What is a Neural Correlate of Consciousness?

David J. Chalmers

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

[[This paper was published in *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (T. Metzinger, ed), published with MIT Press in 2000. It was first presented at Association for the Scientific Study of Consciousness conference on "Neural Correlates of Consciousness" in Bremen, June 1998.]

1 Introduction

The search for neural correlates of consciousness (or NCCs) is arguably the cornerstone in the recent resurgence of the science of consciousness. The search poses many difficult empirical problems, but it seems to be tractable in principle, and some ingenious studies in recent years have led to considerable progress. A number of proposals have been put forward concerning the nature and location of neural correlates of consciousness. A few of these include:

- 40-hertz oscillations in the cerebral cortex (Crick and Koch 1990)
- Intralaminar nuclei in the thalamus (Bogen 1995)
- Re-entrant loops in thalamocortical systems (Edelman 1989)
- 40-hertz rhythmic activity in thalamocortical systems (Llinas et al 1994)
- Extended reticular-thalamic activation system (Newman and Baars 1993)
- Neural assemblies bound by NMDA (Flohr 1995)
- Certain neurochemical levels of activation (Hobson 1997)
- Certain neurons in inferior temporal cortex (Sheinberg and Logothetis 1997)
- Neurons in extrastriate visual cortex projecting to prefrontal areas (Crick and Koch 1995)
- Visual processing within the ventral stream (Milner and Goodale 1995)

(A longer list can be found in Chalmers 1998. Review articles on neural correlates of consciousness, especially visual consciousness, can be found in Crick and Koch 1998 and Milner 1995.)

As the full title of this book ("Neural Correlates of Consciousness: Empirical and Conceptual Issues") suggests, all this activity raises a number of difficult conceptual and foundational issues. I can see at least five sorts of foundational questions in the vicinity.

- (1) What do we mean by 'consciousness'?
- (2) What do we mean by 'neural correlate of consciousness'?
- (3) How can we find the neural correlate(s) of consciousness?
- (4) What will a neural correlate of consciousness explain?
- (5) Is consciousness reducible to its neural correlate(s)?

The first two questions here are conceptual questions, the third is an epistemological or methodological question, the fourth is an explanatory question, and the fifth is an ontological question. The first, fourth, and fifth are versions of general questions that philosophers have discussed for a long time (my own view on them is in Chalmers (1995; 1996)). The second and third questions are more specific to the NCC investigation. I have discussed the third question in Chalmers (1998). Here I want to focus on the second question.

What does it mean to be a neural correlate of consciousness? At first glance, the answer might seem to be so obvious that the question is hardly worth asking. An NCC is just a neural state that directly correlates with a conscious state, or which directly generates consciousness, or something like that. One has a simple image: when your NCC is active, perhaps, your consciousness turns on, and in a corresponding way. But a

moment's reflection suggests that the idea is not completely straightforward, and that the concept needs some clarification.

Here, I will attempt a little conceptual spadework in clarifying the concept of an NCC. I don't know that this is the deepest problem in the area, but it seems to me that if we are looking for an NCC, it makes sense to get clear on what we are looking for. On the way I will try to make contact with some of the empirical work in the area, and see what concept of NCC is at play in some of the central work in the field. I will also draw out some consequences for the methodology of empirical work in the search. Most of this is intended as a first step rather than a last word. Much of what I say will need to be refined, but I hope at least to draw attention to some interesting issues in the vicinity.

As a first pass, we can use the definition of a neural correlate of consciousness given in the program of the ASSC conference. This says a neural correlate of consciousness is a "specific system in the brain whose activity correlates directly with states of conscious experience". This yields something like the following:

> A neural system N is an NCC if the state of N correlates directly with states of consciousness.

There are at least two things to get clear on here. First, what are the relevant "states of consciousness"? Second, what does it mean for a neural state to "correlate directly" with states of consciousness? I'll look into both these things in turn.

2 States of consciousness

I will take it that the states of consciousness we are concerned with here are all states of subjective experience, or equivalently, states of phenomenal consciousness. But what sort of states are relevant? In the NCC literature, I can see a few different classes of state that are sometimes considered.

(i) *Being conscious*

The first option is that the states in question are just those of being conscious and of not being conscious. The corresponding notion of an NCC will be that of a neural system whose state directly correlates with whether a subject is conscious or not. If the NCC is in a particular state, the subject will be conscious. If the NCC is not in that state, the subject will not be conscious.

This is perhaps the idea that first comes to mind when we think about an NCC. We might think about it as the "neural correlate of creature consciousness", where creature consciousness is the property a creature has when it is conscious, and lacks when it is not conscious.

Although this is an interesting notion, it does not seem to capture the sort of NCC that most work in the area is aimed at. As we'll see, most current work is aimed at something more specific. There are some ideas that can be taken as aiming at least in part at this notion, though. For example, the ideas of Bogen (1995) about the intralaminar nucleus seem to be directed at least in part at this sort of NCC.

Examining current work, it's interesting to note that insofar as there is any consensus at all about the location of this sort of NCC, the dominant view seems to be that it should be in or around the thalamus, or at least that it should involve interactions between the thalamic and cortical systems in a central role. Penfield (1937) argued that "the indispensable substratum of consciousness" lies outside the cerebral cortex, and probably lies in the diencephalon (thalamus, hypothalamus, subthalamus, epithalamus). This theme has been taken up in recent years by Bogen, Newman and Baars (1993), and others.

(ii) *Background state of consciousness*

A related idea is that of the neural correlate of what we might call the *background state* of consciousness. A background state is an overall state of consciousness such as being awake, being asleep, dreaming, being under hypnosis, and so on. Exactly what counts as a background state is not entirely clear, as one can divide things up in a number of ways, and with coarser or finer grains, but presumably the class will include a range of normal and of "altered" states.

We can think of this as a slightly more fine-grained version of the previous idea. Creature consciousness is the most coarse-grained background state of consciousness: it is just the state of being conscious. Background states will usually be more fine-grained than this, but they still will not be defined in terms of specific contents or modalities.

A neural correlate of the background state of consciousness, then, will be a neural system N such that the state of N directly correlates with whether a subject is awake, dreaming, under hypnosis, and so on. If N is in state 1, the subject is awake; if N is in state 2, the subject is dreaming; if N is in state 3, the subject is under hypnosis; and so on.

It may well be that some of the thalamocortical proposals discussed above are intended as, or might be extended into proposals about this sort of NCC. A more direct example is given by Hobson's (1997) ideas about neurochemical levels of activation. Hobson holds that these levels can be grouped into a three-dimensional state-space, and that different regions in this space correspond to different overall states of consciousness: wakefulness, REM sleep, nonREM sleep, and so on. When chemical levels are in a particular region in this space, the subject will be awake; when in another region, the subject will be in REM sleep; and so on. On this reading, one might see the neurochemical system as an NCC of the sort characterized above, with the different regions in state space corresponding to correlates of the various specific background states.

(iii) *Contents of consciousness*

There is much more to consciousness than the mere state of being conscious, or the background state of consciousness. Arguably the most interesting states of consciousness are *specific* states of consciousness: the fine-grained states of subjective experience that one is in at any given time. Such states might include the experience of a particular visual image, of a particular sound pattern, of a detailed stream of conscious thought, and so on. A detailed visual experience, for example, might include the experience of certain shapes and colors in one's environment, of specific arrangements of objects, of various relative distances and depths, and so on.

Specific states like these are most often individuated by their *content*. Most conscious states seem to have some sort of specific representational content, representing the world as being one way or another. Much of the specific nature of a visual experience, for example, can be characterized in terms of content. A visual experience typically represents the world as containing various shapes and colors, as containing certain objects standing in certain spatial relations, and so on. If the experience is veridical, the world will be the way the experience represents it as being. If the experience is an illusion or is otherwise misleading, the world will be other than the experience represents it as being. But either way, it seems that visual experiences typically have detailed representational content. The same goes for experiences in other sensory modalities, and arguably for many or most nonsensory experiences as well.

Much of the most interesting work on NCCs is concerned with states like these. This is work on the neural correlates of the contents of consciousness. Much work on the neural correlates of visual consciousness has this character, for example. This work is not concerned merely with the neural states that determine that one *has* visual consciousness; it is concerned with the neural states that determine the specific contents of visual consciousness.

A nice example is supplied by the work of Logothetis and colleagues on the NCC of visual consciousness in monkeys (Logothetis and Schall 1989; Leopold and Logothetis 1996; Sheinberg and Logothetis 1997). In this work, a monkey is trained to press various bars when they are confronted with various sorts of images: horizontal and vertical gratings, for example, or gratings drifting left and right, or faces and sunbursts (I will use horizontal and vertical gratings for the purposes of illustration). After training is complete, the monkey is presented with two stimuli at once, one to each eye. In humans, this usually produces binocular rivalry, with alternating periods of experiencing a definite image, and occasional partial overlap. The monkey responds by pressing bars, in effect "telling" the experimenter what it is seeing: a horizontal grating, or a vertical grating, or an interlocking grid.

At the same time, neurons in the monkey's cortex are being monitored by electrodes. It is first established that certain neurons respond to certain stimuli: to horizontal lines, for example, or to flowers. Then these neurons are monitored in the binocular rivalry situation, to see how well they correlate with what the monkey seems to be seeing. It turns out that cells in primary visual cortex (V1) don't correlate well: when

the monkey is stimulated with horizontal and vertical gratings but "sees" horizontal, a large number of "vertical" cells in V1 fire, as well as "horizontal" cells. At this point, most cells seem to correlate with retinal stimulus, not with visual percept. But further into the visual system, the correlation increases, until in inferior temporal cortex, there is a very strong correlation. When the monkey is stimulated with horizontal and vertical grating but "sees" horizontal, almost all of the relevant horizontal cells in IT fire, and almost none of the vertical cells do. When the monkey's response switches, indicating that it is now "seeing" vertical, the cell response switches accordingly.

These results lend themselves naturally to speculation about the location of a visual NCC. It seems that V1 is unlikely to be or involve an NCC, for example, due to the failure of V1 cells to correlate with the contents of consciousness. Of course there are still the possibilities that some small subset of V1 is an NCC, or that V1 is a neural correlate of some aspects of visual consciousness but not of others, but I leave those aside for now. On the other hand, IT seems to be a natural candidate for the location of an NCC, due to the strong correlation of its cells with the content of consciousness. At least it is natural to suppose that IT is a "lower bound" on the location of a visual NCC (due to the failure of strong correlation before then), though the NCC itself may be further in. None of this evidence is conclusive (and Logothetis and colleagues are appropriately cautious), but it is at least suggestive.

It is clear that this work is concerned with the neural correlates of the *contents* of visual consciousness. We are interested in finding cortical areas whose neural activity correlates with and predicts specific contents of consciousness, such as experiences of horizontal or vertical lines, or of flowers or sunbursts. The ideal is to find a neural system from whose activity we might determine the precise contents of a visual experience, or at least of its contents in certain respects (shape, color, and the like).

Interestingly, it seems that in doing this we are crucially concerned with the the representational contents of the neural systems themselves. In the Logothetis work, for example, it is important to determine the receptive fields of the cells (whether they respond to horizontal or vertical gratings, for example), in order to see whether the receptive fields of active cells matches up with the apparent contents of visual consciousness. In essence, the receptive field is acting at least as a heuristic way of getting at representational content in the neurons in question. Then, the crucial question is whether the representational content in the neural system matches up with the representational content in visual consciousness.

This suggests a natural definition of a neural correlate of the contents of consciousness.

> A neural correlate of the contents of consciousness is a neural representational system N such that representation of a content in N directly correlates with representation of that content in consciousness.

Or, more briefly:

> A content NCC is a neural representational system N such that the content of N directly correlates with the content of consciousness.

For example, the Logothetis work lends itself to the speculation that IT might contain a content NCC for visual consciousness, since the content of cells in IT seems to directly correlate (at least in these experiments) with the contents of visual consciousness. (Much more investigation is required to see whether this correlation holds across the board, of course.)

This definition requires that we have some way of defining the representational content of a neural system independent of the contents of consciousness. There are various ways to do this. Using a cell's receptive field to define its representational content is probably the simplest. A more refined definition might also give a role to a system's projective field, and the sort of behavior that activity in that system typically leads to. And there may be more complex notions of representational content still, based on complex correlations with environment, patterns of behavior, and activity in other cells. But even a crude definition of representational content (e.g., the receptive field definition) is good enough for many purposes, and can yield informative results about the visual NCC.

It's arguable that much work on the visual NCC tacitly invokes this sort of definition. Another example is Milner and Goodale's work on the two pathways of visual perception. They suggest that the ventral stream

is largely for cognitive identification and decision, while the dorsal stream is largely for online motor response, and the visual consciousness correlates with activity in the ventral stream.

Much of the support for this work lies with patients who have dissociations between specific contents of conscious perception and the contents involved in motor response. For example, a subject with visual form agnosia (e.g. Milner and Goodale's patient D.F.) cannot consciously identify a vertical slot, but can "post" an envelope through it without problem; while subjects with optic ataxia (e.g. those with Balint's (1909) syndrome) can identify an object but cannot act appropriately toward it. The dissociations here appear to go along with damage to the ventral and dorsal pathways respectively.

What seems to be going on, on a natural interpretation of these results and of Milner and Goodale's hypothesis, is that for these subjects, there is a dissociation between the contents represented in the ventral pathways and those represented in the dorsal pathway. In these cases, the character of a motor response appears to be determined by the contents represented in the dorsal pathway, but the character of conscious perception appears to be determined by the contents represented in the ventral pathway.

Thus one can see Milner and Goodale's hypothesis as involving the suggestion that the ventral stream contains the neural correlates of the contents of visual consciousness. The hypothesis is quite speculative, of course (though it is interesting to note that IT lies in the ventral stream), but it seems that the content-based analysis provides a natural interpretation of what the hypothesis is implicitly claiming regarding the visual NCC, and of what may follow if the hypothesis turns out to be correct.

One could give a similar analysis of much or most work on the visual NCC. When Crick and Koch (1998) propose that the visual NCC lies outside V1, for example, much of the experimental evidence they appeal to involves cases where some content is represented in consciousness but not in V1, or vice versa. For example, Gur and Snodderly (1997) show that for some quickly alternating isoluminant color stimuli, color cells in V1 flicker back and forth even though a single fused color is consciously perceived. And results by He et al (1996) suggest that orientation of a grating can fade from consciousness even though orientation cells in V1 carry the information. The results are not entirely conclusive, but they suggest a mismatch between the representational content in V1 and the content of consciousness.

One can apply this sort of analysis equally to NCCs in other sensory modalities. An NCC of auditory consciousness, for example, might be defined as a neural representational system whose contents correlate directly with the contents of auditory consciousness: loudness, direction, pitch, tone, and the like. The idea can arguably apply to defining the neural correlates of bodily sensations, of conscious mental imagery, and perhaps of conscious emotion and of the stream of conscious thought. All these aspects of consciousness can be naturally analyzed (at least in part) in terms of their content. In looking for their respective NCCs, we may ultimately be looking for neural systems whose content correlates with the contents of these aspects of consciousness.

(iv) *Arbitrary phenomenal properties*

(This section is a little more technical than those above, and might be skipped by those not interested in philosophical details.)

One might try to give a general definition of an NCC of various states of consciousness, of which each of the above would be a special case. To do this, one would need a general way of thinking about arbitrary states of consciousness. Perhaps the best way is to think in terms of arbitrary *phenomenal properties*. For any distinctive kind of conscious experience, there will be a corresponding phenomenal property: in essence the property of having a conscious experience of that kind. For example, being in a hypnotic state of consciousness is a phenomenal property; having a visual experience of a horizontal line is a phenomenal property; feeling intense happiness is a phenomenal property; feeling a throbbing pain is a phenomenal property; being conscious is a phenomenal property. Phenomenal properties can be as coarse-grained or as fine-grained as you like, as long as they are wholly determined by the current conscious state of the subject.

With this notion in hand, one might try to define the neural correlate of an arbitrary phenomenal property P.

A state N1 of system N is a neural correlate of phenomenal property P if N's being in N1

> directly correlates with the subject having P.

Note that we here talk of a *state* being an NCC. Given a *specific* phenomenal property - experiencing a horizontal line, for example, it is no longer clear that it makes sense to speak of a given system being the NCC of that property. Rather, it will be a particular state of that system. Neural firing in certain horizontal cells in IT (say) might be a neural correlate of seeing a horizontal line, for example; and having one's neurochemical system in a certain region of state space might be a neural correlate of waking consciousness, on Hobson's hypothesis. These are specific states of the neural systems in question.

Most of the time, we are not concerned with neural correlates of single phenomenal properties, but of *families* of phenomenal properties. Hobson is concerned not just with the neural correlate of waking consciousness, for example, but with the neural correlate of the whole family of background states of consciousness. Work on the visual NCC is not concerned with just the neural correlate of horizontal experience, but with the neural correlates of the whole system of visual experiential contents.

We might say a *phenomenal family* is a set of mutually exclusive phenomenal properties that jointly partition the space of conscious experiences, or at least some subset of that space. That is, any subject having an experience (of a certain relevant kind) will have a phenomenal property in the family, and will not have more than one such property. Specific contents of visual consciousness make for a phenomenal family, for example: any visually conscious subject will have some specific visual content, and they will not have two contents at once (given that we are talking about *overall* visual content). The same goes for contents at a particular location in the visual field: anyone with an experience as of a certain location will have some specific content associated with that location (a red horizontal line, say), and not more than one. (Ambiguous experiences are not counterexamples here, as long as we include ambiguous contents as members of the family in question.) The same again goes for color experience at any given location: there will be a phenomenal family (one property for each color quality) for any such location. And the same goes for background states of consciousness. All these sets of phenomenal properties make phenomenal families.

We can then say:

> A neural correlate of a phenomenal family S is a neural system N such that the state of N directly correlates with the subject's phenomenal property in S.

For any phenomenal family S, a subject will have at most one property in S (one background state, or one overall state of visual consciousness, or one color quality at a location). Neural system N will be an NCC of S when there are a corresponding number of states of N, one for every property in P, such that N's being in a given state directly correlates with the subject's having the corresponding phenomenal property. This template can be seen to apply to most of the definitions given above.

For the neural correlate of creature consciousness, we have a simple phenomenal family with two properties: being conscious and not being conscious. An NCC here will be a system with two states that correlate with these two properties.

For the neural correlate of a background state of consciousness, we have a phenomenal family with a few more properties: dreaming, being in an ordinary waking state, being under hypnosis, etc. An NCC here will be a neural system with a few states that correlate directly with these properties. Hobson's neurochemical system would be an example.

For the neural correlate of contents of consciousness, one will have a much more complex phenomenal family (overall states of visual consciousness, or states of color consciousness at a location, or particular conscious occurrent thoughts), and a neural representational system to match. The state of the NCC will directly correlate with the specific phenomenal property.

Notice that in the content case, there is an extra strong requirement on the NCC. In the other cases, we have accepted an arbitrary match of neural states to phenomenal states - any state can serve as the neural correlate of a dreaming state of background consciousness, for example. But where content is concerned, not any neural state will do. We require that the *content* of the neural state in question must match the content of consciousness. This is a much stronger requirement.

It is arguable that this requirement delivers much greater explanatory and predictive power in the case of neural correlates of conscious content. The systematicity in the correlation means that it can be extended to predict the presence or absence of phenomenal features that may not have been present in the initial empirical data set, for example. And it also will dovetail more nicely with finding a mechanism and a functional role for the NCC that matches the role that we associate with a given conscious state.

It is this systematicity in the correlation that makes the current work on neural correlate of visual consciousness particularly interesting. Without it, things would be much more untidy. Imagine that we find arbitrary neural states that correlated directly with the experience of horizontal lines, etc, such that there was no corresponding representational content in the neural state. Instead, we match seemingly arbitrary states N1 with horizontal, N2 with vertical, and so on. Would we count this as a neural correlate of the contents of visual consciousness? If we did, it would be in a much weaker sense, and in a way that would lead to much less explanatory and predictive power.

One might then hope to extend this sort of systematicity to other, non-content-involving phenomenal families. For example, one might find among background states of consciousness some pattern, or some dimension along which they systematically vary (some sort of intensity dimension, for example, or a measure of alertness). If we could then find a neural system whose states do not just arbitrarily correlate with the phenomenal states in question, but which vary along a corresponding systematic dimension, then the NCC in question will have much greater potential explanatory and predictive power. So this sort of systematicity in phenomenal families is something that we should look for, and something that we should look to match in potential neural correlates.

Perhaps one could define a "systematic NCC" as a neural correlate of a phenomenal family such that states correlate with each other in some such systematic way. I will not try to give a general abstract definition here, as things are getting complex enough already, but I think one can see a glimmer of how it might go. I will, however, keep using the case of neural correlate of the contents of consciousness (especially visual consciousness) as the paradigmatic example of an NCC, precisely because its definition builds in this such a notion of systematicity, with the corresponding explanatory and predictive power.

3 Direct correlation

The other thing that we need to clarify is the notion of "direct correlation". We have said that an NCC is a system whose state directly correlates with a state of consciousness, but what does direct correlation involve, exactly? Is it required that the neural system be necessary and sufficient for consciousness, for example, or merely sufficient? And over what range of cases must the correlation obtain for the system to count as an NCC? Any possible case? A relevantly constrained set of cases? And so on.

The paradigmatic case will involve a neural system N with states that correlate with states of consciousness. So we can say that

> state of N -?- state of consciousness

and specifically

> N is in state N1 -?- subject has conscious state C.

In the case of the contents of consciousness, we have a system N such that representing a content in N directly correlates with representation in consciousness. So we can say

> representing C in N -?- representing C in consciousness.

The question in all these cases concerns the nature of the required relation. How strong a relation is required here for N to be a NCC?

(A) *Necessity, sufficiency?*

The first question is whether the NCC state is required to be necessary and sufficient for the conscious state, merely sufficient, or something else in the vicinity.

(A1) *Necessity and sufficiency*. The first possibility is that the state of N is necessary and sufficient for the corresponding state of consciousness. This is an attractive requirement for an NCC, but it is arguably too strong. It might turn out that there is more than one neural correlate of a given conscious state. For example, it may be there there are two systems, M and N, such that a certain state of M suffices for being in pain and a certain state of N also suffices for being in pain, where these two states are not themselves always correlated. In this case, it seems that we would likely say that both M and N (or their corresponding states) are neural correlates of pain. But it is not the case that activity in M is necessary and sufficient for pain (as it is not necessary), and the same goes for N. If both M and N are to count as NCCs here, we cannot require an NCC to be necessary and sufficient.

(A2) *Sufficiency*. From the above, it seems plausible that we require only that an NCC state be *sufficient* for the corresponding state of consciousness, not necessary. But is any sufficient state enough? If it is, then it seems that the whole brain will count as an NCC of any state of consciousness. The whole brain will count as an NCC of pain, for example, since being in a certain total state of the whole brain will suffice for being in pain. Perhaps there is some very weak sense in which this makes sense, but it does not seem to capture what researchers in the field are after when looking for an NCC. So something more than mere sufficiency is required.

(A3) *Minimal sufficiency*. The trouble with requiring mere sufficiency, intuitively, is that it allows irrelevant processes into a NCC. If N is an NCC, then the system obtained by conjoining N with a neighboring system M will also qualify as an NCC by the previous definition, since the state of N+M will suffice for the relevant states of consciousness.

The obvious remedy is to require that an NCC has to be a *minimal sufficient system*: that is, a *minimal* system whose state is sufficient for the corresponding conscious state. By this definition, N will be an NCC when (1) the states of N suffice for the corresponding states of consciousness, and (2) no proper part M of N is such that the states of M suffice for the corresponding states of consciousness. In this way, we pare down any potential NCC to its core: any irrelevant material will be whittled away, and an NCC will be required to contain only the core processes that suffice for the conscious state in question.

Note that on this definition, there may be more than one NCC for a given conscious state. It may be that there is more than one minimal sufficient system for a given state, and both of these will count as a neural correlate of that state. The same goes for systems of phenomenal states. This seems to be the right result: we cannot know a priori that there will be only one NCC for a given state or system of states. Whether there will actually be one or more than one for any given state, however, is something that can be determined only empirically.

There is a technical problem for the minimality requirement. It may turn out that there is significant redundancy in a neural correlate of consciousness, such that for example a given conscious visual content is represented redundantly in many cells in a given area. If this is so, then that visual area as a whole might not qualify as a minimal sufficient system, as various smaller components of it might all themselves correlate with the conscious state. In this case the definition above would imply that various such small components would each be an NCC. One could deal with this sort of case by noting that the problem arises only when the states of the various smaller systems are themselves wholly correlated with each other. (If their mutual correlation can be broken, so will their correlation with consciousness, so that the overall system or some key subsystem will again emerge as the true NCC). Given this, one could stipulate that where states of minimal sufficient systems are wholly correlated with each other, it is the union of the system that should be regarded as an NCC, rather than the individual systems. So an NCC would be a minimal system whose state is sufficient for a given conscious state and whose state is not wholly correlated with the state of any other system. I will pass over this complication in what follows, however.

(B) *What range of cases?*

An NCC will be a minimal neural system N such that the state of N is sufficient for a corresponding conscious state C. This is to say: if the system is in state N1, the subject will have conscious state C. But the question now arises: over what range of cases must the correlation in question hold?

There is sometimes a temptation to say that this question does not need to be answered: all that is required is to say that *in this very case*, neural state N1 suffices for or correlates with conscious state C. But this does not really make sense. There is no such thing as a single-case correlation. Correlation is always

defined with respect to a range of cases. The same goes for sufficiency. To say that neural state N1 suffices for conscious state C is to say that in a range of cases, neural state N1 will always be accompanied by conscious state C. But what is the range of cases?

(B1) *Any possible case*. It is momentarily tempting to suggest that the correlation should range across any possible case: if N is an NCC it should be impossible to be in a relevant state of N without being in the corresponding state of consciousness. But a moment's reflection suggests that this is incompatible with the common usage in the field. NCCs are often supposed to be relatively limited systems, such as the inferior temporal cortex or the intralaminar nucleus. But nobody (or almost nobody) holds that if one excises the entire inferior temporal cortex or intralaminar nucleus and puts it in a jar, and puts the system into a relevant state, it will be accompanied by the corresponding state of consciousness.

That is to say, for a given NCC, it certainly seems *possible* that one can have the NCC state without the corresponding conscious state, for example by performing sufficiently radical lesions. So we cannot require that the correlation range over all possible cases.

Of course, one could always insist that a *true* NCC must be such that it is impossible to have the NCC state without the corresponding conscious state. The consequence of this would be that an NCC would almost certainly be far larger than it is on any current hypothesis, as we would have to build in a large amount of the brain to make sure that all the background conditions are in place. Perhaps it would be some sort of wide-ranging although skeletal brain state, involving aspects of processes from a number of regions of the brain. This might be a valid usage, but it is clear that this is not what researchers in the field are getting at when they are talking about an NCC.

We might call the notion just defined a *total* NCC, as it builds in the totality of physical processes that are absolutely required for a given conscious state. The notion that is current in the field is more akin to that of a *core* NCC. (I adapt this terminology from Shoemaker's (1979) notion of a "total realization" and a "core realization" of a functional mental state.) A total NCC builds in everything and thus automatically suffices for the corresponding conscious states. A core NCC, on the other hand, contains only the "core" processes that correlate with consciousness. The rest of the total NCC will be relegated to some sort of background conditions, required for the correct functioning of the core.

(Philosophical note: The sort of possibility being considered here is natural or nomological possibility, or possibility compatible with the laws of nature. If we required correlation across all *logically* possible cases, there might be no total NCC at all, as it is arguably logically possible (or coherently conceivable) to instantiate any physical process at all without consciousness. But it is probably not naturally possible. It is almost certainly naturally necessary that a being with my brain state will have the same sort as conscious state as me, for example. So natural possibility and necessity is the relevant sort for defining the correlation here.)

The question is then how to distinguish the core from the background. It seems that what is required for an NCC (in the "core" sense) is not that it correlate with consciousness across any possible conditions, but rather that it correlate across some constrained range of cases in which some aspects of normal brain functioning are held constant. The question then becomes, what is to be held constant? Across just what constrained range of cases do we require than an NCC correlate with consciousness?

(B2) *Ordinary functioning brain in ordinary environments*.

One might take the moral of the above to be that one cannot require an NCC to correlate with consciousness in "unnatural" cases. What matters is that the NCC correlates with consciousness is "natural" cases, those that actually occur in the functioning of a normal brain. the most conservative strategy would be to require correlation only across cases involving a normally functioning brain in a normal environment, receiving "ecologically valid" inputs of the sort received in a normal life.

The trouble with this criterion is that it seems too weak to narrow down the NCC. It may turn out that this way, we find NCCs at all stages of the visual system, for example. In normal visual environment, we can expect that the contents of visual systems from V1 through IT will all correlate with the contents of visual consciousness, and that even the contents of the retina will to some extent. The reason is that in normal cases all these will be linked in a straightforward causal chain, and the systems in question will not be dissociated. But it seems wrong to say that merely because of this, all the systems (perhaps even the retina)

should count as an NCC.

The moral of this is that we need a more fine-grained criterion to dissociate these systems and to distinguish the core NCC from processes that are merely causally linked to it. To do this, we are have to require correlation across a range of *unusual cases* as well as across normal cases, as it is these cases that yield interesting dissociations.

(B3) *Normal brain, unusual inputs*. The next most conservative suggestion is that we still require a normal brain for our range of cases, but that we allow any possible inputs, including "ecologically invalid" inputs. This would cover the Logothetis experiments, for example. The inputs that evoke binocular rivalry are certainly unusual, and not encountered in a normal environment. But it is precisely these that allow the experiments to make more fine-grained distinctions than we normally can. The experiments suggest that IT is more likely than V1 to be an NCC precisely because it correlates with consciousness across the wider range of cases. If states of V1 truly do not match up with states of consciousness in this situation, then it seems that V1 cannot be an NCC. If that reasoning is correct, then it seems that we require an NCC to correlate with consciousness across all unusual inputs, and not just across normal environments.

The extension of the correlation requirement from normal environments to unusual inputs is a relatively "safe" extension and seems a reasonable requirement, though those who place a high premium on ecological validity might contest it. But it is arguable that this is still too weak to do the fine-grained work in distinguishing an NCC from systems linked to it. Presumably unusual inputs will go only so far in yielding interesting dissociations, and some systems (particularly those well down the processing pathway) may well stay associated on any unusual inputs. So it is arguable that we will need more fine-grained tools to distinguish the NCC.

(B4) *Normal brain, vary brain stimulation*. The next possibility is to allow cases involving not just unusual inputs, but involving direct stimulation of the brain. Such direct stimulation might include both electrode stimulation and transcranial magnetic stimulation. On this view, we will require that an NCC correlates with consciousness across all cases of brain stimulation, as well as normal functioning. So if we have a potential NCC state that does not correlate with consciousness in the right way in a brain stimulation condition, that state will not be a true NCC.

This requirement seems to fit some methods used in the field. Penfield (e.g. Penfield and Rasmussen 1950) pioneered the use of brain stimulation to draw conclusions about the neural bases of consciousness. Libet (1982) has also used brain stimulation to good effect, and more recently Newsome and colleagues (e.g. Salzman et al 1990) have used brain stimulation to draw some conclusions about neural correlates of motion perception in monkeys. (See also Marge 1991 for a review of transcranial magnetic stimulation in vision.)

Brain stimulation can clearly be used to produce dissociations more fine-grained than can be produced merely with unusual inputs. One might be able to dissociate activity in any system from that in a preceding system by stimulating that system directly, for example, as long as there are not too many backwards connections. Given a candidate NCC - inferior temporal cortex, say - one can test the hypothesis by stimulating an area immediately following the candidate in the processing pathway. If that yields a relevant conscious state without relevant activity in IT (say), that indicates that IT is probably not a true NCC after all. Rather, the NCC may lie in a system further down the processing chain. (I leave aside the possibility that there might be two NCCs at different stages of the chain.)

This reasoning seems sound, suggesting that we may tacitly require an NCC to correlate with consciousness across brain stimulation conditions. There is no immediately obvious problem with the requirement, at least when the stimulation in question is relatively small and localized. If one allows arbitrary large stimulation, there may be problems. For example, one could presumably use brain stimulation at least in principle to disable large areas of the brain (by overstimulating those areas, for example) while leaving NCC activity intact. In this case, it is not implausible to expect that one will have the relevant NCC activity without the usual conscious state (just as in the case where one lesions the whole NCC and puts it in a jar), so the correlation will fail in this case. But intuitively, this does not seem to disprove the claim that the NCC in question is a true NCC, at least before the stimulation. If so, then we cannot allow unlimited brain stimulation in the range of cases relevant to the correlation; and more generally, some of the problems for lesions (discussed below)may apply to reasoning involving brain stimulation. Nevertheless, one might well require than an NCC correlate with consciousness at least across

cases of limited stimulation, in the absence of strong reason to believe otherwise.

(B5) *Abnormal functioning, due to lesions*. In almost all of the cases above, we have retained a normally functioning brain; we have just stimulated it in unusual ways. The next logical step is to allow cases where the brain is not functioning normally, due to lesions in brain systems. Such lesions might be either natural (e.g. due to some sort of brain damage) or artificial (e.g. induced by surgery). On the latest view, we will require that an NCC correlates with states of consciousness not just over cases of normal functioning, but over cases of abnormal functioning as well.

This certainly squares with common practice in the field. Lesion studies are often used to draw conclusions about the neural correlates of consciousness. In Milner and Goodale's work, for example, the fact that consciousness remains much the same upon lesions to the dorsal stream but not to the ventral stream is used to support the conclusion that the NCC lies within the ventral stream. More generally, it is often assumed that if some aspect of consciousness survives relatively intact with a given brain area is damaged, then that brain area is unlikely to be or contain an NCC.

The tacit premise in this research is that an NCC should correlate with consciousness but just in cases of normal functioning but in cases of abnormal functioning as well. Given this premise, it follows that if we find an abnormal case in which neural system N is damaged but a previously corresponding conscious state C is preserved, then N is not a neural correlate of C. Without this premise, or a version of it, it is not clear that any such conclusion can be drawn from lesion studies.

The premise may sound reasonable, but we already have reason to be suspicious of it. We know that for any candidate NCC, sufficiently radical changes can destroy the correlation. Preserving merely system N, cut off from the rest of the brain, for example, is unlikely to yield a corresponding conscious state; but intuitively, this does not imply that N was not an NCC in the original case.

Less radically, one can imagine placing lesions immediately downstream from a candidate NCC N, so that N's effects on the rest of the brain are significantly reduced. In such a case, it is probable that N can be active without the usual behavioral effects associated with consciousness, and quite plausibly without consciousness itself. It's not implausible that an NCC supports consciousness largely in virtue of playing the right functional role in the brain; by virtue of mediating global availability, for example (see Baars 1988 and Chalmers 1998). If so, then if the system is changed so that the NCC no longer plays that functional role, then NCC activity will no longer correlate with consciousness. But the mere fact that correlation can be destroyed by this sort of lesion does not obviously imply that N is not an NCC in a normal brain. If that inference could be made, then almost any candidate NCC could be ruled out by the right sort of lesion.

It may be that even smaller lesions can destroy a correlation in this way. For example, it is not implausible that for any candidate NCC N, there is some other local system in the brain (perhaps a downstream area) whose proper functioning is required for activity in the N to yield the usual effects that go with consciousness, and for N to yield consciousness itself. This second system might not itself be an NCC in any intuitive sense; it might merely play an enabling role, in the way that proper functioning of the heart plays an enabling role for functioning of the brain. If so, then if one lesions this single area downstream, then activity in the N will no longer correlate with consciousness. In this way, any potential NCC might be ruled out by a localized lesion elsewhere.

The trouble is that lesions change the architecture of the brain, and its quite possible that changes to brain architecture can change the very location of an NCC, so that a physical state that was an NCC in a normal brain will not be an NCC in the altered brain. Given this possibility, it seems too strong to require that an NCC correlate with consciousness across arbitrary lesions and changes in brain functioning. We should expect an NCC to be architecture-dependent, not architecture-independent.

So an NCC should not be expected to correlate with consciousness across arbitrary lesion cases. There are now two alternatives. Either we can require correlation across some more restricted range of lesion cases, or one can drop the requirement of correlation in abnormal cases altogether.

For the first alternative to work, we would have to find some way to distinguish a class of "good" lesions from the class of "bad" lesions. An NCC would be expected to correlate with consciousness across the good lesions but not the bad lesions. If one found a "good" lesion case where activity in system N was

present without the corresponding consciousness state, this would imply that N is not an NCC; but no such conclusion could be drawn from a "bad" lesion case.

The trouble is that it is not at all obvious that such a distinction can be drawn. It might be tempting to come up with an after-the-fact distinction, defined as the range of lesions in which correlation with an given NCC N is preserved, but this will not be helpful at all, as we are interested in precisely the criterion that makes N qualify as an NCC in the first place. So a distinction will have to be drawn on relatively a priori grounds (it can then be used to determined whether a given correlation-pattern qualifies an arbitrary system as an NCC or not). But it is not clear how to draw the distinction. One might suggest that correlation should be preserved across small lesions but not large ones; but we have seen above that even small lesions might destroy a potential NCC. Or one might suggest that lesions in downstream areas are illegitimate, but upstream and parallel lesions are legitimate. But even here, it is not clear that indirect interaction with an upstream or parallel area might be required to support proper functioning of an NCC. Perhaps with some ingenuity one might be able to come up with a criterion, but it is not at all obvious how.

The second alternative is to hold that correlation across cases of normal functioning (perhaps with unusual inputs and brain stimulation) is all that is required to be an NCC. If this is so, one can never infer directly from the fact that N fails to correlate with consciousness in a lesion case to the conclusion that N is not an NCC. On this view, the location of an NCC is wholly architecture-dependent, or entirely dependent on the normal functioning of the brain. One cannot expect an NCC to correlate with consciousness in cases with abnormal functioning or different architecture, so no direct conclusion can be drawn from failure of correlation across lesion cases. Of course, on can still appeal to cases with unusual inputs and brain stimulation to make fine-grained distinctions among NCCs.

The main conceptual objection to the second alternative is that one might *need* lesion cases to make the most fine-grained distinctions that are required. Consider a hypothetical case in which we have two linked systems N and M which correlate equally well with consciousness across all normal cases, including all unusual inputs and brain stimulation, but such that in almost all relevant lesion cases, consciousness correlates much better with N than with M. In this case, might we want to say that N rather than M is an NCC? If so, we have to build in some allowance for abnormal cases into the definition of an NCC. An advocate of the second alternative might reply by saying that such cases will be very unusual, and that if N and M are dissociable by lesions, there is likely to be some unusual brain stimulation that will bring out the dissociation as well. In the extreme case where no brain stimulation leads to dissociation, one might simply bite the bullet and say that both N and M are equally good NCCs.

Taking everything into consideration, I am inclined to think the second alternative is better than the first. It seems right to say that "core" NCC location depends on brain architecture and normal functioning, and it is unclear that correlation across abnormal cases should be required, especially given all the associated problems. A problem like the one just mentioned might provide some pressure to investigate the first alternative further, and I do not rule out the possibility that some way of distinguishing "good" from "bad" lesions might be found, but all in all it seems best to say that an NCC cannot be expected to correlate with consciousness across abnormal cases.

Of course this has an impact on the methodology in the search for an NCC. As we have seen, lesion studies are often used to draw conclusions about NCC location (as in the Milner and Goodale research, for example, and also in much research on blindsight), and failure of correlation in lesion cases is often taken to imply that a given system is not an NCC. But we have seen that the tacit premise of this sort of research - that an NCC must correlate across abnormal as well as normal cases - is difficult to support, and leads to significant problems. So it seems that lesion studies are methodologically dangerous here. One should be very cautious in using them to draw conclusions about NCC location.

This is not to say that lesion studies are irrelevant in the search for an NCC. Even if correlation across abnormal cases is not *required* for system N to be an NCC, it may be that correlation across abnormal cases can provide good *evidence* that N is an NCC, and that failure of such correlation in some cases provides good evidence that N is not an NCC. Say we take the second alternative above, and define an NCC as a system that correlates with consciousness across all normal cases (including unusual input and stimulation). It may nevertheless be the case that information about correlations across all these normal cases with unusual stimulation is difficult to come by (due to problems in monitoring brain systems at a fine grain, for example), and that information about correlation across lesion cases is easier to obtain. In

this case, one might sometimes take correlation across abnormal cases as *evidence* that a system will correlate across the normal cases in question, and thus as evidence that the system is an NCC. Similarly, one might take failure of correlation across abnormal cases as evidence that a system will fail to correlate across certain normal cases, and thus as evidence that the system is not an NCC.

The question of whether a given lesion study can serve as evidence in this way needs to be taken on a case-by-case basis. It is clear that some lesion studies will not provide this sort of evidence, as witnessed by the cases of severe lesions and downstream lesions discussed earlier. In the cases, failure of correlation across abnormal cases provides no evidence of failure of correlation across normal cases. On the other hand, it does not seem unreasonable that the Milner and Goodale studies should be taken as evidence that even in normal cases, the ventral stream will correlate better with visual consciousness than the dorsal stream. Of course the real "proof" would come from a careful investigation of the relevant processes across a wide range of "normal" cases involving both standard environments, unusual inputs, and brain stimulation; but in the absence of such a demonstration, the lesion cases at least provide suggestive evidence.

In any case, the moral is that one has to be very cautious when drawing conclusions about NCC location from lesion studies. At best these studies serve as indirect evidence rather than as direct criteria, and even as such there is a chance that the evidence can be misleading. One needs to consider the possibility that the lesion in question is changing brain architecture in such a fashion that what was once an NCC is no longer an NCC, and one needs to look very closely at what is going on to rule out the possibility. It may be that this can sometimes be done, but it is a nontrivial matter.

4 Overall definition

With all this, we have come to a more detailed definition of an NCC. The general case is something like the following:

> 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.

The central case of the neural correlate of the content of consciousness can be put in more specific terms.

> An NCC (for content) is a minimal neural representational system N such that representation of a content in N is sufficient, under conditions C, for representation of that content in consciousness.

One might also give a general definition of the NCC for an arbitrary phenomenal property or for a phenomenal family, but I will leave those aside here.

The "conditions C" clause here represents the relevant range of cases, as discussed above. If the reasoning above is on the right track, then 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. Perhaps one could make a case for including a limited range of lesion cases in the definition. In the other direction, perhaps one might make a case that the requirement of correlation across brain stimulation or unusual inputs is too strong, due to the abnormality of those scenarios. But I think the conditions C proposed here are at least a reasonable first pass, pending further investigation.

Of course, to some extent, defining what "really" counts as an NCC is a terminological matter. One could quite reasonably say that there are multiple different notions of NCC, depending on just how one understands the relevant conditions C, or the matter of necessity and sufficiency, and so on; and not much really rests on which of these is the "right" definition. Still, we have seen that different definitions give very different results, and that many potential definitions have the consequence that systems that intuitively seem to qualify as an NCC do not qualify after all, and that NCC hypotheses put forward by researchers in the field could be ruled out on trivial a priori grounds. Those consequences seem undesirable. It makes sense to have a definition of NCC that fits the way the notion is generally used in the field, and that can make sense of empirical research in the area. At the same time we want a definition of

NCC to be coherent and well-motivated in its own right, such that an NCC is something worth looking for, and such that the definition can itself be used to assess various hypotheses about the identity of an NCC. It seems to me that the definition I have given here is at least a first pass in this direction.

5 Methodological consequences

The discussion so far has been somewhat abstract, and the definitions given above may look like mere words, but from these definitions and the reasoning that went into them, one can straightforward extract some concrete methodological recommendations for the NCC search. Many of these recommendations are plausible or obvious in their own right, but it is interesting to see them emerge from the analysis.

(i) *Lesion studies are methodologically dangerous*. Lesion studies are often used to draw conclusions about neural correlates of consciousness, but we have seen that their use can be problematic. The identity of an NCC is arguably always relative to specific brain architecture and normal brain functioning, and correlation across abnormal cases should not generally be expected. In some cases, lesion studies can change brain architecture so that a system that was previously an NCC is no longer an NCC. So one can never infer directly from failure of correlation between a system and consciousness in a lesion case to the conclusion that that system is an NCC. Sometimes one can infer this indirectly, by using the failure of correlation here as evidence for failure of correlation in normal cases, but one must be cautious.

(ii) *There may be many NCCs*. On the definition above, an NCC is a system whose activity is *sufficient* for certain states of consciousness. This allows for the possibility of multiple NCCs, in at least two ways. First, different sorts of conscious states may have different corresponding NCCs; there may be different NCC for visual and auditory consciousness, for example, and perhaps even for different aspects of visual consciousness. Second, even for a particular sort of conscious state (such as pain), we cannot rule out the possibility that there will be two different systems whose activity is sufficient to produce that state.

Of course it *could* turn out that there is only a small number of NCCs, or perhaps even one. For all that I have said here, it is possible that there is some central system which represents the contents of visual consciousness, auditory consciousness, emotional experience, the stream of conscious thought, the background state of consciousness, and so on. Such a system might be seen as a sort of "consciousness module", or perhaps as a "Cartesian theater" (Dennett 1991) or a "global workspace" (Baars 1988), depending on whether one is a foe or a friend of the idea (see Chalmers 1998 for some discussion). But it is by no means obvious that there will be such a system, and I think the empirical evidence so far is against it. In any case, the matter cannot be decided a priori, so our definition should be compatible with the existence of multiple or many NCCs.

(3) *Minimize size of an NCC*. We have seen that an NCC should be understood as a *minimal* neural system that correlates with consciousness. Given this, we should constrain the search for the NCC by aiming to find a neural correlate that is as small as possible. Given a broad system that appears to correlate with consciousness, we need to isolate the core relevant parts and aspects of that system that underlie the correlation. And given the dual hypotheses that consciousness correlates with a broad system or a narrower system contained within it, we might first investigate the "narrow" hypothesis, as if it correlates with consciousness, the broad system cannot be a true NCC.

So it some extent it makes sense to "start small" in the search for an NCC. This fits the working methodology proposed by Crick and Koch (1998). Crick and Koch suggest that an NCC may perhaps involve a very small number of neurons (perhaps in the thousands) with certain distinctive properties. There is no guarantee that this is correct (and my own money is against it), but it makes a good working hypothesis in the NCC search. Of course one should simultaneously investigate broad systems for correlation with consciousness, so that one can then focus on those areas and try to narrow things down.

(4) *Distinguish NCC for background state and for content*. We have seen that there may be different NCCs for different sorts of states of consciousness. An important distinction in this class is that between the neural correlate of background state of consciousness (wakefulness, dreaming, etc) and the neural correlate of specific contents. It may be that these are quite different systems. It is not implausible on current evidence that an NCC for background state involves processes in the thalamus, or thalamocortical interactions, while an NCC for specific contents of consciousness involves processes in the cortex. These different sorts of NCC will require quite different methods for their investigation.

(5) *NCC studies need to monitor neural representational content*. Arguably the most interesting part of the NCC search is the search for neural determinants of specific contents of consciousness, such as the contents of visual consciousness. We have seen that an NCC here will be a neural representational system whose contents are correlated with the contents of consciousness. To determine whether such a system is truly an NCC, then, we need methods that monitor the representational content of the system. This is just what we find in Logothetis's work, for example, where it is crucial to keep track of activity in neurons with known receptive fields.

This gets at a striking aspect of the NCC search in practice, which is that the most informative and useful results usually come from single-cell studies on monkeys. Large claims are sometimes made for brain imaging on humans, but it is generally difficult to draw solid conclusions from such studies, especially where an NCC is concerned. We can trace the difference to the fact that single-cell studies can monitor representational content in neural systems, whereas imaging studies cannot (or at least usually do not). The power of single-cell studies in the work of Logothetis, Andersen, Newsome and colleagues (e.g. the works of Logothetis and Newsome already cited, and Bradley, Chang, and Andersen 1998) comes precisely from the way that cells can be monitored to keep track of their activity profile of neurons with known representational properties, such as receptive and projective fields. This allows us to track representational content in these neural systems and to correlate it with the apparent contents of consciousness. This is much harder to do in a coarse-grained brain imaging study, which generally tells one that there is an activity in a region while saying nothing about specific contents.

A moral is that it makes sense to concentrate on developing methods that can track neural representational content, especially in humans (where invasive studies are much more problematic, but where evidence for conscious content are much more straightforward). There has been some recent work on the using imaging methods to get at certain aspects of the content of visual consciousness, such as colors and shapes in the visual field (e.g. Engel et al 1997), and different sorts of objects that activate different brain areas (e.g. Tong et al 1998). There is also some current work on using invasive methods in neurosurgery patients to monitor the activity of single cells. One can speculate that if a noninvasive method for monitoring single-cell activity in humans is ever developed, the search for an NCC (like most of neuroscience) will be transformed almost beyond recognition.

(6) *Correlation across a few situations is limited evidence*. According to the definition above, an NCC is a system that correlates with consciousness across arbitrary cases of normal functioning, in any environment, with any unusual input or limited brain stimulation. In practice, though, evidence is far weaker than this. Typically one has a few cases, involving either a few subjects with different lesions, or study in which subjects are given different stimuli, and one notes an apparent correlation. This is only to be expected, given the current technological and ethical constraints on experimental methods. But it does mean that the evidence that current methods give is quite weak. To truly demonstrate that a given system is an NCC, one would need to demonstrate correlation across a far wider range of cases than is currently feasible. Of course current methods may give good *negative* evidence about systems that fail to correlate and thus are not NCCs, but strong positive evidence is harder to find. Positive hypotheses based on current sorts of evidence should probably be considered suggestive but highly speculative.

(7) *We need good criteria for the ascription of consciousness*. To find an NCC, we need to find a neural system that correlates with certain conscious states. To do this, we first need a way to know when a system is in a given conscious state. This is famously problematic, given the privacy of consciousness and the philosophical problem of other minds. In general, we rely on indirect criteria for the ascription of consciousness. The most straightforward of these criteria is verbal report in humans, but other criteria are often required. Where nonhuman subjects are involved, one must rely on quite indirect behavioral signs (voluntary bar-pressing in Logothetis's monkeys, for example).

A deep problem for the field is that our ultimate criteria here are not experimentally testable, as the results of any experiment will itself requires such criteria for its interpretation. (First-person experimentation on oneself as subject may be an exception, but even this has limitations.) So any experimental work implicitly relies on pre-empirical principles (even "philosophical" principles) for its interpretation. Given this, it is vital to refine and justify these pre-empirical principles as well as we can. In the case of verbal report, we may be on relatively safe ground (though even here there may be some grounds for doubt, as witnessed in the debates over "subjective threshold" criteria in unconscious perception research; see e.g. Merikle and Reingold (1992). In other cases, especially nonhuman cases, careful attention to the assumptions involved here are required. I don't think this problem is insurmountable, but it deserves careful attention. Our

conclusions about NCC location will be no better than the pre-experimental assumptions that go into the search. (I consider this problem, and its consequences for the NCC search, in much more detail in Chalmers 1998.)

Methodological summary: We can use all this to sketch a general methodology for the NCC search. First, we need methods for determining the contents of conscious experience in a subject, presumably by indirect behavioral criteria or by first-person phenomenology. Second, we need methods to monitor neural states in a subject, and in particular to monitor neural representational contents. Then we need to perform experiments in a variety of situations to determine which neural systems correlate with conscious states and which do not. Experiments involving normal brain functioning with unusual inputs and limited brain stimulation are particularly crucial here. Direct conclusions cannot be drawn from systems with lesions, but such systems can sometimes serve as indirect evidence. We need to consider multiple hypotheses in order to narrow down a set of minimal neural systems that correlate with consciousness across all relevant scenarios. We may well find many different NCCs in different modalities, and different NCCs for background state and conscious contents, although it is not out of the question that there will be only a small number. If all goes well, we might expect to eventually isolate systems that correlate strongly with consciousness across any normally functioning brain.

6 Should we expect an NCC?

One might well ask: given the notion of an NCC as I have defined it, is it guaranteed that there will *be* a neural correlate of consciousness?

In answering, I will assume that states of consciousness depend systematically in some way on overall states of the brain. If this assumption is false, as is held by some Cartesian dualists (e.g. Eccles 1994) and some phenomenal externalists (e.g. Dretske 1995), then there may be no NCC as defined here, as any given neural state might be instantiated without consciousness. (Even on these positions, an NCC *could* be possible, if it were held that brain states at least correlate with conscious states in ordinary cases). But if the assumption is true, then there will at least be some minimal correlation of neural states with consciousness.

Does it follow that there will be an NCC as defined here? This depends on whether we are talking about neural correlates of arbitrary conscious states, or about the more constrained case of neural correlates of conscious contents. In the first case, it is guaranteed that the brain as a whole will be a neural system which has states that suffice for arbitrary conscious states. So the brain will be one system whose state is sufficient for a given conscious state; and given that there is at least one such system for a given state, there must be at least one such *minimal* system for that state. Such a system will be an NCC for that state. Of course this reasoning does not guarantee that there will be only one NCC for a given state, or that the NCC for one state will be the same as the NCC for another, but we know that an NCC will exist.

In the case of neural correlates of the content of consciousness, things are more constrained, since a neural correlate is required not just to map to a corresponding state of consciousness, but to match it in *content*. This rules out the whole brain as even a non-minimal neural correlate, for example, since representing a content in the brain does not suffice to represent that content in consciousness (much of the brain's representational content is unconscious). Of course we may hope that there will be more constrained neural systems whose content systematically matches the contents of some aspect of consciousness. But one might argue that it is not obvious that such a system *must* exist. It might be held, for example, that the contents of consciousness are an emergent product of the contents of various neural systems, which together suffice for conscious content in question, but none of which precisely mirrors the conscious content.

I think one can plausibly argue that there is reason to expect that conscious contents will be mirrored by the contents of a neural representational system at *some* level of abstraction. In creatures with language, for example, conscious contents correspond well with contents that are made directly available for verbal report; and in conscious creatures more generally, one can argue that the contents of consciousness correspond to contents that are made directly available for the global voluntary control of behavior (see e.g. Chalmers 1998). So there is a correlation between the contents of consciousness and contents revealed or exhibited in certain functional roles within the system.

Given that these contents are revealed in verbal report and are exhibited in the control of behavior, there is reason to believe that they are represented at some point within the cognitive system. Of course this depends to some extent on just what "representation" comes to. On some highly constrained notions of representation - if it is held that the only true representation is symbolic representation, for example - then it is far from clear that the content revealed in behavior must be represented. But on less demanding notions of representation - on which, for example, systems are assigned representational content according to their functional role - then it will be natural to expect that the content revealed in a functional role will be represented in a system that plays that functional role.

This does not guarantee that there will be any single neural system whose content always matches the content of consciousness. It may be that the functional role in question is played by multiple systems, and that a given system may sometimes play the role, and sometimes not. If this is so, we may have to move to a higher level of abstraction. If there is no localizable neural system that qualifies as a correlate of conscious content, we may have to look at a more global system - the "global availability" system, for example, whereby contents are made available for report and global control - and argue that the contents of consciousness correspond to the contents made available in this system. If so, it could turn out that what we are left with is more like a "cognitive correlate of consciousness" (CCC?), since the system may not correspond to any neurobiological system whose nature and boundaries are independently carved out. But it can still function as a correlate in some useful sense.

In this context, it is important to note that an NCC need not be a specific anatomical area in the brain. Some of the existing proposals regarding NCCs involve less localized neurobiological properties. For example, Libet (1994) argues that the neural correlate of consciousness is temporally extended neural firing; Crick and Koch (1998) speculate that the NCC might involve a particular sort of cell, throughout the cortex; Edelman (1988) suggests that the NCC might involve re-entrant thalamocortical loops; and so on. In these cases, NCCs are individuated by temporal properties, or by physiological rather than anatomical properties, or by functional properties, among other possibilities. If so, the "neural representational system" involved in defining a neural correlate of conscious content might also be individuated more abstractly: the relevant neural representational contents might be those represented by temporally extended firings, or by certain sorts of cells, or by re-entrant loops, and so on. So abstractness and failure of localization is not in itself a bar to a system’s qualifying as an NCC.

It seems, then, that there are a range of possibilities for the brain-based correlates of conscious states, ranging from specific anatomical areas, through more abstract neural systems, to purely "cognitive" correlates such as Baars’ (1988) "global workspace". Just how specific an NCC may turn out to be is an empirical question. One might reasonably expect that there will be some biological specificity. Within a given organism or species, one often funds a close match between specific functions and specific physiological systems, and it does not seem unlikely that particular neural systems and properties in the brain should be directly implicated in the mechanisms of availability for global control. If so, then we may expect specific neural correlates even of conscious contents. If not, we may have to settle for more abstract correlates, individuated at least partly at the cognitive level, though even here one will expect that some neural systems will be much more heavily involved than others. In any case it seems reasonable to expect that we will find informative brain-based correlates of consciousness at some level of abstraction in cognitive neurobiology.

Some have argued that we should not expect neural correlates of consciousness. For example, in their discussion of neural "filling-in" in visual perception, Pessoa, Thompson, and Noe (1998) argue against the necessity of what Teller and Pugh (1983) call a "bridge locus" for perception, which closely resembles the notion of a neural correlate of consciousness. Much of their argument is based on the requirement that such a locus must involve a spatiotemporal isomorphism between neural states and conscious states (so a conscious representation of a checkerboard would require a neural state in a checkerboard layout, for example). These arguments do not affect neural correlates of conscious contents as I have defined them, since a match between neural and conscious content does not require such a spatiotemporal correspondence (a neural representation of a shape need not itself have that shape). Pessoa et al also argue more generally against a "uniformity of content" thesis, holding that one should not expect a match between the "personal" contents of consciousness and the "subpersonal" contents of neural systems. I agree that the existence of such a match is not automatic, but as above I think that the fact that conscious contents are mirrored in specific functional roles gives reason to believe that they will be subpersonally represented at least at some level of abstraction.

It has also been argued (e.g. by Güzeldere 1999) that there is probably no neural correlate of consciousness, since there is probably no area of the brain that is specifically dedicated to consciousness as opposed to vision, memory, learning, and so on. One may well agree that there is no such area, but it does not follow that there is no neural correlate of consciousness as defined here. An NCC (as defined here) requires only that a system be correlated with consciousness, not that it be dedicated solely or mainly to consciousness. The alternative notion of an NCC is much more demanding that the notion at issue in most empirical work on the subject, where it is often accepted that an NCC may be closely bound up with visual processing (e.g. Logothetis, Milner and Goodale), memory (e.g. Edelman), and other processes. This becomes particularly clear once one gives up on the requirement that there be a single NCC, and accepts that there may be multiple NCCs in multiple modalities.

7 Conclusion

The discussion in the previous section helps bring out what an NCC is *not*, or at least what it might turn out not to be. An NCC is defined to be a *correlate* of consciousness. From this, it does not automatically follow that an NCC will be a system solely or mainly dedicated to consciousness, or even that an NCC will be the brain system most responsible for the generation of consciousness. It certainly does not follow that an NCC will itself yield an explanation of consciousness, and it is not even guaranteed that identifying an NCC will be the key to understanding the processes underlying consciousness. If were to define an NCC in these stronger terms, it would be far from obvious that there must be an NCC, and it would also be much less clear how to search for an NCC.

Defining an NCC in terms of correlation seems to capture standard usage best, but it also makes the search more clearly defined, and makes the methodology clearer. Correlations are easy for science to study. It also means that the search for an NCC can be to a large extent theoretically neutral, rather than theoretically loaded. Once we have found an NCC, one might hope that it will turn out to be a system dedicated to consciousness, or that it will turn out to yield an explanation of consciousness, but these are further questions. In the meantime the search for an NCC as defined poses a tractable empirical question with relatively clear parameters, one which researchers of widely different theoretical persuasions can engage in.

There are certain rewards of the search for an NCC that one might reasonably expect. For example, these systems might be used to monitor and predict the contents of consciousness in a range of novel situations. For example, we may be able to use them to help reach conclusions about conscious experience in patients under anesthesia, and in subjects with "locked-in syndrome" or in coma. In cases where brain architecture differs significantly from the original cases (perhaps some coma cases, infants, and animals), the evidence will be quite imperfect, but it will at least be suggestive.

These systems might also serve as a crucial step toward a full science of consciousness. Once we know which systems are NCCs, we can investigate the mechanisms by which they work, and how they produce various characteristic functional effects. Just as isolating the DNA basis of the gene helped explain many of the functional phenomena of life, isolating NCC systems may help explain many functional phenomena associated with consciousness. We might also systematize the relationship between NCCs and conscious states, and abstract general principles governing the relationship between them. In this way we might be led to a much greater theoretical understanding.

In the meantime, the search for a neural correlate of consciousness provides a project that is relatively tractable, clearly defined, and theoretically neutral, whose goal seems to be visible somewhere in the middle distance. Because of this, the search makes an appropriate centerpiece for a developing science of consciousness, and an important springboard in the quest for a general theory of the relationship between physical processes and conscious experience.

References

Anderson, R.A. 1997. Neural mechanisms in visual motion perception in primates. Neuron 18:865-872.

Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bogen, J.E. 1995. On the neurophysiology of consciousness, part I: An overview. Consciousness and Cognition 4:52-62.

Bradley, D.C., Chang, G.C. & Andersen, R.A. 1998. Encoding of three-dimensional structure-from-motion by primate area MT neurons. Nature 392:714-17.

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D.J. 1995. Facing up to the problem of consciousness. Journal of Consciousness Studies 2:200-19. Also in (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness* (MIT Press), and in (J. Shear, ed) *Explaining Consciousness: The Hard Problem* (MIT Press).

Chalmers, D.J. 1998. On the search for the neural correlate of consciousness. In (S. Hameroff, A. Kaszniak, & A. Scott, eds) *Toward a Science of Consciousness II*. MIT Press.

Crick, F. and Koch, C. 1990. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2: 263-275.

Crick, F. & Koch, C. 1995. Are we aware of neural activity in primary visual cortex? Nature 375: 121-23.

Crick, F. & Koch, C. 1998. Consciousness and neuroscience. Cerebral Cortex.

Dennett, D.C. 1991. *Consciousness Explain*. Little-Brown.

Dretske, F. 1995. *Naturalizing the Mind*. MIT Press.

Eccles, J.C. 1994. *How the Self Controls its Brain*. New York: Springer-Verlag.

Edelman, G.M. 1989. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Engel, S., Zhang, X. & Wandell, B. 1997. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. Nature 388 (6637):68-71.

Flohr, H. 1995. Sensations and brain processes. Behavioral Brain Research 71:157-61.

Gur, M. & Snodderly, D.M. 1997. A dissociation between brain activity and perception: Chromatically active cortical neurons signal chromatic activity that is not perceived.

Güzeldere, G. 1999. There is no neural correlate of consciousness. Paper presented at "Toward a Science of Consciousness: Fundamental Approaches", Tokyo, May 25-28, 1999.

He, S., Cavanagh, P. & Intriligator, J. 1996. Attentional resolution and the locus of visual awareness. Nature 384:334-37.

Hobson, J.A. 1997. Consciousness as a state-dependent phenomenon. In (J. Cohen & J. Schooler, eds) *Scientific Approaches to Consciousness*. Lawrence Erlbaum.

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

Libet, B. 1982. Brain stimulation in the study of neuronal functions for conscious sensory experiences. Human Neurobiology 1:235-42.

Llinas, R.R., Ribary, U., Joliot, M. & Wang, X.-J. 1994. Content and context in temporal thalamocortical binding. In (G. Buzsaki, R.R. Llinas, & W. Singer, eds.) *Temporal Coding in the Brain*. Berlin: Springer Verlag.

Logothetis, N. & Schall, J. 1989. Neuronal correlates of subjective visual perception. Science 245:761-63.

Marge, E. 1991. Magnetostimulation of vision: Direct noninvasive stimulation of the retina and the visual brain. Optometry and Vision Science 68:427-40.

Merikle, P.M. & Reingold, E.M. 1992. Measuring unconscious processes. In (R. Bornstein & T. Pittman, eds) *Perception without Awareness*. Guilford.

Milner, A.D. 1995. Cerebral correlates of visual awareness. Neuropsychologia 33:1117-30.

Milner, A.D. & Goodale, M.A. 1995. *The Visual Brain in Action*. Oxford University Press.

Newman, J.B. 1997. Putting the puzzle together: Toward a general theory of the neural correlates of consciousness. Journal of Consciousness Studies 4:47-66, 4:100-121.

Penfield, W. 1937. The cerebral cortex and consciousness. In *The Harvey Lectures*. Reprinted in R.H. Wilkins (ed.) *Neurosurgical Classics* (New York: Johnson Reprint Corp., 1965).

Penfield, W. & Rasmussen, T. 1950. *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*.

Pessoa, L., Thompson, E. & Noe, A. 1998. Finding out about filling in: A guide to perceptual completion for visual science and the philosophy of perception. Behavioral and Brain Sciences 21:723-748.

Salzman, C.D., Britten, K.H., & Newsome, W.T. 1990. Cortical microstimulation influences perceptual judgments of motion direction. Nature 346:174-77.

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

Shoemaker, S. 1981. Some varieties of functionalism. Philosophical Topics 12:93-119. Reprinted in *Identity, Cause, and Mind* (Cambridge University Press, 1984).

Teller, D.Y. & Pugh, E.N. 1984. Linking propositions in color vision. In (J.D. Mollon & L.T. Sharpe, eds) *Color Vision: Physiology and Psychophysics*. London: Academic Press.

Tong, F., Nakayama, K., Vaughan, J.T., & Kanwisher N. 1998. Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21:753-759.