Beyond the Neural Correlates of Consciousness

Uriah Kriegel

Forthcoming in *Oxford Handbook of the Philosophy of Consciousness*

Abstract. The centerpiece of the scientific study of consciousness is the search for the neural correlates of consciousness. Yet science is typically interested not only in discovering correlations, but also – and more deeply – in explaining them. When faced with a correlation between two phenomena in nature, we typically want to know *why* they correlate. The purpose of this chapter is twofold. The first half attempts to lay out the various *possible* explanations of the correlation between consciousness and its neural correlate – to provide a sort of "menu" of options from which we probably would ultimately have to choose. The second half raises considerations suggesting that, under certain reasonable assumptions, the choice among these various options may be *in principle* underdetermined by the relevant scientific evidence.

Introduction

The centerpiece of the scientific study of consciousness is the search for the neural correlates of consciousness (Morales & Lau, this volume). Yet science is typically interested not only in correlation relations among natural phenomena, but also in causal and constitutive relations. Often, these causal and constitutive relations are posited as *explanations* of why certain phenomena correlate. To treat correlations as brute and inexplicable is to acquiesce in mysterious aspects of nature, somewhat as the spiritualist revels in "weird coincidences." It is surely the mandate of intellectual inquiry in general and science in particular to address such coincidences and shed light on why they hold.

Consider Leibniz's "pre-established harmony theory" of the connection between mind and body (i.e., the hypothesis that at the beginning of time God established a correlation between the two, so that whenever certain changes occur in some creature's brain activity, certain events will take place simultaneously in the creature's stream of consciousness, and vice versa). In an obvious sense, this is an extremely anti-scientific approach to the correlation between consciousness and brain activity. Yet even this approach ventures *some* kind of explanation. It does not posit the correlation as brute and inexplicable. Instead, it offers a *reason* for the correlation. In so doing, it tries to make it *intelligible*. Insofar as the "brute correlation" approach we find in current scientific research on consciousness does not even attempt to do that, it might be claimed to be even more mysterian.

With this in mind, it is natural for us to hope that the current science of consciousness could offer more than just an *identification* of the neural correlate of consciousness – that it might offer an *explanation* of why the correlation holds. The purpose of the chapter is twofold. In the first half (§§1-2), I want to lay out the various *possible* explanations of the correlation between consciousness and its neural correlate. The idea is to provide a sort of "menu" of options from which we would probably have to choose – and to link it to traditional metaphysical positions on the problem of consciousness. In the chapter's second half (§§3-4), however, I will raise considerations suggesting that, under certain reasonable assumptions, the choice among these various options may be in principle underdetermined by the relevant scientific evidence, so that the traditional metaphysical positions may be *empirically equivalent*. I should stress that I am not entirely persuaded that the claim is true; still, the considerations supporting it strike me as quite powerful and worth contending with. If it is true, however, then the choice between different explanations of phenomenal-cerebral correlations cannot in principle be a scientific one. It must be a matter of *philosophical*-theory choice.

1. Neural Correlates and Explanatory Hypotheses

It is widely thought that materialism and dualism about consciousness are both compatible with the eventual discovery of the neural correlates of consciousness (NCC). One way to think of this is in terms of what we can *infer from* a correlation. Suppose, purely for the sake of exposition, that the NCC is neural synchronization with above-baseline activity in the dorsolateral prefrontal cortex (dlPFC) (Lau and Passingham 2006, Kriegel 2009 Ch.7, Rounis et al. 2010). Often – though, of course, not always – correlation is an indicator of *causation*. When we notice a correlation between the striking of matches and their lighting up, we infer that striking a match *causes* it to light. This is a fairly standard form of so-called inference to the best explanation, arguably the central mode of scientific inference (Harman 1965, Lipton 1992). The reasoning proceeds as follows:

1) Match-striking correlates with match-lighting;
2) The best explanation of this is that match-striking *causes* match-lighting; therefore, plausibly,
3) Match-striking causes match-lighting.

In a similar vein, we might infer from the correlation between neural synchronization with dlPFC activity and consciousness that synchronization with dlPFC activity *causes* consciousness – that this particular neural activity brings about, is responsible for the production of, consciousness. More generally, the reasoning is this:

1) The NCC correlates with consciousness;
2) The best explanation of this is that the NCC *causes* consciousness; therefore, plausibly,
3) The NCC causes consciousness.

This is often the most natural explanatory hypothesis for the correlation between two phenomena: that one is simply the cause of the other.

As is well known, however, the direction of causation is often in question when explanatory inferences are performed. The largest concentration of asthmatics in the US lives in Tucson, Arizona, despite the fact that the Sonora desert's extraordinarily dry air is supposed to *help* with asthma. Obviously, the explanation of this tight correlation between dry air and incidence of asthma is

not that Tucson's dry air causes people to develop asthma. On the contrary, it is that sufficiently severe asthma causes people to relocate to Tucson. By the same token, a perfectly coherent possibility is that synchronization with dlPFC activity is not so much the *cause* of consciousness as its *effect*. In this picture, there is a sort of 'downward causation' by which consciousness alters the state of the brain, a downward causation characteristic of what Chalmers (2002) calls "type-D dualism."[1] It is thus epistemically possible to pursue the following piece of reasoning:

1) The NCC correlates with consciousness;
2) The best explanation of this is that consciousness *causes* the NCC; therefore, plausibly,
3) consciousness causes the NCC.

The difference between this "reverse causal hypothesis" and the "more straightforward" causal explanation concerns what causal direction is taken to *better* explain the correlation between consciousness and the NCC. In this section I do not comment on the question of the possible hypotheses' relative merits; my goal is merely to set out the menu of options.

A further option, when faced with a correlation between two phenomena, is to maintain that there is a *third cause* responsible for the occurrence of each phenomenon independently – and thus responsible for their correlation. The correlation between lightning and thunder, for example, is best explained neither by the hypothesis that lightning causes thunder nor by the hypothesis that thunder causes lightning. Rather, there is a third element that causes both: the collision of ice and water particles inside a cloud causes lightning, on the one hand, and thunder, on the other. Since it causes both, it also causes their correlation. Likewise, one might hold that some third factor might cause the occurrence of the NCC, on the one hand, and consciousness, on the other. Here the general explanatory inference looks like this:

1) The NCC correlates with consciousness;
2) The best explanation of this is that there is some third element that causes both the NCC and consciousness; therefore, plausibly,

3) There is some third element that causes both the NCC and consciousness.

In itself, this explanatory inference is neutral on what the third cause is – what the "X factor" is. This means that there are as many versions of this inference as there are potential X factors. One way to understand "quantum-mechanical approaches" to consciousness (e.g., Hameroff and Penrose 1996) might be a version of the above causal inference. The thesis is that certain quantum-mechanical events cause both changes in the brain and changes in consciousness, thus accounting for the correlation between the two. Another version is of course Leibniz's pre-established harmony theory, where God's will acts as the third cause.

Sometimes causal hypotheses are not the best explanations of correlation at all. There is a tight correlation between lifting something out of a shop and breaking the law. But this is not because shoplifting *causes* lawbreaking, but because shoplifting *is* lawbreaking. We may say that the relation between shoplifting and lawbreaking is not causal but *constitutive*: shoplifting *constitutes* breaking the law. In this case, the shoplifting breaks the law *by definition* rather than by *causation*. But arguably, there are cases where a constitutive hypothesis explains correlation better than a causal hypothesis even where no definitions are involved. When scientists first observed the remarkable correlation between water and the molecular structure known as H_2O, they did not infer that H_2O must *cause* water; instead, they inferred that H_2O must *be* water – that there is nothing more to water over and above H_2O. That is, large enough collections of H_2O molecules *constitutes* bodies of water. Here the inference is from correlation to constitution. The same reasoning can be applied to the correlation between consciousness and the NCC (see Hohwy 2011):

1) The NCC correlates with consciousness;
2) The best explanation of this is that the NCC *constitutes* consciousness; therefore, plausibly,
3) The NCC constitutes consciousness.

There is thus a competition between two interpretations of the correlation between consciousness and its neural correlate: a *causal* interpretation and a

constitutive interpretation. It is the latter that characterizes physicalist theories of consciousness (see Jackson, this volume).

Moreover, just as the causal interpretation admits of two opposing "directions" – the NCC causes consciousness and consciousness causes the NCC – so the constitutive interpretation does. In addition to the above constitutive hypothesis, the opposing hypothesis according to which neural structures are themselves ultimately constituted by consciousness is coherent as well. This is, in effect, the view of idealists, such as Michael Pelczar (2015, this volume), who maintain that ultimate reality is in fact phenomenal. Here the reasoning proceeds as follows:

1) The NCC correlates with consciousness;
2) The best explanation of this is that consciousness *constitutes* the NCC; therefore, plausibly,
3) Consciousness constitutes the NCC.

This reasoning may also be taken to characterize the view of certain panpsychists, such as Greg Rosenberg (2005), who hold that some phenomenal properties underlie all physical properties.

Likewise, corresponding to the "third cause" explanatory hypothesis there is certainly the option of a *third constitutor* hypothesis. That is, there might be an "X factor" that in one manifestation (perhaps in combination with some micro-properties) constitutes the NCC and in another (with other properties) constitutes consciousness. Here the reasoning is:

1) The NCC correlates with consciousness;
2) The best explanation of this is that there is some third element that constitutes both the NCC and consciousness; therefore, plausibly,
3) There is some third element that constitutes both the NCC and consciousness.

As we will see below, certain versions of "neutral monism," a view that goes back at least to Spinoza, are in effect committed to such a third-constitutor view (see Coleman & Goff, this volume).

In summary, I have presented six possible explanations, or interpretations, of the correlation between consciousness and whatever turns out to be its neural correlate (e.g., synchronization with dlPFC). These are:

1) CAUSATION: consciousness is caused by the NCC.
2) REVERSE CAUSATION: the NCC is caused by consciousness.
3) THIRD CAUSE: consciousness and the NCC are both caused by some third element.
4) CONSTITUTION: consciousness is constituted by the NCC.
5) REVERSE CONSTITUTION: the NCC is constituted by consciousness.
6) THIRD CONSTITUTOR: consciousness and the NCC are both constituted by some third element.

In the next section, I link this menu to contemporary philosophical theories about the mind-body problem.

2. Explanatory Hypotheses and Metaphysical Positions

Of the six explanatory hypotheses just laid out, the most important may well be CAUSATION and CONSTITUTION. For something like it appears to mark the crucial disagreement between moderate forms of materialism and moderate forms of dualism. There are radical forms of materialism according to which consciousness does not exist (see Irvine & Sprevak, this volume), and therefore has no correlates, as well as radical forms of dualism according to which conscious activity unfolds in complete freedom from brain activity. We may call these *eliminative materialism* and *non-naturalistic dualism*. In modern philosophy of mind, *non-eliminative* forms of materialism and *naturalistic* forms of dualism are the more dominant views. Both agree that conscious activity depends in some way on brain activity. The disagreement concerns the type of dependence involved. For the materialist, consciousness depends ontologically, metaphysically, or constitutively, on brain activity. For the dualist, the dependence is merely "natural," nomological, or causal.

There are many different ways to try to capture this difference more precisely. In late twentieth-century philosophy of mind, the notion of

supervenience played a crucial role in this area. In particular, the distinction between (non-eliminative) materialism and (naturalist) dualism was taken to come down to the choice between *metaphysical* and merely *nomological* supervenience (Chalmers 1996). Roughly, the idea is that for the materialist, in no metaphysically possible world could there be variation in conscious activity without corresponding variation in brain activity, whereas for the dualist, this could happen in some metaphysically possible world, though not in any nomologically possible world (i.e., in any world that has the same laws of nature as the actual world).

In more recent philosophy of mind, the notion of supervenience is often taken to point at a mere *formal symptom* of underlying substantive connections between the supervenient and the subvenient. We might distinguish between "metaphysical grounding" and "natural grounding" as the two substantive connections underlying metaphysical and nomological supervenience; in which case (non-eliminative materialism) would claim that consciousness is metaphysically grounded in the NCC whereas (naturalistic) dualism would claim that consciousness is naturally grounded in the NCC. Alternatively, we might call "grounding" the connection whose diagnostic symptom is metaphysical supervenience and "emergence" the connection whose diagnostic symptom is nomological supervenience; in which case materialism would claim that consciousness is grounded in the NCC whereas dualism would claim that consciousness merely emerges from the NCC.

However we mark this difference, the distinction between causal and constitutive connections seems to go to the core of the distinction. The notion that a conscious state is *constituted* by its neural correlate is of a piece with the ideas that the former is grounded in, or metaphysically supervenes upon, the latter. The notion that a conscious state is *caused* by its neural correlate is of a piece with the ideas that the former emerges from, or nomologically supervenes upon, the latter. Indeed, a constitutive connection would presumably have metaphysical supervenience for a symptom and a causal connection would presumably have nomological supervenience for a symptom (since causation certainly obeys the principle "same causes, same effects"). Thus we may take

CONSTITUTION and CAUSATION to capture the key difference between non-eliminative materialism and naturalistic dualism.

As for non-naturalistic dualism, it can play out in two very different ways. The first is that there is no correlation whatsoever between consciousness and the NCC. This view can be seen as making the prediction that the scientific search for the NCC will end with failure. The second option, however, is that although conscious activity unfolds in complete freedom from brain activity, the opposite does not hold, perhaps because with REVERSE CAUSATION is true: conscious activity casually determines what the neural process our brain undergoes. Both views, to repeat, are non-naturalistic forms of dualism. They do not represent a significant background positions on the mind-body problem in contemporary philosophy and cognitive science. I am airing them here just for the sake of completeness.

In the same vein, we may note that REVERSE CONSTITUTION is a very natural development of traditional *idealism*. Some care is needed here, however. Some idealists, such as Berkeley, hold that the physical does not exist – nothing is physical (Berkeley 1710). A fortiori, then, there are no brains and no neural processes. Accordingly, there is neural correlate of consciousness. All there is is consciousness. This is to be distinguished from the view, perhaps Plato's and/or Leibniz's, that the physical does exist but is ultimately constituted by the phenomenal. We may distinguish the two views by calling the former view *eliminative idealism* and the latter *non-eliminative idealism*. Now, while eliminative idealism denies the existence of phenomenal–neural correlations (just as eliminative materialism does), non-eliminative materialism appears to be committed to REVERSE CONSTITUTION.

What about THIRD CAUSE and THIRD CONSTITUTOR? The latter fits rather naturally with certain forms of Russellian monism, namely, those that posit fundamental properties of the universe that are neither phenomenal nor physical but at the same time are both *proto*-phenomenal and *proto*-physical: some combinations or aggregates of them somehow constitute phenomenal properties, others somehow constitute physical properties (including neural properties). If this is one's view of the ultimate connection between the

consciousness and brain activity, then one takes there to be a third type of ingredient in the universe that constitutively underlies both consciousness and its correlated neural processes. That is, one is committed to THIRD CONSTITUTOR. Now, there are also forms of version of Russellian monism in which the proto-physical properties are taken to be phenomenal rather than *proto*-phenomenal properties. Those versions are rather committed to REVERSE CAUSATION, and essentially collapse into non-eliminative idealism.

Insofar as THIRD CONSTITUTOR is a kind of Russellian monism, we could think of THIRD CAUSE as corresponding to a kind of "Russellian dualism." The idea is that there is some third type of property, neither phenomenal nor physical, different combinations of which somehow *causally* bring about instantiations of phenomenal properties and instantiations of physical properties. As noted, quantum theories of consciousness may be seen to be committed to something like this.

Thus we can map the six explanatory hypotheses laid out at the end of the previous section onto six metaphysical positions on the ultimate connection between phenomenal and neural properties: CAUSATION corresponds to naturalistic dualism; REVERSE CAUSATION corresponds to (some versions of) non-naturalistic dualism; THIRD CAUSE corresponds to certain quantum theories; CONSTITUTION corresponds to non-eliminative materialism; REVERSE CONSTITUTION corresponds to non-eliminative idealism; THIRD CONSTITUTOR corresponds to (certain versions of) Russellian monism. What this suggests is that many of the competing metaphysical positions on the mind-body problem can be paired with specific interpretations of the correlation between consciousness and the NCC.[2]

In this way, our sixfold scheme allows us to see how the choice among various metaphysical positions on the problem of consciousness reduces to a choice among different explanatory hypotheses regarding the correlation between consciousness and its NCC. It allows us to reframe the philosophical problem of consciousness as the following question: Which is the best explanatory inference to make from the correlation between consciousness and

the NCC? Which offers the most plausible explanatory hypothesis about why this correlation exists?

3. Explanatory Hypotheses and Empirical Equivalence

How should we go about choosing among the options before us? In general, choosing among alternative explanatory hypotheses is based on two kinds of consideration. First, there is the question of *empirical adequacy*: which of the competing hypotheses accommodates the empirical data best. Secondly, there is the question of *theoretical adequacy*: which of the competing hypotheses scores highest with respect to the theoretical (or "superempirical") virtues, such as simplicity, parsimony, conservatism, modesty, cohesion/coherence, unity, elegance, fecundity, testability, and so on (see Quine and Ullian 1970). In this section, I want to raise the epistemic possibility that at least CONSTITUTION and CAUSATION – arguably, the two leading explanatory hypotheses in our scheme – may be empirically equivalent, in the sense of being exactly equal in empirical adequacy. In the next section I will briefly consider the consequences such empirical equivalence would have for the choice between them.

In trying to pull CAUSATION and CONSTITUTION apart experimentally, the first order of business should be to seek empirical symptoms of the difference between causal and constitutive relations in general – in the hope that we might be able to exploit these in the present context as well. There are two main empirical symptoms of the causal/constitutive difference: one has to do with time lag, the other with mediating mechanism. The hope is that at least one of these could help us produce *discordant predictions* out of CAUSATION and CONSTITUTION.

Let us start with the issue of time lag. It is plausible to suppose that, while there is always a time lag between cause and effect (the former *precedes* the latter), constitutor and constitutee (if you will) are always *simultaneous*. Thus, the presence of H_2O in some location does not precede the presence of water, but the striking of a match does precede its lighting up. Of course, like everything else in philosophy, the temporal lag between cause and effect *has* been

contested (Huemer and Kovitz 2003). If there is no temporal lap between cause and effect, then temporal considerations will offer no empirical symptom of the difference between causal and constitutive connection. Let us grant for the sake of argument, however, that causes do precede their effects, whereas constitutors do not precede their constitutees. Applied to the choice between CAUSATION and CONSTITUTION, we might suppose that if the NCC is *causally* connected to consciousness, then its occurrence will precede the onset of consciousness ever so slightly, whereas if it is *constitutively* connected to consciousness, it will be strictly simultaneous therewith. This is one empirical symptom of the difference between causal and constitutive connections.

As for mechanism, causal connections are typically mediated by a mechanism, whereas constitutive connections are not. Thus, when investigating the connection between match-striking and match-lighting, it is possible to "go deeper" and discover the mechanism that mediates the causing of the latter by the former. Typically, this means exposing a series of intermediary causal transactions at a more fundamental level of reality – in this case, chemical interactions involving sulfur, phosphorus, oxygen, and so on. In general, when A causes B, it is often the case that this is mediated through a series of finer-grained causal transactions – A causes E_1, E_1 causes E_2, E_2 causes E_3, … , E_{n-1} causes E_n, and E_n causes B. The only exception to the existence of a mediating mechanism concerns causal transactions at the "bottom level" of reality, which must be brute and unmediated, since we cannot "go deeper" and seek even more fundamental transactions mediating them. (More on that presently.) In contrast with all this, when A *constitutes* B, such that B is *nothing but* A, there is no expectation that there be "intermediate stages" of "nothing-but-ness." For A to constitute B, it is not necessary that there be some series X_1 … X_n such that A constitutes X_1, X_1 causes X_2, … , X_{n-1} causes X_n, and X_n constitutes B. Accordingly, to choose between CAUSATION and CONSTITUTION, we might seek a series of intermediary correlates between consciousness and the NCC. If such a series can be found, however short, this could indicate a causal connection between the two. If none can be found (despite sustained attempts to reveal one), that could indicate a constitutive connection.

Unfortunately, I think both of these empirical symptoms of the causal/constitutive distinction – temporal lag and mediating mechanism – face outstanding challenges when applied to the case of consciousness and the NCC. These challenges make it unlikely that they can help us discriminate between CAUSATION and CONSTITUTION.

When it comes to temporal lag, there is of course the problem that no technology we can envisage at present has the sort of temporal resolution necessary to tell apart the difference between exact simultaneity and slight precedence at the time scales with which we are concerned here. (Certainly fMRI and EEG do not, but nor does optical imaging.[3]) More importantly, there are at least two *more principled* problems with appeal to temporal lag in the present context.

An initial problem is this. Imagine a time lag characteristic of the relevant kind of causal transaction – a lag between times t_1 and t_2. Imagine also that at t_1 the neural state N_1 occurs and the phenomenal state P_1 does, and that at t_2 neural state N_2 occurs and phenomenal state P_2 does. Here there are both materialist/constitutive and dualist/causal hypotheses regarding the neural correlate of P_2. The materialist hypothesis is that the neural correlate of P_2 is N_2, which is simultaneous with P_2 and indeed constitutes it. The dualist hypothesis is that the neural correlate of P_2 is N_1, which precedes it, because it is its cause. At the time scales we are talking about, P_2 is likely to be systematically correlated across different contexts with both N_1 and N_2.[4]

The same point applies to the very onset of consciousness. Suppose two mental states M_1 and M_2 occur at t_1 and t_2, such that M_2 is phenomenally conscious but M_1 is not. Suppose also that N_1 is a neural state exactly contemporaneous with M_1 and N_2 a neural state exactly contemporaneous with M_2. Again, we can hypothesize that N_1 is the neural correlate of M_2, hence a cause of consciousness, or that N_2 is the neural correlate of M_2, hence a constitutor of consciousness. Both hypotheses accommodate the timed observations of N_1, N_2, M_1, and M_2.

There is a further problem, which may be tougher yet. When trying to pinpoint the exact time of two kinds of event, with an eye to comparing these

times, it is crucial that we know how much time the measuring instruments take to produce their timing verdicts. Otherwise, there will be an irresolvable confound. If a time lag is detected between A and B, all we know immediately is that the detecting of A preceded the detecting of B. This is consistent with both (a) A really preceding B and (b) A and B being simultaneous but the detecting of B taking longer than the detecting of A. The only way to remove this confound is by having an independent measure of the time it takes each instrument to time its target. Ideally, this problem would be bypassed by using the very same measuring tool for both, or at least overcome by using measuring tools that demonstrably take the same amount of time to do the measuring. Clearly, however, in the present case this ideal set-up is unavailable: the timing of conscious states must ultimately rely on introspection, since introspection is our only direct access to conscious states, whereas the timing of neural states cannot use introspection, since introspection affords us no access to neural states.[5] Sub-ideally, then, we might use two different measuring instruments and find an independent way to measure the time it takes each measuring instrument to detect its target, subtracting this time to identify the likely time of occurrence of the target event. This approach may apply well to the timing of neural states: measuring the time it takes a measuring instrument to time the occurrence of a neural event may be fairly straightforward in principle (if technically challenging in practice). The real problem with the approach, however, is that when it comes to the timing of conscious states by introspection, the approach becomes circular, since introspective states are themselves conscious. We can imagine a future in which we have fully specified the neural mechanisms subserving introspection, and where we have measured precisely the time it takes for information to "travel up" to the "introspection center" and trigger the neural state underlying the introspective state. But unless we know whether there is a further bit of travel to be done, because that neural state merely *causes* the introspective state, or the travelling is finished, because the neural state *constitutes* the introspective state, we cannot be certain of the exact time it takes to introspectively detect that which is introspected.

If all this is correct, we are bound to remain stuck with our confound, and therefore with two empirically indistinguishable interpretations of any time lag between the detecting of the NCC and the detecting of consciousness.[6] Bearing in mind Wittgenstein's alleged remark that it is nonsense to suppose that humans will some day walk on the moon, and adopting in consequence a more diffident cast of mind toward the deliverances of armchair reasoning, I hesitate to rule out a priori the idea of a future time in which the timing of corresponding neural and conscious states has been established, in a way as yet elusive to our imagination, and is used to empirically distinguish CAUSATION and CONSTITUTION. Nonetheless, the above challenge to the very possibility of such a future looms large.

So much for using temporal lags to empirically distinguish causal and constitutive hypotheses. What about mediating mechanism? The idea was that causal transactions are mediated by mechanisms involving finer-grained causal transactions, whereas constitutive connections are not normally mediated by a series of finer-grained constitutions. Recall, however, that there was an exception to the rule that causal transactions are mediated by finer-grained transactions. The exception was causal transactions at the fundamental level of reality. At the bottom level of reality, there *are* no finer-grained causal transactions for us to seek. We must treat such transactions as metaphysically brute and ungrounded – somewhat as we treat the gravitational constant, the Avogadro constant, and other fundamental physical constants. Nothing *underlies* the fact that the gravitational constant is approximately 6.673×10^{-11} N·(m/kg)^2, and likewise nothing *underlies* the causal process by which a lepton absorbs a boson and converts into a neutrino. There are causal laws governing such causal transactions, but there are no finer-grained transactions mediating them. The problem this presents in the present context is that according to mainstream versions of naturalistic dualism, consciousness occurs precisely at the fundamental level of reality, where no mediating mechanism is to be found. If so, the fact that CONSTITUTION does not make room for a mediating mechanism linking the NCC and consciousness does not distinguish it from CAUSATION.

Consider Chalmers' (1996) version of naturalistic dualism. Chalmers reasons that since dualists, unlike materialists, hold that phenomenal properties are irreducible to any microphysical properties (or any other fundamental properties there might be), they must posit phenomenal properties as fundamental. If charm and spin are fundamental, then phenomenal consciousness must be construed as belonging at the same level of reality as charm and spin – the "bottom level." This means that any causal transactions between microphysical events and phenomenal events are effectively transactions at the "bottom level" of reality. That, in turn, means that there will be no *more* fundamental transactions mediating them – no mechanism connecting cause and effect. In consequence, the materialist's CONSTITUTION and the naturalistic dualist's CAUSATION make the exact same prediction here: that there will be no "intermediate correlates" between consciousness and the NCC, or more accurately between consciousness and the microphysical processes that constitute the NCC. Since they make the same prediction, they are empirically equivalent on this score.

There may be some other empirical symptoms of the difference between causation and constitution, other than temporal lag and mediating mechanism. But for my part, I cannot think of any. It would certainly be of great value to identify such potential empirical symptoms. Unless we can do so, we may have to acquiesce in the empirical indistinguishability of CONSTITUTION and CAUSATION, hence of non-eliminative materialism and naturalistic dualism.

4. Empirical Equivalence and the Science of Consciousness

As noted, in addition to *empirical* adequacy, scientific theories are also assessed for their *theoretical* adequacy. One could therefore suggest that CAUSATION and CONSTITUTION may yet be evaluated and compared with respect to the superempirical virtues. Certainly parsimony seems to tell in favor of CONSTITUTION, or materialism more generally, since 1 < 2 (see Smart 1959)… This approach raises a number of difficulties, however.

First, when two theories are perfectly empirically equivalent, there is an important sense in which choosing among them on the basis of superempirical virtues is a nonscientific endeavor. Doctoral students and postdocs in cognitive neuroscience laboratories are trained by their advisors in the designing and carrying out of experiments, not in the judicious comparison of experimentally indistinguishable hypotheses along superempirical dimensions.

More deeply, there is an ongoing debate in philosophy of science about the proper doxastic attitude toward the superempirical or theoretical virtues (see van Fraassen 1980, Churchland 1982). Consider simplicity. It is intuitive that *mutatis mutandis* we should always prefer simpler theories to more complicated ones. It is thus natural to count the simplicity of a theory as a reason for believing it. However, it is not obvious *why* the simplicity of a theory counts in its favor. In particular, it is very unclear why, indeed whether, a theory's simplicity means that it is more likely to be true. As van Fraassen (1980: 90) puts it, "it is surely absurd to think that the world is more likely to be simple than complicated."

Several philosophical ideas underlie this challenge to understanding the status of simplicity in theory evaluation. One idea is that ultimately, the only reason to *believe* a theory is that the theory is likely to be true. We may decide to *adopt* a theory, for pragmatic, aesthetic, or other reasons. But that is not quite the same as *believing* a theory. To believe a theory is to adopt it for *epistemic* reasons, more specifically for the reason that we think it *likely to be true*. Second, what makes a theory true is that it represents correctly the way the world is. Accordingly, for a theory to be more likely to come out true, it must be more likely that it represents the world the way it really is. Third, we do not actually have an independent handle on the objective degree of nature's complexity, in a way that would allow us to compare the complexity of nature and the complexity of theories that purport to describe it. If we take on board all three ideas, it would seem that simplicity is not a reason to *believe* a scientific theory – though it may well be a reason to adopt it on some non-epistemic grounds (i.e., for the sake of other purposes than knowing how the world is).

The worry is that this kind of reasoning may generalize to the other central theoretical virtues, especially parsimony and unity. We do not have some independent grip on the cosmos' degree of unity, one that suggests the cosmos is so inherently unified that the more unified our theory of it, the more likely the theory is to represent the world correctly. Likewise, we do not have an independent handle on the number of entities in the world that recommends keeping the number of posits in our theory thereof to a minimum. Thus whatever the force of unity and parsimony – and it is an open question both what that force exactly is and what is it based on – it cannot be due to their augmenting a theory's likely truth (i.e., the likelihood that it represents the world the way it really is). One way to put the challenge is that the theoretical virtues may not be *truth-conducive*: that a theory T exhibits the theoretical virtues does not make it more likely that T correctly represents the way things are (Beebe 2009, Kriegel 2013).

It might be objected that parsimony reasoning is often used in scientific theory building in what appears to be a truth-conducive way (Sober 2009). Consider this piece of reasoning from evolutionary biology: both humans and monkeys have tailbones; if humans and monkeys have no common ancestry, the tailbone would have had to originate twice; if they have common ancestry, it only had to originate once; the latter hypothesis is thus more parsimonious than the former, and more likely to be true. The idea here is that the occurrence of two independent events is less probable than the occurrence of one (other things being equal). Suppose the probability of E_1 occurring is 70% and that of E_2 is 50%. Then the probability of *both* occurring is 35% – lower than either. Accordingly, the single-event hypothesis is more probable than the dual-event one. So parsimony tracks likely truth.

Observe, however, that the kind of parsimony invoked here is not the kind invoked by materialism and CONSTITUTION. Both evolutionary hypotheses under consideration posit the same types of entity – humans, monkeys, tailbones, originations. They only differ on the *distribution* of those entity types: one posits one token event where the other posits two tokens. Thus the two hypotheses differ in what we might call *token*-parsimony, but are equal in *type*-parsimony: they differ in the number of token tailbone-origination events they

posit, but both are ontologically committed to tailbones, originations, and indeed tailbone-originations. In contrast, materialism and dualism differ in *type*-parsimony: they disagree on the *kinds* of things there are in the world.[7] The point is: while it is clear how token-parsimony can be truth-conducive, it is much more mysterious how type-parsimony might be. Yet it is the latter that separates CAUSATION and CONSTITUTION.

If all this is right, then there may be no epistemic grounds for preferring CAUSATION or CONSTITUTION, qua scientific hypotheses about the relationship between consciousness and its neural correlate. On the one hand, there appear to be no way to experimentally disentangle them. On the other, the theoretical virtues do not seem to apply to them in a way that renders one more likely to be true than the other. As already noted, it may well be that our present difficulties in envisaging an experimental test that could separate predictions by CAUSATION and CONSTITUTION are but failures of imagination. It is also possible, of course, that a demonstration of the truth-conduciveness of (some) superempirical virtues will emerge at some point. Still, the considerations above do cast a worrisome shadow over the hope for a scientific resolution of the dualism/materialism debate. If so, the right attitude may be to withhold scientific judgment on whether CAUSATION or CONSTITUTION (or one of the other four hypotheses formulated in §1) is most likely to be true.

This line of reasoning is in some ways disappointing. But in other ways, it may be thought liberating. The problem of consciousness has led many scientists to ignore consciousness as an improper subject of scientific investigation. Some have even been led to deny the existence of consciousness, more or less to protect the Enlightenment notion that science can account for every aspect of reality. Others have admitted the existence of consciousness and refused to ignore it, but there is a stubborn sense that they nonetheless have deflated somewhat the phenomenon, turning it into a purely functional phenomenon thin on intrinsic subjective character. The above reflections recommend a humbler approach that relinquishes the mentioned Enlightenment ideal and concedes that we may be unable in principle to reach a scientific resolution of the problem of consciousness. There may be principled methodological and epistemological reasons why we cannot choose among the

various possible explanations of the correlation between consciousness and the NCC. Indeed, the above reflections may be seen to offer a *diagnosis* of the elusiveness of scientific progress on the ultimate question of consciousness' place in nature.[8]

References

- Beebe, J.R. (2009), 'The Abductivist Reply to Skepticism', in *Philosophy and Phenomenological Research* 79: 605-636.
- Berkeley, G. (1710), *A Treatise Concerning the Principles of Human Knowledge*.
- Chalmers, D.J. (1996), *The Conscious Mind*. (Oxford UP).
- Chalmers, D.J. (2002), 'Consciousness and Its Place in Nature', in D.J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (Oxford UP).
- Churchland, P.M. (1979), *Scientific Realism and the Plasticity of Mind* (Cambridge UP).
- Churchland, P.M. (1982), 'The Ontological Status of Observables: In Praise of the Superempirical Virtues', in *Pacific Philosophical Quarterly* 63: 226-236.
- van Fraassen, B.C. (1980), *The Scientific Image*. (Oxford UP).
- Gratton, G., Fabiani M., Elbert, T., and Rockstroh, B. (2003), 'Seeing Right through You: Applications of Optical Imaging to the Study of the Human Brain', in *Psychophysiology* 40: 487-491.
- Hameroff, S.R., and Penrose, R. (1996), 'Conscious Events as Orchestrated Spacetime Selections', in *Journal of Consciousness Studies* 3: 36–53.
- Harman, G. (1965), 'The Inference to the Best Explanation', in *Philosophical Review* 74: 88-95.
- Hohwy, J. (2011), 'Mind-Brain Identity and Evidential Insulation', in *Philosophical Studies* 153: 377-395.

- Huemer, M. and Kovitz, B. (2003), 'Causation as Simultaneous and Continuous', in *Philosophical Quarterly* 53: 556-565.
- Kriegel, U. (2009), *Subjective Consciousness: A Self-Representational Theory* (Oxford UP).
- Kriegel, U. (2013), 'The epistemological challenge of revisionary metaphysics', in *Philosophers' Imprint* 12 (June): 1-30.
- Lau, H.C. and Passingham, R.E. (2006), 'Relative Blindsight in Normal Observers and the Neural Correlate of Visual Consciousness' in *Proceedings of the National Academy of Science USA* 103: 18763-18768.
- Lipton, P. (1991), *Inference to the Best Explanation* (Routledge).
- Lockwood, M. (1989), *Mind, Brain and the Quantum* (Blackwell).
- Quine, W.V.O., and Ullian, J.S. (1970), *The Web of Belief* (Random House).
- Rosenberg, G. (2005), *A Place for Consciousness: Probing the Deep Structure of the Natural World* (Oxford UP).
- Pelczar, M.W. (2015), *Sensorama: A Phenomenalist Analysis of Spacetime and Its Contents* (Oxford UP).
- Rounis E., Maniscalco, B., Rothwell, J., Passingham, R.E., and Lau, H.C. (2010), 'Theta-Burst Transcranial Magnetic Stimulation to the Prefrontal Cortex Impairs Metacognitive Visual Awareness', in *Cognitive Neuroscience* 1: 165-175.
- Russell, B. (1927), *The Analysis of Matter* (Kegan Paul).
- Smart, J.J.C. (1959), 'Sensations and Brain Processes', *Philosophical Review* 68: 141-156.
- Sober, E. (2009), 'Parsimony Arguments in Science and Philosophy – a Test Case for Naturalism$_P$', in *Proceedings and Addresses of the American Philosophical Association* 82: 117-155.

[1] At this point, I do not wish to comment on the plausibility of this view. The present discussion is intended merely to lay out the *possible* explanations.

[2] Note, though, that our sixfold distinction is not exhaustive of the logical landscape on the mind-body problem. It notably leaves out eliminative materialism and eliminative idealism. It also fails to make certain distinctions, e.g. between reductive and non-reductive materialism and between epiphenomenalist and downward-causation versions of dualism.

[3] On optical imaging and its resolution, see Graton et al. 2003.

[4] The problem would be solved if we could somehow "observe" not only both correlates, but also the actual connection between them. But it is widely held that the connection between cause and effect is unobservable (and presumably so is the connection between constitutor and constitute).

[5] This too has been contested by some philosophers. For example, Churchland (1979) claims that with better (more scientifically based) primary and secondary education, future generations of humans will learn to introspect their conscious life in neural terms. I am going to assume here that this is false.

[6] The confound, to repeat, is this. Suppose we detect consciousness and the NCC, and the detecting of the latter suitably precedes the detecting of the former. One interpretation of this time lag between the two detections is that there was a real time lag between the NCC and consciousness. The other is that there was no time lag between consciousness and the NCC and the lag between their detections is due entirely to the different speeds of operation of our timing devices. The specter I am raising here is that there is no way to experimentally pull apart these two interpretations. (Conversely, suppose we find no time lag between the detecting of the NCC and the detecting of consciousness. This too is consistent with at least two interpretations. One is that the two are simultaneous. The other is that the NCC precedes consciousness but its precedence is masked by a compensatory difference in the speed of timing consciousness and timing the NCC.)

[7] Furthermore, the dualist does not posit her two types of property – physical and phenomenal – to explain a single explanandum (as is the case with the tailbones). Rather, she posits the physical brain property to explain neurological data or third-person overt behavior, but phenomenal properties to explain our introspective impressions or first-person grasp on our mental life.

[8] I would like to thank David Chalmers, Jakob Hohwy, Benji Kozuch, and Farid Masrour for comments on a previous draft, and Benji Kozuch and Rachel Schneebaum for useful

conversations. I have also benefited from presenting one incarnation or another of the paper at the Berlin School of Mind and Brain, Boston University, CREA, the University of Arizona, the University of Copenhagen, and the University of Strasbourg. I am grateful to the audiences there, in particular Michel Bitbol, Rosa Cao, Carolyn Dicey-Jennings, Ellen Fridland, Rik Hine, Shaun Nichols, Michael Pauen, and Sebastian Watzl. This work was supported by the French National Research Agency's grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC, as well as by grant 675415 of the European Union's Horizon 2020 Research and Innovation program.