" (Gloor, 1994, p.256). In the experiments conducted by Logothetis et al, the subjectivity (or 'experientiality') of visual consciousness is clearly baseline. In the binocular rivalry situation, there is plenty of conscious experience going on beyond the changes measured. In all the experimental conditions, subjective point of view obtains, and is independent of the binocular rivalry. The 'well-defined psychological function' is the changing

perception of the preferred stimulus. The method isolates specific areas activated during a specific and measurable change in conscious visual perception. But this certainly does *not* mean that other areas are not involved in the global task.

3. The search for causes

So the problem with the NCC concept is that the contents of consciousness are always subjective. Even under conditions C, neural activity in IT and STS just is not sufficient for any complete visual experience. But this problem with the NCC concept need not be a problem with neurobiological research on consciousness per se. Logothetis et al have succeeded brilliantly in marking one particular neural difference between the two complete visual experiences. But there is no need to interpret this kind of study as attempting to identify a *contentful correlate* for any actual experience. So where does this leave us? In the final section I argue that the search for the NCC can be reconceptualized as the search for a highly specialized working part of a complex mechanism that governs global subjective consciousness.

3.i INUS conditions and the NCC

Mackie (1974, 1980) introduced the notion of an INUS condition in his analysis of casuality. The INUS condition is the *i*nsufficient but *n*ecessary component of an *u*nnecessary but *s*ufficient set of conditions. Thus, a lit match can be an INUS condition for a forest fire, despite the fact that it is neither necessary nor sufficient for the forest fire. In the explanation of the actual fire, the lit match plays a special role within the context of a set of conditions that are jointly sufficient for the fire. "What we call a cause typically is, and is recognized as being, only a partial cause; it is what *makes the difference* in

relation to some background or causal field" (1980 p.xi, emphasis added). Causal claims based on the INUS analysis are to be tested by constructing experimental conditions that manipulate the putative INUS condition in order to observe what difference this makes in the system ('field') as a whole.[8] And this is in fact a very precise description of the experiments here in view. As discussed above, Logothetis et al identified the neural activations that *make the difference to*, and hence are partial causes of, the experience of the preferred stimulus.

Note however, that the INUS condition cannot be the essence of the casual relation because it is symmetrical, while casual relations are not. So, on Mackie's analysis, it will also be true that B is the INUS condition for A - the forest fire is the INUS condition for the lighting of the match, as well as the reverse. In good Humean tradition, it appears that the best an INUS analysis can provide is correlation, not causation. But Mackie offered the INUS relation merely as an analysis of the *form* of a complex causal relation, describing part of an empirical method for 'hunting causes' (Cartwright, 2007). Mackie buttressed his analysis by adding that the INUS condition picks out a cause (and not a mere correlate) only insofar as it identifies an *effective* relation (or one in which A 'fixes' B, A 'produces' B, etc.). Roughly, A effects B when <u>preventing</u> A *prevents* B. Logothetis' binocular rivalry paradigm is an ingenious way of testing precisely this: the rivaling stimulus intermittently prevents the neural activation (A), and in turn this is shown to prevent the behavioral report (B).[9]

Seeing Logothetis et al.'s research as part of the hunt for causes renders it continuous with the rest of neuroscience.[10] Explanation in neuroscience generally takes the form of a search for mechanisms that underlie cognitive and behavioral phenomena, and these mechanistic explanations are offered on the basis of evidence about *manipulability* (Craver, 2007). This is very much in tune with the idea that the INUS condition identifies causes in terms of their effectiveness. Further, Craver argues that a neural mechanism offered in an explanation is often held to be (partly) *constitutive* of its explanandum. This

means that a putative causal mechanism can be explanatory when it is the INUS condition in a system that exhibits the explanandum. The explanandum itself can be a phenomenon observed at another level. Thus, the search for the NCC may be understood as a search for the neural parts of a complex mechanism - an organism - which is conscious. The global and subjective content of consciousness, then, is understood as constituted (enacted) at the complex, holistic level of the embodied animal, not at the level of the working neural mechanism that governs it.

One more important point must be clarified. I do not suggest that this research has the resources to bridge the so-called explanatory gap and solve the problem about qualia (the qualitative properties of conscious experience). I follow Levine (2001) in holding that we will not have a complete explanation of qualia until we have a functional analysis of what qualia do. But Levine also holds that we *can* in fact locate the causes of consciousness, even in the absence of a functional-role analysis of qualia. We can have - and in this case we *do* have - good empirical reason to believe that particular neural activations are what bring about particular features of experience. There is no conceptual problem in conducting this sort of investigation. What we are at a loss to do is to explain why these neural mechanisms (or any other) should give rise to the particular qualia they do (rather than some other other). Levine even argues that it is precisely our ability to localize the mechanisms of consciousness that motivates the absent qualia & conceivability arguments in the first place! The moment we identify c-fibers (say) as the mechanism of pain, we immediately find that we can imagine them working in the absence of qualia. So the explanatory gap argument is not a denial of the claim that c-fibers (say) are the causal mechanism for pain. Rather it is a denial of the claim that we understand why this should be the case, since we have no analysis that connects what c-fibers do with what pain feels like.

But in addition to our perplexity about the functional role of qualia, Noe & Thompson have shown that we cannot directly map abstracted bits of *experienced intentional content* to momentary

activation in isolated neural populations. Again, this is because actual experience is a holistic gestalt, not a patchwork of independent bits, some of them the non-subjective visual bits and *others* the subjective bits. There is no non-subjective *bit* of visual experience to which the activity in IT and STS could conceivably correspond. But there is a precisely measurable difference between the neural conditions that underlie two distinct, globally subjective moments of conscious experience. The research identifies the neural activity that *makes the difference* between these moments. This relation of 'making the difference for' is better analyzed in terms of the INUS condition than in terms of minimal sufficiency. In turn, the INUS relation is best understood as a causal relation, not as a merely correlative one. Thus, the research aims at localizing the working parts of an underlying mechanism, and in this respect it is continuous with much of the rest of neuroscience as well as the long tradition of psychophysical research.

In light of the above, I attempt the following rewrite of the definition of the NCC:

An NCC is a minimal neural system N such that states of N are the partial causes of a distinguishable state of consciousness, where a given state of N, as a working part of an embodied mechanism, is the INUS condition for the effected aspect of the conscious state.

Works Cited

Boring, E.G. (1950, [1929]). *A History of Experimental Psychology.* New York: Appleton-Century-Crofts.

Cartwright, N. (2007). *Hunting Causes and Using Them*. New York, NY: Cambridge University Press

Chalmers, D. (2000). What is a neural correlate of consciousness? In Metzinger (Ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions.* Cambridge, MA: MIT Press

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience.* Oxford & New York: Oxford University Press.

Dreyfus, H. (1992). *What Computers Still Can't Do.* Cambridge, MA: MIT Press.

Gloor, P. (1994). Is Berger's dream coming true? *Electroencephalography and Clinical Neurophysiology,* 90, 253-266.

Glymour, C. (1994). On the methods of cognitive neuropsychology. *British Journal for the Philosophy of Science,* 45: 815-45.

Hilgard, E. (1987). *Psychology in America: A historical survey.* New York & Orlando, FL: Harcourt Brace Jovanovich Publishers.

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach.* Englewood, CO: Roberts & Co.

Kriegel, U. (2006) Philosophical theories of consciousness: Contemporary Western perspectives. In *Cambridge Handbook of Consciousness* (Moscovitch, M, Thompson, E., & Zelato, P., Eds). New York: Cambridge University Press.

Leopold, D.A. & Logothetis, N.K. (1996). Activity Changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature,* 379, 549-553.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness.* New York: Oxford University Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Mackie, J.L. (1974, [1980]). *The Cement of the Universe: A study of Causation.* Oxford & New York: Clarendon Press.

Metzinger, T. (2004). Appearance is not knowledge: The incoherent straw man, content-content confusions and mindless conscious subjects. *Journal of Consciousness Studies,*11(1), 67-72.

Nagel, T. (1974). What is it like to be bat? *Philosophical Review,* 83, 435-450.

Noe, A. & Thompson, E. (2004a). Are there neural correlates of consciousness? *Journal of Consciousness Studies,*11(1), 3-28.

Noe, A. & Thompson, E. (2004b). Sorting out the neural basis of consciousness: Authors' reply to commentators. *Journal of Consciousness Studies,*11(1), 87-98.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press.

Sheinberg, D.L. & Logothetis, N.K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences,* 94, 3408-3413.

Spivey, M. (2007). *The Continuity of Mind.* New York: Oxford University Press.

Van Fraassen, B. (1980). *The Scientific Image.* Oxford & New York: Oxford University Press.

Vanderweele & Robins (2009). Minimal sufficient causation and directed acyclic graphs. *Annals of Statistics,* Vol. 37, No. 3, 1437–1465.

Notes

[1] Thanks to John Fennell for feedback on an earlier version of this paper.

[2] Christof Koch (2004) is another contemporary example of a neuroscientist researching content consciousness in the visual system.

[3] Significant complications arise when it is asked, under what conditions will the neural state suffice? Presumably, if neurons are excised from the brain, placed in a dish, and given electrochemical stimulation, their activity will not suffice for content consciousness. Here Chalmers resorts to the ceteris paribus clause *under conditions C.* But defining the exact range of cases in which a putative NCC ought to be sufficient for the corresponding conscious state presents a thicket of difficulties. After dedicating a sizable chunk of his paper to discussing this issue, Chalmers tentatively concludes that *". . . conditions C might be seen as conditions involving normal brain functioning, allowing unusual inputs and limited brain stimulation, but not lesions or other changes in architecture. Of course the precise nature of conditions C is still debatable. . . .But I think the conditions C proposed here are at least a reasonable first pass, pending further investigation"* (Chalmers, 2000, p.31). In part, the difficulty arises from the functional and developmental plasticity of the cortex. Brains with lesions or other structural alterations may find unique ways to function and may have unique NCC profiles. Thus, data from lesion studies must be interpreted very carefully, since observations of a heavily damaged brain may not identify activation that is sufficient in a more typical neural system. Also note that seeking the specification of the range of conditions under which a neural state correlates with a particular conscious content is very much like asking for a specification of the conditions under which a data structure in an information system bears its content. The answer is that a data representation bears its content in *that* system and in other, relevantly similar systems. Specification of the relevance parameters for the interpretation of information has, of course, confounded research in cognitive science and artificial intelligence for decades (Dreyfus, 1992). One motivation for rethinking the correlate idea is that much contemporary neuroscience also demands a rethinking of the classical conception of representation that it takes for granted. The internal temporal dynamics of the brain rarely fit the description of well-specified discrete representational contents. Instead, dynamic models of neural activity suggest that we conceive of a trajectory through state space as "a temporally drawn-out pattern of multiple 'representations' being simultaneously partially active" (Spivey, 2007, p.5).

[4] The three global features they discuss are closely related. Each is meant to show that conscious visual content appears as part of a larger structure of perception and action, as a particular figure against a larger background frame. Thus, visual consciousness is not simply a patchwork of independent atoms, each of which corresponds to some particular neural activation in visual cortex. This is a familiar phenomenological and gestaltist point, but none the worse for wear. See below.

[5] E.g. Metzinger (2000), Kriegel (2006). Levin (2001) has argued that the 'real' hard problem of consciousness is to understand how anything like a subjective point of view could arise in the world.

[6] Nagel (1974) characterized individual consciousness as *what-it-is-like for* that individual. Many writers focus on the 'what' part of Nagel's phrase. Accordingly, there is a large literature on *qualia*, in which, e.g. the greeneness of green is the topic of thought experiments about spectrum inversion and color-deprived neuroscientists. But the contents of consciousness are always *for* some subject, and this raises the problem that Noe & Thompson call 'experientiality'.

[7] In his comments on Noe & Thompson's focus piece, Metzinger defends the neural correlate idea by denying that it is committed to the matching content assumption. I agree that the neurobiological research itself is not committed to the matching content assumption. But Chalmers' definition remains committed to it in virtue of the sufficiency requirement. Without the matching content assumption, the correlation relation obtains only between global subjective states and global states of the nervous system (Metzinger 2004, p.71).

[8] The recent literature on causality distinguishes between counterfactual approaches, which use directed acyclic graphs, and mechanistic approaches which appeal to intervention and manipulation. Mackie's work is typically filed under the latter heading, as discussion will indicate. But Vanderweele & Robins (2009) argue that the two kinds of account can be integrated into a single framework. Space does not allow discussion here.

[9] Even with the addition of the criterion that the INUS condition must also be 'effective' in this way, it remains conceivable that B is in reality the *cause* of A. Here again the analysis actually fits the data very nicely. The

experiment itself cannot distinguish between two logical possibilities: that the neural state causes the conscious state, and that the conscious state causes the neural state.

[10] Note that I adopt Mackie's terminology only to emphasize clearly that the relation picked out by the experiments is a causal/mechanical one. The language of minimal sufficiency could certainly be retained, and it is commonly used in statistical models. But that usage does not typically carry the mechanistic meaning that is so central to explanation in neuroscience. cf. Craver (2007), Cartwright (2007), Pearl (2000), Glymour (1994), Lewis (1973). See also van Fraassen (1980) for more on the notion of an *effectual difference*.