Beyond the Neural Correlates of Consciousness

Uriah Kriegel

Forthcoming in U. Kriegel (ed.), *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 correlation relations, but also – and more deeply – in causal and constitutive relations. When faced with a correlation between two phenomena in nature, we typically feel compelled to produce an *explanation* of why the two 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 would ultimately have to choose. The second half then raises considerations suggesting that, under certain reasonable assumptions, the choice among these various options may be *in principle* underdetermined by the relevant scientific evidence, in the sense that the traditional metaphysical positions may be strictly *empirically equivalent*. If so, the choice between them cannot in principle be a scientific one – it must be a matter of *philosophical*-theory choice.

Introduction

The centerpiece of the scientific study of consciousness is the search for the neural correlates of consciousness (Lau, this volume). Yet science is typically interested not only in correlation relations, but also – and more deeply – in causal and constitutive relations. When faced with a correlation between two phenomena in nature, we typically feel compelled to produce an *explanation* of why the two correlate. To posit brute and inexplicable correlations 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 them, by providing explanations of them.

It is worth noting, in this context, Leibniz's "pre-established harmony theory" of the connection between the mind and the body. Leibniz may have been the first philosopher to articulate the problem of the explanatory gap explicitly:

> It must be confessed, however, that *perception*, and that which depends upon it, *are inexplicable by mechanical causes*, that is to say, by figures and motions. Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior, as he would into a mill. Now, on going into it he would find only pieces working upon one another, but never would he find anything to explain perception. (Leibniz 1714, §17; italics original)

At the same time, Leibniz recognizes that events in the domain of consciousness appear to correlate with events in the physical domain. His hypothesis about the ultimate relation between consciousness and its neural correlates is highly speculative and mysterian: 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 simltaneously 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 scientific work on consciousness does not even attempt to do that, it would appear to be even more mysterian and anti-scientific than Leibniz's outlandish hypothesis.

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. Ideally, the explanation in question would bear on the perennial philosophical problem of the ultimate connection between mind and body. In this chapter, however, I will voice skepticism about the satisfiability of this hope.

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 ultimately 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, in the sense 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 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 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 (see 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.

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. This is often the best explanation of such 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 – 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), for example, is as 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*, not 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, H_2O *constitutes* 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).

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 John Foster (1982), 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 also characterizes 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 itself. Here the reasoning is this:

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 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.

Above, these are occasionally illustrated using synchronization with dlPFC as a potential NCC, but one could plug in any other neural structure, depending on one's favored NCC hypothesis.

2. Explanatory Hypotheses and Metaphysical Positions

Each of the six explanatory hypotheses just laid out corresponds to a *metaphysical position* on the ultimate connection between consciousness and the relevant part of the material brain, and more generally between the phenomenal and physical aspects of reality. Modern philosophical discussions of such metaphysical positions tend to characterize them in terms of *supervenience*. In particular, the distinction between materialist and dualist positions is often taken to come down to the choice between *metaphysical* and merely *nomological* supervenience (Chalmers 1996). The latter grounds consciousness in brain activity via laws of nature – presumably *causal* laws. The former guarantees a connection that goes beyond causal laws – it posits a *constitutive* connection between consciousness and the brain that must hold regardless of the causal laws of nature.

It should be noted that modern philosophical discussions focus on nomological supervenience because it captures a particularly respectable variety of dualism, a kind of dualism that does not treat consciousness as entirely brute and inexplicable, but as one explained causally rather than constitutively by brain

activity. This is what Chalmers (1996) calls "naturalistic dualism." But there are, of course, more non-naturalistic versions of dualism that deny even this sort of causal grounding of the phenomenal in the physical. According to some views, the physical world is not causally closed: not every physical event or fact has a physical cause. This is because some phenomenal properties can emerge in nature which are endowed with downward-causal powers. Such powers are not fixed by the causal powers of the underlying physical properties, and they allow for (without quite ensuring) REVERSE CAUSATION.[2] Since the phenomenal properties' causal powers outstrip those of their physical correlates, there can be two possible worlds in which all the same physical properties are instantiated at a time *t*, but the phenomenal properties instantiated at *t* have slightly different causal effects, leading to a different state of the universe at the next moment. Under reasonable assumptions, this kind of scenario constitutes a violation of the nomological supervenience of the phenomenal on the physical. But in any case, it seems "non-naturalistic" in an important sense.[3] (To say that the sense is important is not to say that it is *pejorative*; the view has been ably defended.)

It would seem, then, that CAUSATION, CONSTITUTION, and REVERSE CAUSATION can be framed in terms of theses about supervenience relations between the phenomenal and the physical. According to CONSTITUTION, phenomenal properties supervene metaphysically on physical properties; according to CAUSATION, the former supervene merely nomologically on the latter; according to REVERSE CAUSATION, phenomenal properties do not even nomologically supervene upon physical properties.

Clearly, REVERSE CAUSATION and REVERSE CONSTITUTION can likewise be framed in terms of supervenience – though this time the supervenience of the relevant sub-class of physical properties (those implicated in the NCC) upon phenomenal properties. The view corresponds, after all, to 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). This is importantly different from the view that the physical exists but is ultimately constituted by the phenomenal.

The latter view is idealist as well, however. We may distinguish them by calling the former *eliminative idealism* and the latter *reductive idealism*, somewhat as we commonly distinguish between eliminative and reductive materialism (see Sprevak, this volume). Just as the reductive materialist attempts to reduce consciousness to matter rather than heavy-handedly deny the very existence of consciousness, so the reductive idealist attempts to reduce matter to consciousness rather than implausibly deny matter's existence. Likewise, just as the eliminative materialist thinks that the kind of consciousness we retain in a purely physicalist worldview is forsooth consciousness in name only (Churchland 1981), so the eliminative materialist thinks that we would need to twist our own words to retain a place for matter in our ultimate theory of the world. Interestingly, although the historically best known version of idealism – Berkeley's – is straightforwardly eliminativist, other historically prominent versions of idealism – such as Plato's and Leibniz's – are arguably reductivist.[4] The point, in any case, is that reductive idealism, in paralleling reductive materialism, is committed to the metaphysical supervenience of physical properties upon phenomenal properties. (Eliminative idealism, meanwhile, is committed to the nonexistence, or at least non-instantiation, of physical properties.)

The other two explanatory hypotheses discussed above were THIRD CAUSE and THIRD CONSTITUTOR. The metaphysical position committed to the latter is in effect the traditional view known as *neutral monism*, defended by Spinoza and Russell *inter alia*. The view has enjoyed a resurgence in more recent philosophy of mind, often under the name *Russellian monism* (see Lockwood 1989). In Spinoza, the idea is that there is some third substance, more primordial than either mind or matter, which manifests itself as matter in one context and as mind in another. The more modern version of the view is framed in terms of *properties* rather than *substances*. The idea is that there must exist certain fundamental properties of the universe that are neither phenomenal nor physical but which are both *proto*-phenomenal and *proto*-physical: different combinations or aggregates of them somehow ground phenomenal properties and physical properties. The motivation for this is often the

observation that physics only tells us about physical properties' (actual and potential) causal relations to each other, but is silent on their (categorical) intrinsic natures (Russell 1927). The observations paves the way to the speculation that the intrinsic, categorical *je-ne-sais-quoi* of the properties invoked in physics is *one and the same* as that of mental properties. Accordingly, grounding both physical and mental properties are more basic properties which in themselves are neither mental nor physical. If we want to put the view in terms of supervenience, we might say that both physical and phenomenal properties metaphysically supervene upon those intrinsic, categorical, 'proto,' *je-ne-sais-quoi* properties.

Insofar as THIRD CONSTITUTOR is a kind of neutral or Russellian monism, we should think of THIRD CAUSE as a kind of "neutral dualism" or "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. The upshot is that both physical and phenomenal properties merely *nomologically* supervene upon the underlying 'proto' properties.

Thus we can map the six explanatory hypotheses laid out at the end of the previous section onto six traditional metaphysical positions concerning the supervenience relations holding between phenomenal and physical properties:

1) NATURALISTIC DUALISM: phenomenal properties *merely nomologically* supervene on physical properties. (This corresponds to CAUSATION.)
2) NON-NATURALISTIC DUALISM: phenomenal properties fail to supervene even nomologically on physical properties. (This corresponds roughly to REVERSE CAUSATION.)
3) NEUTRAL DUALISM: phenomenal properties and physical properties merely nomologically supervene on *proto-phenomenal* properties. (This corresponds to THIRD CAUSATION.)
4) MATERIALISM (MATERIALISTIC MONISM): phenomenal properties *metaphysically* supervene on physical properties. (This corresponds to CONSTITUTION.)

5) IDEALISM (IDEALIST MONISM): physical properties metaphysically supervene on *phenomenal* properties. (This corresponds to REVERSE CONSTITUTION.)
6) NEUTRAL MONISM: phenomenal properties and physical properties metaphysically supervene on proto-phenomenal properties. (This corresponds to THIRD CONSTITUTOR.)

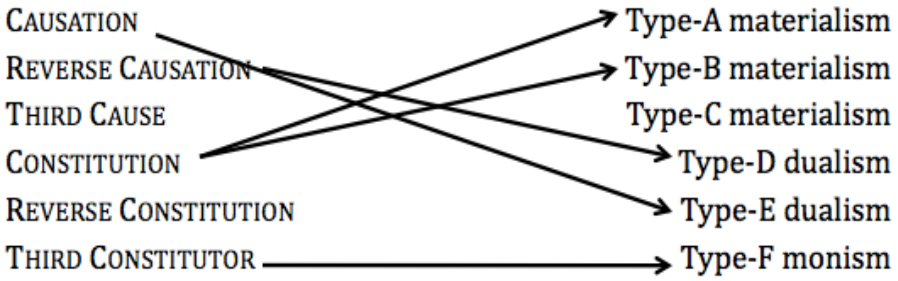
What this suggests is that the main views on the mind-body problem, or rather the part of it we may call the 'consciousness-brain problem,' trace back to different hypotheses about what best explain the evident, scientifically established correlation between phenomenal and physical properties.

It is interesting to compare the above sixfold scheme for categorizing views on the problem of consciousness with David Chalmers' (2002) oft-cited sixfold scheme. Chalmers divides the landscapes into six positions, which he calls type-A materialism, type-B materialism, type-C materialism, type-D dualism, type-E dualism, and type-F monism. However, as Chalmers notes, type-C materialism is a somewhat unstable position, and in all likelihood collapses to one of the other positions once sufficiently developed.[5] Type-F monism, meanwhile, is just Russellian monism. As for Chalmers' two kinds of dualism, type-D gives consciousness 'downward' causal powers not inherited from its neural correlate, as in REVERSE CAUSATION and NON-NATURALISTIC DUALISM; type-E deprives consciousness of such causal powers – indeed, in order to respect physical causal closure, of any causal powers – and so remains within the bounds of NATURALISTIC DUALISM and CAUSATION.

The one distinction in Chalmers' scheme not reflected in ours is between type-A and type-B materialism. There are two ways to characterize this distinction, one epistemic and one metaphysical. The epistemic characterization is in terms of the explanatory gap (Levine 1983): type-A materialism denies that it exists, or that it is unbridgeable (see Jackson, this volume); type-B materialism concedes its existence but denies that it forces us to reject physicalism (see Levine, this volume). Another characterization of the type-A/type-B the difference, however, is metaphysical and appeals again to supervenience relations: type-A materialism

implies that the supervenience of phenomenal upon physical properties is even stronger than merely metaphysical supervenience, and amounts to *conceptual*, *logical*, or otherwise *a priori* supervenience; type-B materialism insists on *merely* metaphysical, and hence *a posteriori*, supervenience. According to type-A materialism, then, it is a conceptually necessary truth that consciousness supervenes on its neural correlate. According to type-B, it is only a Kripkean metaphysically necessary truth (see Kripke 1972).

The mapping of these views onto the ones proposed in the above scheme is not straightforward. The above remarks suggest that type-A and type-B materialism are both versions of MATERIALISM, and therefore are both committed to CONSTITUTION. I have also suggested that type-D dualism maps onto NATURALISTIC DUALISM, to which REVERSE CAUSATION leads, while type-E dualism is a version of NATURALISTIC DUALISM, led to by CAUSATION. It is also clear that type-F monism is essentially NEUTRAL MONISM (committed to THIRD CONSTITUTOR). In any case, type-C materialism fits none of the items in the present paper's sixfold scheme, while IDEALISM and NEUTRAL DUALISM (with their respective commitments to REVERSE CONSTITUTION and THIRD CAUSE, respectively) are absent in Chalmers' scheme. These mapping relationships may be summarily represented as follows:



The lacunae in Chalmers' scheme are not really oversights. The scheme attempts to classify the main views in play in live debates in contemporary philosophy of mind – it is not intended as a catalog of logically coherent positions. However, there is also value in widening our view enough to appreciate all the coherent, epistemically

possible positions on the matter. And it is worth keeping in mind that certain views currently widely considered hopeless may at some point enjoy a sociological renaissance; indeed, this appears to have happened with Russellian monism over the past two decades. This is what the present sixfold scheme attempts to do. An advantage of our alternative scheme is that it is ultimately grounded in six possible interpretation of scientifically established correlations.

There are also advantages to Chalmers' scheme, of course. The most important is its distinction between two importantly different kinds of materialism. However, this advantage can be co-opted, by simply distinguishing two "grades" of MATERIALISM, conceptual and merely metaphysical (*a priori* and *a posteriori*):

4a) TYPE-A MATERIALISM: phenomenal properties *conceptually* supervene on physical properties.

4b) TYPE-B MATERIALISM: phenomenal properties *merely* metaphysically supervene on physical properties.

Corresponding to the difference between type-A and type-B materialism are two different explanatory hypotheses about why consciousness correlates with the right neural event – two different versions of CONSTITUTION. One version will claim that consciousness is constituted by its NCC as a matter of a priori conceptual links between the notions of consciousness and its neural correlate (this is type-A materialism's commitment). The other will hold that it is a matter of Kripkean a posteriori necessity (this is type-B's).

For complete comprehensiveness, we might make a parallel distinction between two different versions of idealism, as well as add to the menu of options eliminative materialism and (what I have called) eliminative idealism. The upshot would be a menu of ten options.

The most important feature of the present scheme, however, is not its comprehensiveness. It is that it 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 phenomenal consciousness and its NCC. The parallelism between the sixfold schemes of the present and previous sections suggests that we may be able reframe the philosophical problem of the ultimate metaphysical relation between consciousness and the brain, and more generally the phenomenal and the physical, as the following question: Which is the most reasonable explanatory inference to make from the correlation between consciousness and the NCC? That is: Which is 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 the explanatory hypotheses we have encountered are all 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 among them.

I will conduct the discussion by focusing on the distinction between CONSTITUTION and CAUSATION, hence between materialism and naturalistic dualism. I do so because of the special status of these two hypotheses as the leading competitors in the extant literature (and because discussing *all* explanatory hypotheses mentioned above would be unmanageable!). But the points I will make should extend to the other explanatory hypotheses as well.

In order for there to be an empirical difference between CAUSATION and CONSTITUTION, there would have to be some (*possible*) empirical data that one accommodates (*could* accommodate) better than the other. In cognitive-scientific practice, two types of data are typically recognized: behavioral and neural. More controversially, some hold that in the area of consciousness research a third type of datum must be admitted – introspective data. The problem, however, is that it is unclear what behavioral, neural, or introspective differences there might be between causal and constitutive explanations of phenomenal-neural correlations. It is hard to imagine a neurological study producing such data, given that neurological studies target concern directly only the neural relatum of the correlation. At the same time, introspective evidence concerns directly only the phenomenal relatum of that correlation, so it is hard to see how it could fare better. Meanwhile, behavioral studies target directly neither relatum, trying instead to shed indirect light on both; it is in any case hard to imagine a behavioral study the results of which might suggest that the phenomenal is caused by the neural rather than constituted by it (or constituted rather than caused). Thus it is very unclear how *any* of these types of datum, individually or in combination, might discriminate between CAUSATION and CONSTITUTION.

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 symptoms 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 constotitors 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" at *any* level of reality. For A to constitute B, it is not necessary that there be some series $X_1 \ldots X_n$ 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, 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. 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.[6]) 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 thus can be taken to constitute it. The dualist hypothesis is that the neural correlate of P_2 is N_1, which precedes it in such a way that it can be taken to be its cause. At the time scales we are talking about, P_2 is likely to be systematically correlated across different context with both N_1 and N_2.[7] Unless we can somehow "observe" not only both correlates, but also the actual connection between them, both hypotheses accommodate the observational data equally. Yet accordingly to a widely held and highly plausible Humean view, the actual connection between cause and effect is unobservable – and equal plausibility attaches to the unobservability of the actual connection between constitutor and constitutee. If so, the temporal symptom of the difference between causation and constitution may not be exploitable here.

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. Since arguably the connections among them (causal or constitutive) cannot be themselves observed, it is not clear what observational data could separate the two hypotheses.

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 phenomenal states must ultimately rely on introspection, since introspection is our only direct access to phenomenal states, whereas the timing of neural states cannot use introspection, since introspection affords us no access to neural states.[8] 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 problem real problem with the approach, however, is that when it comes to the timing of phenomenal states by introspection, the approach is circular. 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.[9] Bearing in mind Wittgenstein's remark that it is nonsensical to suppose that people sometimes go to 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 phenomenal has been established, in a way as yet elusive to our imagination, in such a way as 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 differences 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 some lepton absorbs a boson and converts into a neutrino. There are 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 for that matter any other fundamental properties there might be), they must posit phenomenal properties as fundamental alongside such microphysical properties. For example, assuming that such quantum-mechanical properties as charm and spin are fundamental, 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 at least 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. For my part, I suspect that ultimately the

fundamental problem in distinguishing a causal and a constitutive reading of the correlation between consciousness and the NCC concerns the different access we have to each: we have only introspective access to consciousness, and only non-introspective access to its neural correlate. The original introspectionists were keenly aware of this, allowing introspection to play a role only in *describing* consciousness but not in explaining it (see Titchener 1912). The chasm between our modes of access to the two relata of the correlation relation frustrates a clear conception of the connection between them.

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. It is not qua scientist that one compares the relative simplicity, parsimony, conservatism, and so on of two experimentally indistinguishable theories. After all, doctoral students and postdoctoral researchers 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) points out, "it is surely absurd to think that the world is more likely to be simple than complicated."

Several philosophical ideas underlie this challenge to the 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 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, so 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 other grounds and for the sake of other purposes than knowing how the world is.

The worry is that this is kind of reasoning may generalize to the other central theoretical virtues, especially parsimony and unity. 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).

Consider the virtue variously referred to as unity, coherence, or cohesion. Intuitively, the more interconnected and unified a theory's tenets, the more compelling the theory. However, Bovens and Hartmann (2003), working within Bayesian probability theory, have offered a formal proof that the internal coherence of a theory is not truth-conducive: a more coherent theory is not more likely to be

true than a less coherent one. The debate on Bovens and Hartmann's proof is ongoing, and far from settled, but nobody as managed as yet to establish the truth-conduciveness of unity or coherence.

The case of parsimony is trickier. In some scientific contexts, it is clear that parsimony is truth-conducive. 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 to originate twice; if they have common ancestry, it only has to originate once; the latter hypothesis is thus more parsimonious than the former, and more likely to be true (Sober 2009). 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 posit the same entities – humans, monkeys, tailbones, originations. They only differ on the *distribution* of those entities. Or perhaps more accurately: the two hypotheses differ in what we might call *token*-parsimony, but are equal in *type*-parsimony. One posits only one tailbone-origination event where the other posits two, but both are ontologically committed to tailbones and originations. In contrast, materialism and dualism differ in *type*-parsimony: they disagree on *kinds* of things there are in the world. 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.

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. At the same time, 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, science has in the past proven able to outstrip the imaginative capacities of its detractors time and again, and 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. The two may simply be both empirically and superempirically equivalent. As far as science is concerned, then, it would seem that we should simply withhold judgment on whether CAUSATION or CONSTITUTION (or one of the other four hypotheses formulated in §1) is most likely to be true.

Conclusion

The above 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 offer a *diagnosis* of the elusiveness of scientific progress on the ultimate question of consciousness' place in nature.

At the same time, the implications for a *philosophical* inquiry into consciousness are less clear. For all we have said, armchair reasoning and disciplined metaphysical speculation may still be appropriate where experimental testing and superempirical evaluation are not. In the background are vexed questions about philosophical methodology, the relationship between science and philosophy, and the aims of intellectual inquiry.[10]

References

- Beebe, J.R. (2009). The Abductivist reply to skepticism. *Philosophy and Phenomenological Research* 79:605-636.
- Berkeley, G. 1710. *A Treatise Concerning the Principles of Human Knowledge*.
- Boutel, A. (2013). How to be a type-C physicalist. *Philosophical Studies* 164: 301-320.
- Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford UP.
- Chalmers, D.J. (1996). *The Conscious Mind*. Oxford and New York: 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 and New York: Oxford UP.
- Churchland, P.M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.
- Churchland, P.M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67-90.
- Churchland, P.M. (1982). The ontological status of observables: In praise of the superempirical virtues. *Pacific Philosophical Quarterly* 63:226-236.
- Foster, J. (1982). *The Case for Idealism*. London: Routledge & Kegan Paul
- van Fraassen, B.C. (1980). *The Scientific Image*. Oxford and New York: 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. *Psychophysiology* 40:487-491.
- Hameroff, S.R., and Penrose, R. (1996). Conscious events as orchestrated spacetime selections. *Journal of Consciousness Studies* 3:36–53.
- Harman, G. (1965). The inference to the best explanation. *Philosophical Review* 74:88-95.
- Hohwy, J. (2011). Mind-brain identity and evidential insulation. *Philosophical Studies* 153: 377-395.
- Huemer, M. and Kovitz, B. (2003). Causation as simultaneous and continuous. *Philosophical Quarterly* 53:556-565.
- Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford UP.
- Kriegel, U. (2013). The epistemological challenge of revisionary metaphysics. *Philosophers' Imprint* 12 (June): 1-30.
- Kripke, S. (1972). Naming and necessity. In D. Davidson and G. Harman (eds.), *Semantics of Natural Language*. Dordrecht: Reidel.
- Lau, H.C. and Passingham, R.E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Science USA* 103:18763-18768.
- Leibniz, G.W. 1714. *The Monadology*.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354-361.
- Lipton, P. (1991). *Inference to the Best Explanation*. London: Routledge.
- Lockwood, M. (1989). *Mind, Brain and the Quantum*. Oxford: Blackwell.
- Quine, W.V.O., and Ullian, J.S. (1970). *The Web of Belief*. New York: Random House.
- Rosenberg, G. (2005). *A Place for Consciousness: Probing the Deep Structure of the Natural World*. Oxford: 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. *Cognitive Neuroscience* 1:165-175.
- Russell, B. 1927. *The Analysis of Matter*. London: 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$_P$.' *Proceedings and Addresses of the American Philosophical Association* 82: 117-155.
- Titchener, E.B. (1912). The schema of introspection. *American Journal of Psychology* 23:485-508.

---

[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] The view allows, but does not guarantee, REVERSE CAUSATION, because while it allows that phenomenal properties downward-cause some neural events, it does not ensure that the NCC is one of those downward-caused events. That is, it is not committed to the idea that consciousness downward-causes its own neural correlates, as REVERSE CAUSATION requires. At the same time, it is *compatible* with such downward causation of the NCC.

[3] The assumption I have in mind is that properties individuate sensitively to their causal powers, so that any properties F and G with different causal powers are different properties. If the phenomenal properties in the relevant pair of worlds have different causal powers, on this principle about individuation they are different properties. And yet the physical properties involved are *ex hypothesi* the same. Yet the laws of nature are the same. So there is failure of nomological supervenience.

[4] In many places, Leibniz (1714) puts his view by saying that matter is a 'well-founded phenomenon' that can be reductively explained in terms of immaterial 'monads.' Plato, meanwhile, appears to grant only Ideas the highest ontological status, that of being 'fully real' (whatever that means), but not to altogether deny an existential status of the spatiotemporal world's goings-on.

[5] This has been challenged, but I will assume here is ultimately correct. For discussion, see Boutel 2013.

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

[7] Thus, suppose dualism is true and N_1 is actually the cause of P_2. Then not only N_1 is likely to correlate systematically in a variety of different contexts with P_2, but also N_2.

[8] 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.

[9] 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.)

[10] 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, and the University of Copenhagen. 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.