

Correlation, Causation, Constitution:

On the Interplay between the Science and Philosophy of Consciousness

Benjamin Kozuch & Uriah Kriegel

Forthcoming in S.M. Miller (ed.), *The Constitution of Consciousness*, John Benjamins

Abstract :: Consciousness is a natural phenomenon, the object of a flourishing area of research in the natural sciences – research whose primary goal is to identify the neural correlates of consciousness. This raises the question: why is there need for a *philosophy* of consciousness? After all, most natural phenomena do not invite, in addition to empirical scientific study, a dedicated philosophical investigation as well. As we see things, the need for a philosophy of consciousness arises for two reasons. First, as a young and energetic science operating as yet under no guiding paradigm, the science of consciousness has been subject to considerable confusion regarding its methodological, conceptual, and philosophical foundations, and it is philosophy's mandate to address such confusion in an attempt to regiment scientific practice. Secondly, the identification of a neural feature that correlated perfectly with consciousness would still leave open a certain metaphysical question: is the relation between consciousness and the relevant neural feature *merely* correlation, or is that correlation indicative of a deeper, more intimate relation between the two? In this paper, we offer an opinionated overview of the philosophy of consciousness as it addresses these two dimensions of consciousness research.

1. Introduction

Consciousness is a natural phenomenon, the object of a flourishing area of research in the natural sciences – research whose primary goal is to identify the “neural correlates” of consciousness. This raises the question: why is there need for a *philosophy* of consciousness? After all, most natural phenomena do not invite, in addition to empirical scientific study, a dedicated philosophical investigation as well.

As we see things, the need for a philosophy of consciousness to supplement the science of consciousness arises for two reasons. First, as a young and energetic science operating as yet under no guiding paradigm, the science of consciousness has been subject to a somewhat uncontrolled proliferation of approaches and presuppositions regarding the methodological, conceptual, and philosophical foundations of the search for the neural correlates of consciousness. It is philosophy’s mandate to address such methodological confusion in an attempt to regiment scientific practice, and many philosophers have indeed weighed in on these foundational issues for consciousness science.

Secondly, the identification of a neural feature that correlated perfectly with consciousness would still leave open a certain metaphysical question: is the relation between consciousness and the relevant neural feature *merely* correlation, or is that correlation indicative of a deeper, more intimate relation between the two? Work addressing this further question can be thought of as attempting a philosophical interpretation of scientific theories, somewhat on a par, say, with philosophical interpretations of quantum mechanics: in both cases, philosophy has to take over where science proper ends in order to articulate an intelligible conception of how the world must be given what the science suggests.

Thus a philosophy of consciousness appears to be necessary both to firm up the foundations of the science of consciousness and to take over where the science proper leaves off. In what follows, we offer an opinionated overview of the

philosophy of consciousness as it addresses these two dimensions of consciousness research: §2 discusses conceptual issues regarding the notion of neural correlates of consciousness and methodological approaches to the search for these neural correlates; §3 discusses the variety of metaphysical relationships neural correlation might suggest.

2. From Philosophy to Science: Neural Correlates of Consciousness and the Content-Matching Method

It has sometimes been claimed that consciousness is now “largely a scientific problem,” one that might proceed in complete isolation from philosophy (Crick 1996). This looks implausible on the face of it, given the vexing metaphysical issues attendant to the study of consciousness (see §3). But philosophy appears to have a role in the scientific study of consciousness itself, by way of clarifying the conceptual and methodological foundations for a scientific study of consciousness. Our aim in this section is to survey and (briefly) extend some of this work.

Scientific research into consciousness has had two research foci: the neural correlates of consciousness (NCC), and the function of consciousness. In this paper, we discuss the NCC exclusively. More specifically yet, we are concerned with what is sometimes referred to as the *content* NCC, as opposed to the *background* NCC (Frith et al 1999; Chalmers 2000; Rees *et al* 2002; Crick & Koch 2003).¹ Although there has been important research on the background NCC, the better-known work concerns the content NCC, and it is the content NCC with which we are here concerned. We begin with a look at the standard explication and regimentation of the concept of an

¹ The content of a conscious experience is given by whatever one is conscious of. (In an experience as of a blue surface, for example, the blueness the person experiences is part of the content of her experience.) Whatever neural systems correlate with the content of experience are the content NCC. The background NCC are whatever neural systems correlate with general modes of consciousness, such as dreaming, being wide awake, or being disoriented. Some potential neural correlates of background consciousness have been proposed, but there is as yet no consensus on what they are.

NCC, as this suggests an initially plausible methodology for (content) NCC research. (Henceforth, by “NCC” we will mean specifically *content* NCC.)

The regimentation we have in mind is due to Chalmers (2000: 31), who defines an NCC as follows: “An NCC (for content) is a minimal neural representational system N such that representation of a content in N is sufficient...for representation of that content in consciousness.” Three central features of this definition should be noted.

A first feature of this definition is that it requires neural representation of a content to be only *sufficient*, not *necessary*, for conscious representation of that content (contra Teller & Pugh 1983 and Kanwisher 2001). This leaves open the possibility of there being more than one neural system able to represent some content in consciousness.

Secondly, and relatedly, the definition also states that the NCC will be the *minimal* neural representational system sufficient for representation of the content in consciousness. This delimits the NCC so that they contain “only the core processes that suffice for the conscious state in question” (Chalmers 2000: 9). Suppose N is a neural system that represents content C and is sufficient for some conscious experience E to represent C. Typically there will be some other, vaster neural system N*, such that N is a proper part of N*. It will then be trivial both that N* contains a representation of C and that N*'s representation of C is sufficient for E's representing C.² Yet it seems that the parts of N* that do not have to do with N are in some sense irrelevant to the NCC. It is to rule out of the NCC these parts that the definition appeals to a *minimal* sufficient condition. This feature of the definition shows sensitivity to the idea that the search for the NCC is a search for the *basis* of consciousness. (Thus a distinction is sometimes made between the notions of a neural “correlate” and a neural “basis,” with the thought that the neuroscience of

² Indeed, there should also be a yet vaster not-only-neural biological system, encompassing N* as a part, that would trivially contain a representation of C. (So-called extended consciousness theorists might hold that the biological correlates of consciousness are of this sort.)

consciousness should seek the latter.³ The sensibility behind this thought is partially spoken to by the minimality condition, whether it is considered a condition on a neural “correlate” (as in Chalmers’ definition) or only on a neural “basis.” Below, we use the two terms interchangeably,).

Finally, Chalmers’ definition requires an NCC for some experience E to match E in content. This requirement seems driven by a certain metaphysical assumption regarding the neural basis of the content of an experience: that as a matter of nomological necessity, any neural system forming the basis of E must have the same representational content as E. Noë and Thompson (2004) have dubbed this the “isomorphism constraint.” To see the rationale for the isomorphism constraint, consider how strange instances of its violation would be. Suppose some neural system N was the neural basis of an experience as of a vertical line, but N represented a horizontal rather than vertical line? To think that an experience and its neural basis could fail to match in content in this way would be “to suppose that there was no intelligible connection (beyond brute correlation) between the experience and the neural locus in question” (Noë & Thompson 2004: 5).⁴

The isomorphism constraint suggests conditions under which we would be justified in thinking that we had found an NCC. Noë and Thompson (2004: 7) write:

Suppose one discovered a neural representational system N such that (i) N represents that p, and (ii) N’s activity is correlated with the occurrence of a perceptual experience with the content that p. If one discovered such a neural representational system, it might seem

³ Among authors who make this distinction, or ones very like it, are Kanwisher (2001), Crick and Koch (2003), Block (2007), and Miller (2007). While these commentators do not all use specifically the term “neural basis,” it is clear that they have something along these lines in mind. Crick and Koch, for example, speak of looking for “the minimal set of neuronal events that gives rise to a specific aspect of a conscious experience” (2003:119). Kanwisher argues that we should seek those “patterns of neural activity [that] are necessary and/or sufficient for perceptual awareness” (2001:98). Block claims that “at a minimum, one wants the neural underpinnings of a match of content between the mental and neural state[s]” (2007:481). And Miller writes, “not every neural correlate of a conscious state is necessarily constitutive of that state” (2007: 161).

⁴ Beyond the matter of intelligibility, or perhaps partly in light of it, the isomorphism constraint seems to be a ground-floor metaphysical assumption for any materialist theory of consciousness, as well as the kind of naturalistic dualism propounded by Chalmers (1996).

reasonable to think...one had discovered the place in the brain where the conscious experience happens.

This suggests a certain methodology for identifying the NCC, to which Noë and Thompson (2004: 4) refer as the *matching-content doctrine*: “The first task of the neuroscience of consciousness is to uncover the neural representation systems whose contents systematically match the contents of consciousness.” The matching-content doctrine extends the insight provided by the isomorphism constraint into a paradigm for NCC research: To find the NCC, we should look for matches in content between neural systems and experiences. We will refer to this paradigm as the *content-matching method*.

In the remainder of this section, we take a critical look at the content-matching method. We argue that content matches provide but weak justification for taking some neural system to be an NCC, since a neural system can match an experience in content and still fail to be the neural basis of that experience; namely, in case there is some other neural system whose content matches the experience and which is actually operative in yielding the experience.⁵ Instead of the content-matching method, we advocate a *content-mismatch method*, which decisively shows a neural system to be *not* the neural basis of an experience. Our conclusion will be that a content-matching method—at least as conceived of above—is a mistaken picture of what successful NCC research will look like.

Some important results in NCC research appear based on a content-matching method. Consider the following study by Tong *et al.* (1998). If a subject is given a different stimulus to each eye, this can result in so-called binocular rivalry: Instead of experiencing a single image that fuses the stimuli, the subject’s experience oscillates between the two. In the Tong *et al.* study, subjects were fed an image of a face to one eye and a house to the other, inducing binocular rivalry. The experimenters employed fMRI, with special attention to two temporal areas: the

⁵ Here we intend the term “yield” to be neutral as between causal and constitutive readings; more on this below.

fusiform face area (FFA), thought to specialize in the processing of faces; and the parahippocampal place area (PPA), thought to specialize in the processing of locales (such as houses). Tong *et al.* found increased activity in the FFA when the subject experienced the face and in the PPA when the subject experienced the house. Such results seem to indicate a content match, in that activity in the FFA (which represents faces) increased when and only when the subject saw the face, while activity in the PPA (which represents house, *inter alia*) increased when and only when the subject saw the house. Likely, it is this apparent content match that brought Tong *et al.* (1998: 75) to claim that this experiment “support[s] the notion that multiple extrastriate regions [i.e., the FFA and PPA]... participate in our awareness of...the visual world.”

Let us assume that this experiment presents an instance of a content match between the FFA and the face experience. We maintain that this match in content, on its own, provides little justification for thinking that the FFA is the neural basis of face experience. To be well-justified in thinking this, one would need reason for thinking that there is no other neural system(s) also matching the face experience in content. For if some other neural system also matches the face experience in content, then it is possible that this other neural system (and *not* the FFA) is actually the neural basis of the face experience. And so it seems that, even if Tong *et al.*'s experiment demonstrates a content match, it gives but limited support (on its own) to the hypothesis that the FFA is an NCC.

To clarify, we do not wish to deny that the FFA is in fact the neural basis of face experience (we take no stand on that here). For there may be good reasons for thinking that no other neural system besides the FFA matches the experience in content. But it is the absence of other content matches, rather than the presence of this content match, that we believe is central for supporting the hypothesis that the FFA is in fact the neural basis of face experience. Thus we think that the justification for believing the FFA to be an NCC should come almost entirely from the fact that other neural systems have been ruled out as potential content-matchers. The positive content match provided by the Tong *et al.* study would be but subsidiary evidence for this hypothesis, since (as pointed out above) a content match between

the FFA and face experience is consistent with the FFA not being the neural basis of that experience.

This observation will apply, of course, to all content matches: Any content match between some neural system N and experience E does little to support the idea that N is the neural basis of E , since the content match (on its own) does nothing to rule out there being some other neural system N^* that also matches E in content and is the real neural basis of E . Because of this, content matches look evidentially weak. If so, the content-matching method—in which one tries to find the NCC by looking for content matches—looks like the wrong way to conceive of a method for finding the NCC. As we now argue, there is reason for thinking that there should instead be an emphasis on content *mismatches*.

To see why, let us look again at what inspired the content-matching method, the isomorphism constraint. The isomorphism constraint states that a neural system N is the basis of an experience E only if N matches E in content. Assuming the isomorphism constraint, if some neural system N and particular experience E mismatch in content, N can *definitely* not be the neural basis of E . Content mismatches appear of much higher evidential value than content matches. If we could rule out sufficiently many neural systems using content mismatches of this sort, narrowing in on a small number of potential neural bases – or, in the ideal scenario, just one potential basis – we would have much stronger evidence for the hypothesis that the relevant neural systems or system are or is the neural basis of E . (More on this when we discuss the potential for an “eliminative inference” below.)

To put a face on this content-mismatch method, let us look at a study carried out by Zeki (1983; see also Zeki 1982). Zeki took single-cell recordings from monkeys' V1 (primary visual cortex) while they viewed a Mondrian (an arrangement of contiguous rectangles of various colors, resembling the work of painter Piet Mondrian). The stimulus was presented in either standard lighting conditions (bathed in “white” light) or aberrant lighting conditions (bathed in, e.g., red light). Because of the primate visual system's ability to maintain color constancy (Land 1974), the aberrant lighting brings about no significant difference in the appearance of the colors of the Mondrian, even though it greatly changes the

composition of light reflected from it. A tan area on the Mondrian, for example, will continue to look tan, rather than taking on a red hue because of the lighting.⁶ Zeki found, however, that activity of cells in V1 was affected by the aberrant lighting conditions. A cell that had a preference for red light, for example, showed increased spiking even though its receptive field fell on a tan area.⁷ Since the monkey experienced that part of the Mondrian as being tan, but V1 represented it as being red, there is a content mismatch between V1 and the experience of the monkey. This rules out V1 as a neural basis of the monkey's color experience, since (according to the isomorphism constraint) the neural basis of an experience must match that experience in content.

One could debate, of course, whether this experiment successfully reveals a content mismatch.⁸ But if, after careful consideration, we found that there was strong reason for thinking that the Zeki experiment successfully presents a content mismatch, we would have strong reason for thinking that V1 is not the neural basis of color experience. Compare this to a content match: Even if, after careful consideration, we accepted that the aforementioned Tong *et al.* experiment presents a content match, this would provide but weak reason for thinking the FFA is the neural basis of face experience, as it would leave open the possibility of there being another neural system matching the face experience in content.

There appears, then, to be a stark asymmetry in the kind of justification that content matches and mismatches can provide. To us, this suggests that it is content mismatches rather than content matches that should undergird NCC research.

⁶ Clearly, we do not know this on the basis of the monkey's verbal report; rather it is something inferred from how a normal human observer experiences the colors of the Mondrian (under aberrant lighting), along with great similarities between the monkey and human visual systems.

⁷ The term "receptive field" refers to the part of the visual field to which a cell is responsive. The receptive fields of cells in earlier parts of the cortical visual system (like V1) are rather small, with them becoming gradually larger as one ascends to higher parts.

⁸ As Zeki himself points out, these results do not necessarily "imply that there are not other wavelength selective cells in monkey striate cortex whose responses do correlate with colours as perceived by us" (1982: 58).

As noted above, content mismatches make possible a method for finding the NCC, based on what is sometimes called “eliminative inference” (Mill 1843; Platt 1964; see also Kitcher 1993). In eliminative inference, support for some theory T is gained by falsifying all of the competitors of T: Through a process of elimination, one shows that theory T must be true. In the present context, the idea would be that content mismatches (perhaps along with lesion studies) could be used to eliminate hypotheses concerning the neural basis of various types of conscious experience, until only one candidate for the neural basis is left open. In any case we are able to do so, it seems we would have a much stronger kind of justification than could be provided by content matches.

Naturally, the above is merely an outline of a method for finding the NCC (for more on this, see Kozuch forthcoming). Even this brief presentation, however, suggests that the idea expressed by the matching-content doctrine is most likely off-track. According to the matching-content doctrine, the most important task facing a neuroscience of consciousness is that of finding those neural systems that match experiences in content. But there appears to be a *prima facie* case for preferring a search for content mismatches rather than content matches.

3. From Science to Philosophy: Neural Correlates and the Metaphysics of Consciousness

Assume that we are at the end of neuroscientific inquiry, and the NCC have been fully and accurately identified. It would seem that certain questions remain regarding the exact relationship between consciousness and the relevant neural structures. It is natural to take the correlation between a conscious content and a neural structure to be indicative of a more intimate connection between them. Perhaps the most sanguine view in this area is that the conscious content is in fact *identical* to the relevant neural structure – that every phenomenal property is *identical* to some neural property it correlates with. This *identity thesis* is sometimes

recommended on the grounds that it is the best explanation of the correlation, being the most parsimonious (Smart 1959). We may formulate the thesis as follows:

(I) Phenomenal properties are identical to neural properties.

Just as the correlation between Mark Twain and Samuel Clemens is no coincidence, but rather due to the fact that the man is one, so the correlation between some phenomenal property P and some neural property N is most straightforwardly explained by the hypothesis that P=N. This kind of identity thesis is the hallmark of *reductive materialism* about consciousness.

As elegant as the identity thesis is, many philosophers have taken it to be disproved by the fact that one and the same phenomenal property can be multiply realized by different neural properties (Putnam 1967). This fact is taken to disprove reductive materialism about consciousness, but not materialism as such. For it does not threaten the notion that the phenomenal facts are fully *fixed* by the neural facts. This fixing relation is sometimes captured by the notion of *metaphysical supervenience*, the idea that variation in phenomenal properties must involve variation in neural properties as well. Call this the thesis of metaphysical supervenience:

(MS) Phenomenal properties merely metaphysically supervene upon neural properties.

Such metaphysical supervenience is the hallmark of *non-reductive materialism* about consciousness.⁹

⁹ Two qualifications are in order. First, arguably metaphysical supervenience is not sufficient for materialism. At the very least, one needs to add that the particulars that have phenomenal and neural properties are all material. For a view according to which mind and matter are two separate substances, each having its own distinctive properties but in such a way that mind's properties supervene metaphysically upon matter's properties, would not be a materialist view (Papineau 1993). Secondly, as Horgan (1993) has argued, quite convincingly, taking the metaphysical supervenience of phenomenal on neural properties to be primitive and inexplicable seems offensive to the spirit of materialism. A true materialist position would have not only to posit such metaphysical supervenience, but also explain it. Horgan calls the relations of 'explained

The qualifier “merely” is needed because, at least on the standard conception of supervenience, metaphysical supervenience of A on B does not preclude *identity* of A and B; any such precluding would have to be explicit. On this conception, supervenience strictly so-called is a purely logical relation, mandating the sufficiency in all metaphysically possible worlds of B for A. This is to be contrasted with a more robust implicature of the term “supervenience,” whereby the supervenience of A on B entails the *ontological priority* of B to A. This implicature casts supervenience as an *anti-symmetrical* relation: if A supervenes on B, then B does not supervene on A.¹⁰ Strictly speaking, however, supervenience is only an *asymmetric* relation, in the sense that the supervenience of A on B entails neither the presence nor the absence of supervenience of B on A. It therefore does not preclude identity. In fact, we may plausibly see the identity relation as nothing but, or at least as underlying, two-way metaphysical supervenience: A=B just in case A metaphysically supervenes on B and B metaphysically supervenes on A.

To repeat, what motivates the move from identity, or two-way metaphysical supervenience, to one-way metaphysical supervenience is the apparent fact that a single phenomenal property can be realized by multiple neural properties. It is worth distinguishing two scenarios here. One is where a phenomenal property has a different neural realizer in the actual world, the other where it has a different realizer only in some counterfactual world. The former is *multiple realization*, the latter *multiple realizability*. Putnam’s (1967) case against reductive materialism asserted multiple realization (humans and octopi were claimed to have different realizers for pain). However, Putnam’s claim has been challenged on empirical and methodological grounds, and many philosophers have maintained that mental

supervenience’ superdupervenience. His claim is thus that materialism requires metaphysical superdupervenience of phenomenal on neural properties.

¹⁰ This implicature does preclude identity, which is symmetrical. This kind of anti-symmetrical supervenience-cum-priority relations is closely associated with what is sometimes called a “grounding relation” (see, e.g., Fine 2001).

properties probably have unique realizers in the actual world, though may well have other realizers in counterfactual worlds.¹¹

In any case, as is well-known, the metaphysical supervenience of consciousness on neural properties is also controversial – indeed arguably the central controversy of the philosophy of mind of the past decade. Wielding a variety of (mostly epistemic) arguments, assorted dualists have claimed that phenomenal properties cannot ultimately metaphysically supervene on neural properties (see, e.g., Chalmers 1996). Nonetheless, they insist that it is possible to explain the correlation between phenomenal and neural properties, namely, as due to certain primitives laws of nature that dictate the co-instantiation of neural and phenomenal properties. The existence of such laws of nature guarantees that even though phenomenal properties do not *metaphysically* supervene on neural properties, they nonetheless *nomologically* supervene on them. Call this the thesis of nomological supervenience:

(NS) Phenomenal properties merely nomologically supervene upon neural properties.

Such nomological supervenience is the hallmark of what Chalmers (1996) calls *naturalistic dualism*.¹² It is plausible to construe the relevant laws of nature as *causal* laws, laws of the form “under conditions C, neural feature N causes phenomenal property P.” This causal version of naturalistic dualism is reminiscent of traditional *emergentist* views of consciousness, according to which phenomenal property instantiations causally emerge from neural property instantiation. In this emergentist variety, naturalistic dualism attempts to explain correlation by

¹¹ Such multiple realizability would suffice to undermine the identity of phenomenal and neural properties, given that identity is in all likelihood a necessary relation (if it holds at all, it holds necessarily). If it turns out some identity is contingent (Gibbard 1975), however, then multiple realizability may cohabit with the identity thesis.

¹² As before, the “merely” is needed because the obtaining of nomological supervenience relation between two properties does not preclude the obtaining of a stronger supervenience relation. Also, here too the view would qualify as genuinely naturalistic only if the particulars that have the phenomenal and neural properties are all material particulars.

causation: the thought is that the best explanation of neural-phenomenal correlation is neural-phenomenal causation.

It is interesting to note here that, although philosophers take the obtaining of *mere* nomological supervenience to demonstrate the truth of dualism, scientists typically regard it as underwriting materialism. The guiding idea seems to be that any phenomenon that could be shown to be causally integrated in an ordinary way into the web of natural laws connecting natural phenomena should be regarded as a physical phenomenon. Thus insofar as consciousness can be embedded into the causal web of the material world, it ceases to present a challenge to a materialist conception of the world. This kind of “inclusive materialism” (Kriegel 2007) closer to naturalistic dualism than to reductive materialism in some respects but closer to reductive materialism in others. Thus, unlike dualism both inclusive and reductive materialism deny that anything is non-physical; on the other hand, unlike reductive materialism both inclusive materialism and naturalistic dualism deny that consciousness is *nothing but* some already familiar physical property.

We have surveyed a number of explanations of neural correlation, in a decreasing order of metaphysical exaction: reductive materialism appeals to identity, non-reductive materialism to mere metaphysical supervenience, and naturalistic dualism to mere nomological supervenience. A final view worth stating explicitly is *non-naturalistic* dualism, according to which any supervenience of the phenomenal on the neural is merely *contingent*: not grounded in any natural laws, it is purely *accidental*. Call this the thesis of *contingent supervenience*:

(CS) Phenomenal properties merely contingently supervene upon neural properties.

Although conceding that neural properties are contingently sufficient for phenomenal properties, this view (in virtue of the “merely”) deems this sufficiency positively inexplicable. It is in this sense that the view is non-naturalistic.¹³

Given the various option surveyed in this section, the metaphysical question left over once the science of consciousness has identified the NCC can be posed succinctly as follows: what is the best metaphysical explanation of the correlation between phenomenal and neural properties? Non-naturalistic dualism offers no explanation of this, of course, but reductive materialism, non-reductive materialism, and naturalistic dualism can each be thought of as doing so. According to reductive materialism, what explains the correlation is the identity of phenomenal and neural properties. For non-reductive materialism and naturalistic dualism, the explanation is in terms of metaphysical and nomological supervenience, respectively.¹⁴

One problem here, however, is that, as logical relations, it is hard to see in what sense metaphysical or nomological supervenience of the phenomenal on the neural can be said to *explain* the correlation between them. Metaphysical and nomological supervenience *entail* contingent supervenience, but entailment is not yet explanation. To address this issue, non-reductive materialism and naturalistic dualism could identify a specific relation that would underlie metaphysical/nomological supervenience and would be the *reason* why the supervenience holds. For naturalistic dualism, the relation of *causation* seems to be

¹³ Interestingly, non-naturalistic dualism not only fails to explain neural correlation, in a sense it fails to describe it as well. For correlation between A and B implies more than the contingent sufficiency of A for B – it involves the contingent sufficiency of B for A as well. That is, correlation is a symmetric relation. For just as identity can be seen as (underlying) two-way metaphysical supervenience, correlation is naturally understood as two-way contingent supervenience. (This is so, of course, only so long as supervenience is not construed as involving priority relations.)

¹⁴ However, if we do construe correlation as two-way contingent supervenience, as per the previous note, it becomes a problem that metaphysical and nomological supervenience are one-way supervenience, whereas correlation is two-way supervenience. To address this issue, both non-reductive materialism and naturalistic dualism would have to add to MS and NS a claim of contingent supervenience going the other way – essentially, making the neural properties that are metaphysically or nomologically sufficient for phenomenal properties also contingently necessary for them.

of the right kind: the reason phenomenal properties nomologically supervene on neural properties is that they are *caused* by them. For non-reductive materialism, it is natural to appeal to a relation of *constitution* as underlying metaphysical supervenience: phenomenal properties supervene on neural ones because they are *constituted* by those neural properties.

The relation of constitution, which is supposed to be stronger (more intimate) than causation but weaker (less intimate) than identity is also more *prima facie* mysterious than identity and causation, because less familiar from other contexts of inquiry.¹⁵ Nonetheless, pending a full analysis of the nature of this relation, the philosopher can maintain that something like it *must* be posited to capture the possibility of mere metaphysical supervenience – something stronger than nomological supervenience but weaker than two-way metaphysical supervenience (i.e., identity). Short of providing a full analysis of the relevant notion of constitution, we can enhance its intelligibility of through certain examples and analogies (e.g., to material constitution of the statue by the clay). We may also hold that in its intrinsic opacity constitution is forsooth on a par with causation: just as the causal “secret connexion” is a sort of metaphysical *je ne sais qua* that underlies constant conjunction, so the constitutive connection is a sort of metaphysical *je ne sais qua* that underlies metaphysical supervenience.

With these supplementations in place, we now have three competing “metaphysical hypotheses” to explain the fact that phenomenal properties correlate with neural properties:

(H1) Phenomenal properties are identical to neural properties.

(H2) Phenomenal properties are constituted by neural properties.

(H3) Phenomenal properties are caused by neural properties.

¹⁵ Indeed, this is part of the *raison d'être* of the present volume.

The three hypotheses appear to be *empirically equivalent*: they make the same predictions (and retrodictions) regarding what phenomenal properties would be instantiated in what neural conditions.¹⁶ What will distinguish them, then, are how they “score” on the so-called “super-empirical virtues”: parsimony, modesty, conservatism, unity, simplicity, and so on.¹⁷

This is not the place to conduct a wide-ranging assessment of the overall comparative strengths and weaknesses of H1-H3 along these various dimensions. We will restrict ourselves to preliminary remarks here, indicating where our own preferences lie.

As far as parsimony is concerned, it is fairly clear that H1 fares the best, and H3 the worst. The status of H2 is harder to assess. On the one hand, constitution does not amount to identity, so when A constitutes B, A and B are numerically distinct, and we have here as many entities as we do when A causes B. On the other hand, on a natural understanding of the notion of constitution, when A constitutes B we are justified in saying that B is “nothing but” A, which may suggest that B is an “ontological free lunch.”¹⁸ It may well be that the notion of parsimony itself would have to be disambiguated, such that H2 is as parsimonious as H1 on some disambiguations but as H3 on others.

¹⁶ According to some views in the philosophy of science – most notably, logical positivism – when two theories are empirically equivalent in this way, their disagreement cannot be substantive. If this is right, then there is no place for a metaphysics of consciousness in addition to the science of consciousness, and in general logical positivists thought metaphysics was nonsense (Carnap 1932). This is often regarded as untenable, for reasons we cannot go into here. A gentler kind of deflationary attitude can be found in van Fraassen (1980), according to whom empirically equivalent scientific theories are such that we may choose to *accept* one over the other for non-cognitive reasons, but we have no reason to *believe* one more than the other. This would deflate the metaphysical issue under consideration without quite dismissing it as non-substantive. This is much more plausible than logical positivism (Kriegel 2011), though of course highly controversial. More on that below.

¹⁷ For a fuller development of an approach to the metaphysics of consciousness that emphasizes comparisons of theoretical virtues in this way, see Biggs forthcoming. Biggs also offers some explicit statements of the nature of some of these virtues. For a classic discussion of the nature of these virtues, see Quine and Ullian 1970.

¹⁸ Some potentially imperfect analogies here might be the idea that the table is constituted by its legs, its top, and their spatial arrangement, and is therefore nothing but all that, or that the table’s redness is constituted by its vermilion-ness, and is therefore nothing but it.

The situation with modesty appears to be the converse of that with parsimony: since identity is the strongest, most demanding of the three relations, and causation the weakest, it would seem that H3 is the most modest of the three hypotheses and H1 the least modest, with H2 lying somewhere in-between. Note well: this does not imply that parsimony and modesty cancel each other out, as one may weigh more in a final “tally” of comparative virtues.

How the hypotheses rank in terms of conservatism, thus in terms of continuity with the reigning belief, depends on what one designates as the reigning belief in the relevant sense. As far as *folk* belief is concerned, there is good reason to think it generally dualist (Bloom 2004). Even though it is not committed to the machinery of nomological supervenience, causal laws, etc., folk belief is thus most closely aligned with H3, and farthest removed from H1. It is not entirely clear, however, that folk belief should function here as the kind of belief the hypotheses should attempt to depart least from, other things being equal. One might wish to designate the dominant philosophical conviction as that belief – the belief in need of conserving. From this perspective, it would seem that H2 is the most conservative, as non-reductive materialism has approached the status of orthodoxy since the late sixties.¹⁹ In any event, conservatism does not seem to be of central importance (great weight) in the present context, as philosophical theorization does not always exhibit the pattern of continuity and directedness that, say, the history of physics does.

A theoretical virtue that *is* clearly of great importance is *unity*: the more unified an overall theory of the world, the more virtuous it is.²⁰ Here it is clear that,

¹⁹ Relatedly, in a recent survey of 3226 philosophers, 56.4% of respondents said they were physicalists, 27% that they were “non-physicalists” (which presumably covers not only dualists, but also neutral monists and the like). Among philosophers of mind specifically, the proportion was even more acute: 61.2% physicalists and 21.9% non-physicalists (see <http://philpapers.org/surveys/results.pl>).

²⁰ There is a legitimate question of why this is so, which we will not take up here, except to note that according to Kitcher (1981), unity enhances explanatoriness; that the Humean principle of the ‘unity of nature’ recommends a unified theory of nature; and that insofar as we are willing to grant esthetic

since all other non-microphysical properties are thought to be identical to or constituted by microphysical properties, any theory of the world that denies this of phenomenal properties would be less unified than one that does not. Likewise, our overall image of the world seems to cast properties from the manifest image of the world as generally nothing but some properties from the scientific image of the world (Stoljar 2006); excepting phenomenal properties diminishes the unity of our overall image of the world. Thus H3 scores much lower on unity than H1 and H2. Which of H1 and H2 scores higher depends on whether we take the general rule to involve identity or constitution. Is water, for example, strictly identical to H₂O or merely constituted thereby? Most philosophers go for identity here, though there are very good arguments for mere constitution – see especially Johnston 1997. Furthermore, if (as is plausible) multiple realizability applies not only to phenomenal properties but to all special-science properties (Fodor 1974), then a constitution-based theory of the world is more unified than an identity-based one, making H2 score higher than H1 on unity.

The foregoing discussion has treated the super-empirical virtues as recommending belief in the truth of the theories that exhibit them: the more virtuous a theory or hypothesis, the higher our credence in it should be. It should be pointed out, however, that some philosophers deny this. In van Fraassen's (1980) constructive empiricism, for instance, the super-empirical virtues are taken to provide pragmatic or instrumental reasons for *adopting* theories, but not epistemic reasons for *believing* them. While this is quite a radical view, its fundamental attraction is straightforward: for no super-empirical virtue is it particularly clear how it is supposed to be truth-conducive; if it is not truth-conducive, the fact that a theory exemplifies does not make it more likely that the world is the way the theory says it is; and if it does not make it more likely that the world is the way the theory says it is, then it is unclear why we should *believe* the theory. We have considerable

virtues of a theory, such as unity, simplicity, and symmetry, a role in theory construction, unity would certainly enhance esthetics.

sympathy toward this line of reasoning and hope to explore it more fully in future work. Its upshot would be that there may be no way to choose among H1-H3: since they are empirically equivalent, and the super-empirical virtues are not truth-conducive, there is no epistemic reason to believe one more than the others.

In any case, this is not the place to produce a final verdict on the matter. The main purpose of this section has been to point out one way a philosophy of consciousness is required to go beyond the science of consciousness: namely, by producing metaphysical hypotheses about the ultimate relationship between consciousness and neural activity that would *explain* the correlation between them.

4. Conclusion

In this paper, we have discussed two ways in which the philosophy of consciousness is relevant to the science of consciousness. The first concerns the precise analysis of the notion of a neural correlate of consciousness (understood as a neural basis for it) and the proper methodology for studying it; this was the subject of §2. The second concerns potential inferences to the best explanation from phenomenal-neural correlation to various more intimate phenomenal-neural relations; this was the topic of §3. Our main claims have been two: that a methodology for NCC research that focuses on content mismatches may be more epistemically sound than one relying on content matches, and that the choice between dualist and materialist theories of consciousness can be profitably cast as a choice between different potential explanations of the correlation between consciousness and its NCC, whatever it turns out to be.²¹

References

²¹ For comments on a previous draft, we are very grateful to Michael Bruno.

- Biggs, S. Forthcoming. "Abduction and Modality." *Philosophy and Phenomenological Research*.
- Block, N.J. 1995. "On a Confusion About the Function of Consciousness." *Behavioral and Brain Sciences* 18: 227-247.
- Block, N.J. 2007. "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience." *Behavioural and Brain Sciences* 30: 481-499.
- Bloom, P. 2004. *Descartes' Baby*. New York: Basic Books.
- Carnap, R. 1932. "The Elimination of Metaphysics through Logical Analysis of Language." Trans. by A. Pap. In A.J. Ayer (ed.), *Logical Positivism*. New York: Free Press, 1959.
- Chalmers, D. J. 1996. *The Conscious Mind*. Oxford and New York: Oxford UP.
- Chalmers, D.J. 2000. "What Is a Neural Correlate of Consciousness?" In T. Metzinger (ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Cambridge MA: MIT Press.
- Crick, F. 1994. *The Astonishing Hypothesis*. New York: Charles Scribner's Sons.
- Crick, F. 1996. "Visual perception: rivalry and consciousness." *Nature* 379: 485-6
- Crick, F. and C. Koch 2003. "A framework for consciousness." *Nature Neuroscience* 6: 119-125.
- Fine, K. 2001. "The Question of Realism." *Philosophers' Imprint* 1: 1-30.
- Frith, C., R. Perry, and E. Lumer. "The neural correlates of conscious experience: an experimental framework." *Trends in Cognitive Sciences* 3: 105-114.
- Gibbard, A. 1975. "Contingent Identity." *Journal of Philosophical Logic* 4: 187-221.
- Horgan, T. 1993. "From Supervenience to Superdupervenience: Meeting the demands of a Material World." *Mind* 102: 555-86.
- Johnston, M. 1997. "Manifest Kinds." *Journal of Philosophy* 94: 564-583.
- Kanwisher, N. 2001. "Neural events and perceptual awareness." *Cognition* 79: 89-113.
- Kitcher, P. 1981. "Explanatory Unification." *Philosophy of Science* 4: 507-531.
- Kitcher, P. 1993. *The Advancement of Science*. Oxford: Oxford University Press
- Kozuch, B.P. Forthcoming. *Using Eliminative Methodology in the Scientific and Philosophical Studies of Consciousness*. PhD Dissertation, University of Arizona.
- Kriegel, U. 2007. "Gray Matters." *Journal of Consciousness Studies* 14: 96-116.

- Kriegel, U. 2009. *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford UP.
- Kriegel, U. 2011. "Two Defenses of Common-Sense Ontology." *Dialectica* 65.
- Land, E.H. 1974. "The retinex theory of colour vision." *Proceeding of the Royal Institution Great Britain* 47: 23-57.
- Mill, J.S. 1843/2002. *A System of Logic*. Honolulu: University Press of the Pacific.
- Miller, S.M. 2007. "On the correlation/constitution distinction problem (and other hard problems) in the scientific study of consciousness." *Acta Neuropsychiatrica* 19: 159–176.
- Noë, A. and E. Thompson 2004. "Are There Neural Correlates of Consciousness?" *Journal of Consciousness Studies* 11: 3-28.
- Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.
- Platt, J. R. 1964. "Strong Inference." *Science* 146: 347-53.
- Putnam, H. 1967. "The Nature of Mental States." Originally published as "Psychological Predicates," in W.H. Capitan and D.D. Merrill (eds.), *Art, Mind, and Religion*. Reprinted in D.M. Rosenthal (ed.), *The Nature of Mind*. Oxford: Oxford University Press.
- Quine, W.V.O. and J.S. Ullian 1970. *The Web of Belief*. New York: Random House.
- Rees, G., G. Kreiman, and C. Koch 2002. "Neural correlates of consciousness in humans." *Nature Reviews Neuroscience* 3: 261-70.
- Smart, J.J.C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68: 141-156.
- Stoljar, D. 2006. *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*. Oxford: Oxford University Press.
- Teller, D.Y. and E.N. Pugh 1983. "Linking propositions in color vision." In J. Mollon and L. Sharpe (eds.), *Colour Vision*. London: Academic Press.
- Tong, F., K. Nakayama, J.T. Vaughan, and N. Kanwisher 1998. "Binocular Rivalry and Visual Awareness in Human Extrastriate Cortex." *Neuron* 21: 753–759.
- van Fraassen, B.C. 1980. *The Scientific Image*. Oxford and New York: Oxford University Press.
- Zeki S. 1982. "Do wavelength selective cells in monkey striate cortex respond to colours?" *Journal of Phvsiology* 330: 57-58P.
- Zeki, S. 1983. "Colour coding in the cerebral cortex: the reaction of cells in the monkey visual cortex to wavelengths and colours." *Neuroscience* 9: 741-65.