THE CAUSAL ROLE OF STATE CONSCIOUSNESS

by

GEORGE A. SELI

A dissertation submitted to the Graduate Faculty in philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2013

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Barbara Montero

_____ _____

Date Chair of the Examining Committee

Iakovos Vasiliou

_____ _____

Date Executive Officer

Barbara Montero

Michael Levin

David Rosenthal

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE CAUSAL ROLE OF STATE CONSCIOUSNESS

by

George A. Seli

Adviser: Professor Barbara Montero

Mental states often occur consciously. We regularly have conscious perceptions in different modalities, for example. A thought process usually involves several conscious beliefs, perhaps conscious doubts or desires. It is generally assumed that a mental state affects cognition and behavior in virtue of its psychological properties. But is a state's *being conscious* – what I call the c-property – causally relevant? If so, does the efficacy of that property benefit the creature that is in the conscious state? I argue for an affirmative reply, based on a higher-order theory of consciousness. Such a theory claims that for a mental state to be conscious is for the agent to be aware of being in it, via a suitable mental representation of the first-order state.

In Chapter 1, I develop an account of causally relevant properties of events, since I construe consciousness as a property of a mental event. In Chapter 2, I review traditional problems for mental causation, such as the contention that neural states are causally sufficient for all cognition and behavior. I show how proposed solutions to these problems would also establish that it is possible for the c-property to be efficacious. Even so, the property may happen to be epiphenomenal. Accordingly, in Chapter 3 I give the conditions that an event property must meet in order to count as epiphenomenal. In particular, I argue that such a property's

epiphenomenality is consistent with the *necessity* of its instantiation to the outcome of a causal process.

I devote Chapter 4 to the function of consciousness with regard to the feeling of agency. I argue that the higher-order representation that makes such a feeling conscious can be deployed in reasoning about the feeling at the time it occurs. I extend this account in Chapter 5 to conscious perceptions, volitions, and thoughts, arguing that the utility of these first-order states' *being represented* is to enable reasoning about whether to be in the states. I conclude by showing how my proposal explains the correlation between consciousness and (i) perceptions and volitions that guide non-routine behavior, and (ii) thinking that intellectually challenges the agent.

TABLE OF CONTENTS

I. INTRODUCTION

1. Folk Psychological Views

We are in a variety of mental states throughout our daily lives: a state of elation over a promotion, a state of hearing an alarm, a state of doubt about the competency of a mechanic, a state of expecting a taxi to arrive. It is commonly assumed that these states make things happen, both other mental states and physical behavior. For example, a worry gives rise to deliberation as to how to improve a situation, seeing an acrobatic feat engenders applause, a belief that proposition *P* causes an assertion that *P* in certain situations, and so forth. We also think that some mental states occur consciously, while others do not. One might have a standing belief that the local museum is to the north of one's home, but only consciously believe that when giving directions to a guest. One might be jealous of one's brother, but only become consciously jealous after several sessions with a psychologist. If there are both conscious and nonconscious mental states, the view that mental states are efficacious does not address the question whether a state's *being conscious* – what we may call "state consciousness" – is efficacious. Does that property causally impact cognition and behavior along with the states that instantiate it, or is it epiphenomenal? That is, do conscious mental states cause things in only virtue of *other* psychological properties they instantiate, not their being conscious? And, more interestingly, are the effects (if any) of state consciousness useful? In this project, I aim to develop a theory that answers the latter question in the affirmative, based on one particular theory of consciousness.

My fundamental assumptions, then, are shared by the folk: there is such a phenomenon as consciousness, there are mental states, and consciousness, in one sense of the term, is a property

of some (but not all) mental states. Yet realism about state consciousness leaves a very open question as to its function. The claim that the value of consciousness is obvious – since we're unable to do much when we're unconscious – is founded in a distinct notion of consciousness: *creature consciousness*. Per David Rosenthal's definition (1997, p. 730), for a creature to be conscious is simply for it to be awake and responsive to stimuli, which of course is critical to its survival. But is having conscious mental *states* critical, or even helpful, to its survival? After all, there is evidence that nonconscious states, such as subconscious fears and subliminal perceptions, do impact our cognition and guide behavior. Moreover, not all theories of state consciousness make it plain that it is beneficial for a creature to be in conscious mental states. For example, on higher-order theories, as I will discuss, it seems that thinking, perceiving, and willing could proceed just as well nonconsciously. If so, then consciousness is a causally inert property of mental states. Our cognitive, conative, and perceptual states, when conscious, would not *depend* on being conscious in order to generate or optimize behavior. And thus, theoretically, we could be creature conscious – responsive to stimuli in all the usual ways – while lacking state consciousness.

But folk psychology does not appear to admit the possibility of inefficacious state consciousness. The folk have a concept of conscious and nonconscious mental states; particularly since Freud's work, both "versions" of fears and desires are generally acknowledged, for example. But we tend to hold that a state's being conscious makes an important difference. As Owen Flanagan remarks, "Even if we no longer believe that consciousness is involved in all mental activity, we tend to think that when it is involved, it is centrally involved" (1997, p. 360). For example, we accord a special status to "conscious decisions" as being more reflective and able to overcome reflexive or impulsive behavior, which may explain the idiomatic force of

"raising consciousness" as a means of altering such behavior. We regard nonconscious cognition, in contrast, as resulting in more "automatic" behavior: My nonconscious volitions may result in poor posture, but I can make a "conscious effort" to maintain better posture. Moreover, "focusing" or "concentrating" – pretheoretically understood as willfully increasing one's degree of consciousness – is also thought to improve performance in many cases.

The putative causal difference that consciousness makes is not only important in terms of *what it causes* (e.g., adaptive behavior) but also in terms of *what is doing the causing*. Rightly or not, we tend to identify our mental selves primarily with our conscious mentality, likely because conscious states are the only ones we have introspective access to and are thus more acquainted with in a day-to-day context. So we tend to think that when these states are exerting their distinct causal influence, *we* are more in control of our behavior than otherwise. That is, the actions that result are more "agentive" than those resulting from nonconscious mentality. As such, we regard actions with this causal origin as particularly incurring of responsibility. For example, as an excuse for leaving the milk carton on the counter instead of returning it to the refrigerator, one might say, "I didn't do it consciously," meaning that one didn't consciously will to do it, or arrive at that volition via conscious thought. One isn't usually completely relieved of responsibility in such a case, of course, since the nonconscious cognition that yielded the act is still regarded as part of one's mental self. But on the assumption that conscious states constitute one's essential mental self, one is metaphysically "less involved" with the act, and so less responsible. At least that seems to be the folk intuition. Now, by "I didn't do it consciously," one might just mean that one didn't do it deliberately. Here one would be equating (or associating) a conscious way of doing things with an *intentional* way of doing things. Conscious causation on

this construal would still be important in terms of what is doing the causing: it alone entails the production of behavior by one's will.

Agency issues aside, the folk view that consciousness adds efficacy to mental processing, while pretheoretic, is not mere intuition. It seems to have some empirical support, beginning with introspection. One can, supposedly, experience one's conscious states as causally active. Max Velmans, for example, considers consciousness inefficacious, but only from the third-person perspective. "From a first-person perspective, things look very different," he writes. "Consciousness appears to exert a *central* influence on human affairs" (1991, sect. 9.3). I agree that many of our mental states, particularly desires and volitions, can be introspected as causally active. But I distinguish the property *being conscious* from the states that instantiate it, and I doubt that the property itself is introspected as causally active.[1] This is not to deny that we feel agentive or in control of our actions. However, I don't think that this phenomenon is the feeling of volitions (or beliefs, desires, etc.) bringing about acts *in virtue of being conscious*. Indeed, as I will argue in Chapter 4, we typically feel in control of our acts without being aware of such forgoing states at all. In any case, first-person experience remains a *possible* source of evidence the folk admits for the idea that consciousness is efficacious, rendering that idea more than intuition.

Another empirical source is the observation that adaptive and/or complex actions, both bodily and mental, typically stem from conscious cognition. One doesn't learn to play a new piano piece, or solve a challenging math problem, without any conscious perceptual attention or conscious thought about the activity. Indeed, we find it hard to fathom such a feat. This correlation suggests the "central influence" of consciousness on our affairs: consciousness is recruited for our most demanding cognitive and behavioral tasks. But it does not *prove* that

[1] My argument that the property can be introspectively distinguished is found in Ch. 2, sect. 5.

influence. The epiphenomenality of consciousness is consistent with this correlation. That is, consciousness may arise from cognition that controls adaptive and/or complex action while being causally irrelevant to that cognition. As I argue in Chapter 3, it may even be a *necessary* concomitant of such a process while still being epiphenomenal. And despite the importance we accord consciousness in our self-concept, the epiphenomenality of a property that some (likely relatively few) mental states instantiate seems more credible than the epiphenomenality of the mental as a whole.

Yet the causal efficacy of the mental itself has been theoretically challenged in various ways, which I review in Chapter 2. It has been argued that (i) the mental is "causally excluded" by the physical, meaning that any mental or behavioral event has a sufficient neural cause that preempts a mental one; (ii) mental states are anomalous, which is to say they do not fall into law-like regularities; and (iii) representational content is inefficacious within the mind/brain, as it is constituted by external conditions. These arguments aim to show that a mental state, or mental content, isn't the sort of thing metaphysically that can affect the mind/brain. Anyone who argues that consciousness is efficacious, as I ultimately will, must hold that these challenges can be met, as the efficacy of the mental is *necessary* for the efficacy of consciousness. How can a mental state property be efficacious without mental states being so? Mental efficacy isn't *sufficient* for conscious efficacy, however, since consciousness may be a causally irrelevant property of causally active mental states. Now, it may be that the same solution to the problem of mental causation can be applied to the problem of whether consciousness can be causally relevant. But this will depend on what it is for a state to be conscious. Suppose, for example, that a state's being conscious is just its having a kind of representational content; then, establishing that representational content can be efficacious within the mind/brain establishes that consciousness

can be efficacious. Suppose that a state's being conscious consists in its being represented by a separate, higher-order mental state; then, establishing that mental states are not causally excluded by neural states establishes that higher-order states – and thus instances of consciousness – can be "causally included" as well.

2. Theory Informs Function

Showing that consciousness is the kind of phenomenon that *can* affect the mind/brain is not to show that it actually does, nor what its effects are. Naturally, inquiry into function must be informed by a theory of the nature of consciousness. Uriah Kriegel puts it succinctly: "In order to understand what consciousness *does* … we must have an agreement on what consciousness *is*" (2004, p. 171). I will assume a higher-order theory, which generally says that for a state to be conscious is for it to be represented by the mind/brain. This metarepresenting is taken to explain a distinctive feature of conscious states, namely, that they are states that not only make us *aware of things*, but that we're *aware of being in*. In particular, I subscribe to what Ned Block has called an "ambitious" version of higher-order theory, one that takes higher-order representation to constitute phenomenality – there being "something it is like" to be in a mental state.[2] In contrast, a "modest" higher-order theory simply claims that higher-order representation makes for *one kind* of consciousness (introspective or "monitoring" consciousness), allowing a different explanation altogether for phenomenality.[3] So in arguing for a certain function of higher-order representation, as I do in Chapters 4 and 5, I intend to be arguing for a function of

[2] Thomas Nagel (1974) put forth this notion with regard to what it's like to be a particular creature, while Block (1978) extended the notion to what it's like to be in a particular mental state.
[3] See Block (2009).

phenomenality. But clearly, establishing that higher-order representation generates phenomenality is a project in its own right.

If that project is successful, theorizing the utility of higher-order representation is no easy next step, as compared to theorizing the utility of other proposed reductive bases of phenomenal consciousness. On first-order theories, for instance, it is clear that having conscious states greatly benefits a creature. A state is conscious simply in virtue of making the creature in that state aware of something – an external object, in the case of sense perception. Such a state is an *act* of awareness, according to Fred Dretske. "If we agree about this," he writes, "then the function, the good, of state consciousness is evident. It is to make creatures conscious." And creature consciousness – its responsiveness to stimuli – is of course crucial to survival. "If there is a biological advantage in gazelles being aware of prowling lions, then there is a purpose in gazelles having conscious experiences," says Dretske (1997, sect. 2). However, evidence for *nonconscious* states that make us aware of things in the environment poses a problem for this sort of account. Subliminal perception is a case in point. Experiments involving normal subjects presented with masked primes[4] or special kinds of patient with impaired visual systems (e.g., blindsighters, prosopagnosiacs[5]) have shown that participants can perform tasks where certain perceptual information would be required, while reporting no awareness of that information.

The higher-order theorist's conviction that we would not count these subjects' perceptual states as conscious weighs in here, and motivates inquiry into a reductive base for state

[4] These are briefly presented targets, such as a word or color patch, that are (supposedly) perceived nonconsciously due to "masks," irrelevant stimulus patterns presented before and/or after the target. The targets can nonetheless impact perceptual areas and "prime" (facilitate) certain behavior, despite the subjects' reporting that they perceived nothing.

[5] The former have damage to primary visual cortex (area V1) and thus experience a "blind field," but presumably can still nonconsciously see targets in that field, as they can respond to them with certain proficiency. See, for example, Lawrence Weiskrantz (1986). The latter have damage to fusiform gyrus, located in the temporal lobe of the visual system, and are thus unable to recognize faces. Still, there is evidence that they can process face information nonconsciously, as they seem to exhibit different emotional responses to familiar and unfamiliar faces. See, for example, R.M. Bauer (1984).

consciousness different from – or in addition to – first-order representation. Many theorists posit a *functional* property that distinguishes conscious states, presumably one that would not be instantiated in the case of subliminal perception. According to Global Workspace Theory, that property is the state's being globally accessible,[6] which is for its content to be widely broadcast in the brain – particularly to one's belief/desire complex and memory systems. This "accessibility" can be explained in terms of the state's dispositions. Michael Tye, for example, stresses that a state is conscious in virtue of (inter alia) *being poised*; that is, standing "ready to make a direct impact on beliefs and/or desires" (2000, p. 62). Similarly, for Jesse Prinz, representations are conscious in virtue of being *attended,* which is to say they are available to working memory. Information so available can be manipulated in nonroutine ways, enabling flexible responses to our environment. This capacity is both clearly valuable to the creature and something that subliminal perceivers appear not to have, as I discuss in Chapter 5.

But it remains theoretically possible that the reductive base of phenomenal consciousness is higher-order representation, and that first-order states accessible to working memory tend to be targeted. The utility of consciousness is then far less obvious than if the property *were* accessibility and/or some type of first-order representing. For it's the latter properties that constitute the processing of information about a creature's environment: representing external conditions and then transferring that information to higher cognitive areas. The creature's representing the first-order state – and thereby being in a *conscious* first-order state – seems psychologically irrelevant to that process. It also seems causally irrelevant, as a state's *being represented* is a relational property and so unable to add to its causal powers – rather like calling a tennis match does nothing to alter the causal powers of a serve or a volley. As Dretske observes, on the higher-order view, "Mental states and processes would be no less effective in

[6] See, for example, B.J. Baars (1997).

doing their job – whatever, exactly, we take that job to be – if they were all unconscious." It follows that state consciousness would fail to contribute to a creature's fitness, as Dretske illustrates: "Since the gazelle … can be conscious of a lion without being conscious of the internal states that make it conscious of the lion, it can be conscious of the lion – i.e., see, smell, feel and hear the lion – while occupying no conscious states at all. This being so, what is the purpose, the biological point, of conscious states?" (1997, sect. 2).

But as I argue in Chapter 2, the causal irrelevance of consciousness to mental processing does not follow from the higher-order view. Higher-order states, though they cannot alter the causal powers of their "target" states, will have causal powers of their own, grounded in their representational content. That is, the information about one's current mental state provided by a state's being conscious will have cognitive effects, perhaps beneficial ones that qualify as a *function*. Kriegel (2004), for example, essentially argues that this function is to provide just enough information about one's current mental state so that one may easily obtain more information if needed – a function he claims is similar to that of peripheral visual information about one's environment. But I think that more work is needed to show *why* information about one's current mental state is useful, by showing how it would be used – whether the information is about an occurrent perception, emotion, volition, thought, and so forth. That is largely my focus in this project, beginning with Chapter 4, sect. 9, where I argue that higher-order representation enables metacognitive reasoning about an affective state: the feeling of control over one's actions. I extend this account in Chapter 5, discussing the value of metacognition with regard to other kinds of mental state. I also explain how this function accounts for the general correlation between conscious perception and flexible behavior (something that access-oriented

theories can readily explain), between conscious volition and nonroutine behavior, and between conscious thought and challenging thinking.

3. Event Metaphysics

As mentioned in sect. 1, investigating the function of state consciousness clearly commits me to there being mental states or events, and that they are the sorts of things that enter causal relations. I also assume that events, like objects, have properties; for example, some mental events have the property *being conscious*. And it is in virtue of some of an event's properties, but not necessarily all, that it causes another event. This is the "qua issue" I take up in Chapter 2. For example, the brick's flight causes the window's breaking qua the brick's hardness, but not its brown color. Strictly speaking, however, hardness and brownness are not properties of the *event* (the brick's flight), but of a *constituent* of the event (the brick). I follow Jaegwon Kim (1973) in construing events as property instantiations over time[7]: They are made up of states of affairs – in turn made up of objects and their properties – extended in time. So here the brick and *being brown* are the constituent object and property, respectively, of the brick's flight. The event itself instantiates certain spatiotemporal properties.

In this section I examine whether consciousness can be understood as a property of a mental event or a property of a mental event constituent. And if the latter, how is it that consciousness can add to the causal powers of the event? Such powers must supervene on *event* properties, and properties of an event constituent aren't (necessarily) properties of the event. Consider the following two claims regarding the causal powers of any event *e*:

[7] For more discussion on the nature of events, see Ch. 2, sect. 2.

(i) Every constituent of *e* adds causal powers to *e*.

(ii) Every causal power of *e* is grounded in one of *e*'s properties.

Let's assess these in turn, with *e* being a particular cruise of some given ship. Regarding (i), some of *e*'s constituents are clearly efficacious: They are exercising their causal powers, which in some cases entails that *e* is exercising those powers. The shape of the hull, for instance, is causally relevant to producing the pattern of waves around the ship, and so the cruise is causing that pattern.[8] But a constituent need not be *exercising* a causal power in order to count as *adding* that power to *e*, for causal powers are conditional. If, for example, Bill Clinton is a constituent of the cruise, he adds the causal power of producing tabloid coverage under conditions that include, inter alia, the press's knowing that he is on the cruise. Once we recognize this conditionality, it's difficult to think of a constituent of the cruise that does *not* add to *e*'s causal powers. So (i) is plausible.[9] Less support is needed for (ii); indeed, it seems axiomatic, as a thing's causal powers supervene on its properties. What else *could* ground *e*'s causal powers?

Holding both (i) and (ii), however, appears vulnerable to a reductio, as follows: Per (i), if *wearing a tropical shirt* is a constituent of the cruise (because it is a property of Bill Clinton),

[8] I do not claim that *all* of a constituent's causal powers transfer to the event. The effect of any causal power that does so transfer must be such that it obtains, or would obtain, outside the spatial region of the event. So if a passenger on the ship – a constituent of the cruise – lifts a fork, the cruise does not lift a fork. Being a constituent state of affairs of the cruise, the rising fork is not outside the cruise's spatial region. It is an *ontological* consequence of that cruise, not a causal one. On the other hand, if the passenger knocks a glass into the sea, then it is correct – if unusual – to say the cruise caused the glass's falling into the sea, insofar as the cruise subsumes the passenger's moving elbow.

[9] But note that its converse – namely that all of *e*'s causal powers are grounded in constituents of *e* – may not be. The cruise's spatiotemporal properties are arguably not among its constituents: It *takes up* certain times and places, but it is not *made up* of them. Yet it has certain causal powers in virtue of its spatiotemporal properties: to cause waves at certain times and places, for example.

then *wearing a tropical shirt* adds some causal powers to *e*, say the power to produce, under certain conditions, a representation of the shirt in a tabloid picture of Clinton.[10] Call this causal power *P*. Per (ii), *P* must be grounded in a property of *e*. So, *wearing a tropical shirt* is a property of *e*. But this result is clearly absurd, as the cruise doesn't wear a shirt.

I propose to resist the last step in the reductio: It is true that the constituent *wearing a tropical shirt* grounds *P*. It is also true that *P* must be grounded in a property of *e*. What doesn't follow is that *wearing a tropical shirt* is the property of *e* that grounds *P*; rather, it is this property: *being (partly) constituted by the property <wearing a tropical shirt>*. The latter property *is* instantiated by *e*. It's an example of what I call a "property of constitution." In general, each constituent of an event *y* corresponds to one of *y*'s properties of constitution. So if an object *x* is a constituent of *y*, then *y* has the property *being constituted by x*. And if *Q* is a property of *x*, then *Q* is also a constituent of *y*, and *y* has the property *being constituted by Q* (though *y* does not necessarily itself instantiate *Q*). Furthermore, if an event's spatiotemporal properties are not among its constituents – as they do not seem to be – then they do not correspond to any of the event's properties of constitution. Rather, they are simply properties of the event, along with its properties of constitution.

Now, each of an event's properties of constitution affords the event the same set of causal powers afforded by the property's associated constituent. Each property of constitution is therefore causally relevant to some possible effects of the event; and for any actual effect, a given property of constitution may or may not be relevant to it. For example, the cruise's being constituted by the property *wearing a tropical shirt* is causally relevant to the publication of the

[10] I merely assume here that the cruise would become part of the causal history of the shirt's being represented in the tabloid. There are, of course, more proximate and perhaps interesting causes, such as the editor's decision to run the picture. There are also more specific events we would normally *cite* as causes in lieu of the cruise, such as Clinton's sunning himself in the area of the ship where the photojournalist happened to be. But that does not negate the fact that the broader event, the cruise itself, is a cause of the image.

Clinton picture, but irrelevant to the pattern of waves around the ship. Similarly, the WTC attack (understood as the terrorists' flying the planes toward the buildings) has the power of destroying the WTC qua certain properties of constitution but not others: For example, its being constituted by the terrorists, and by the terrorists' having working eyes, are efficacious, but not the event's being constituted by the paint on the planes.[11]

Since the property *being conscious* may be intrinsic to a conscious state or extrinsic to it (depending on one's theory of consciousness), it's important to understand how the intrinsic/extrinsic distinction applies to events in general. Like an object's extrinsic properties, an event's extrinsic properties are its relations to things apart from it, in the spatiotemporal sense. The Civil War, for example, has the extrinsic property *occurring before WWI*. An event's *being constituted by x*, however, is clearly not a relation to anything apart from that event. So an event's properties of constitution are intrinsic to it.

Now let's apply the intrinsic/extrinsic property distinction to a conscious mental event *m*. If *being conscious* is intrinsic to *m*, it's a property of constitution: *being constituted by being conscious*. To see why, first consider the metaphysical structure of *m*: Like a physical event, it's a property instantiation (or set thereof) over time. Naturally, *m* is the person's instantiating some mental property over time. For example, if *m* is a particular state of anxiety, the person and *being anxious* are its constituents.[12] But more specifically, the person is *consciously* anxious, meaning

[11] Furthermore, since properties of constitution are relative to certain spatiotemporal points within the event, apparently contradictory ones can be instantiated by the same event. For example, the dying of the WTC inhabitants has the property *being constituted by being a mother*, and *being constituted by not being a mother* – in different spatial regions of the event.

[12] Let me address a couple of possible objections to this analysis. First, it seems that the anxiety is both a mental event and a mental property, and a thing can't be both an event and a property. But there's no incoherence: it's simply convenient to use the same term for two metaphysically different but intimately related phenomena. We call the person's being anxious at *t* (the mental event) "anxiety" and *being anxious* (the constituent mental property) "anxiety." The second objection is perhaps more serious. If we make the physicalist assumption that a mental event occurs entirely in some region of the brain, then its constituents must exist entirely in that region. But if a mental

that his being anxious has a certain phenomenal character. *Being conscious*, then, is a higher-order mental property, i.e., a property of a mental property.[13] As such, it's a constituent of *m*, along with the person and *being anxious*. So when we say that *m* is conscious on an intrinsic understanding of that property, we must mean that *m* is (partly) constituted by *being conscious*. This makes sense if we think about what specifically has the phenomenal character: it's not *the person's being anxious at t* that has it, it's the *being anxious*. The question of causal role then becomes: what causal powers, if any, does *being partly constituted by being conscious* add to *m*?

On the other hand, if *being conscious* is extrinsic to *m*, the *event* instantiates consciousness, not its mental property constituent. So here *being conscious* is not a property of constitution. On this kind of view, *m* is conscious in virtue of being represented by another mental event, or in virtue of playing a certain functional role in the mind/brain, depending on the theory. Regarding the first case, consider the higher-order thought version: the HOT would have the representational content *I am feeling anxious*, so clearly it's *m* that's being represented, i.e., the person's instantiating anxiety now. Regarding the second case, since I'm taking events to be causal relata, and since functional properties are causal properties, it's the *event* of anxiety that instantiates whatever functional properties such a theory equates with consciousness.

---

event is *a person's* instantiating mental property *P* at *t*, then the person, as a constituent of the event, must exist entirely in one of his brain areas, which is false on any natural understanding of what a person is.

There are several responses a physicalist can give here: First, she can reject the idea that a mental event occurs entirely in the brain, but nevertheless insist that the mental event *supervenes* on a neural event. So the person's being anxious at time *t* might supervene on the limbic system's having some activation property *A* at *t*. Alternatively, she may hold to identity theory instead of supervenience theory, but insist only on mental/neural *property* identity, not *event* identity. So being anxious = *A*, and both properties occur entirely in the brain; but the person's being anxious at *t* is *not* identical with the limbic system's being *A* at *t*, since these events are not spatially coextensive. Third, the physicalist may argue that the person's being anxious at *t* is not identical with a *strictly* neural event, but with a broader physical one: the person's *body* exhibiting *A* at *t*.

[13] Not to be confused with higher-order theories of consciousness: the property that *being anxious* instantiates – *having phenomenal character* – doesn't represent the lower-order property. It's higher order metaphysically, not representationally.

II. MENTAL CAUSATION AND CONSCIOUSNESS

1. Introduction

What is it for a mental state's *being conscious* – what I will sometimes refer to as a state's *c-property* – to be efficacious within the mind/brain? It would seem that instantiating the property makes a difference to a state's ability to cause mental or bodily events. In Kriegel's words, consciousness would contribute to or modify the state's "fund of causal powers" (2004).[14] More formally, let F be the set of causal powers a mental state m possesses, let G be the set of causal powers of that same mental state when it is conscious, m_c, and assume $F \neq G$. Since differences in causal powers supervene on differences in properties, $F \neq G$ entails that m_c's distinct causal potentiality is due to its c-property. A theory that holds consciousness to be causally relevant claims, at minimum, that at least once during an agent's lifetime, she is in a mental state that is like m_c, i.e., one whose causal potentiality depends (in part) on its being conscious. A more theoretically robust and appealing claim is that an agent is regularly in such states, and that the causal powers afforded (or negated) by consciousness significantly affect thought and behavior. There are, of course, further questions as to the kind of causal difference the property makes, the means by which it makes that difference, the cases in which it facilitates mental or physical performance, and so on. Yet a theory of "conscious causation" may well be a nonstarter if mental states themselves are causally inert: How can consciousness make mental

[14] A theory that specifies this contribution meets what Kriegel calls the "singularity requirement," which is to "distinguish between the causal powers that a conscious state has and the causal powers it has precisely in virtue of being conscious" (p. 174). Rosenthal makes the same point with regard to the state's function: "[F]or states that are conscious, we must distinguish the function that is specifically due to its being conscious from the function that results from others of its psychological properties" (2008, p. 830).

processing more (or less) efficacious if it is not mental processing, but *neural* processing, that does the causal work?

In this chapter I examine the issue of conscious efficacy in the context of more fundamental problems for mental causation. In brief, the three problems I will primarily deal with pose challenges to the efficacy of mental events as follows:

(i) Content Externalism: Mental events, at least intentional ones, have the content they do in virtue of relations they bear to external conditions, as Hilary Putnam (1975) originally argued. And content that is externally constituted cannot be efficacious within the mind/brain.

(ii) Anomalism: A causal relation must subsume its relata under strict laws, and, according to Donald Davidson (1991), there are no such laws correlating mental and neural events. So mental events, it seems, cannot cause neural events.

(iii) Exclusion: The fact that physical events are part of a causally closed system entails that every neural event has a sufficient physical cause. So, barring overdetermination, mental events are unable to cause neural events. This result is essentially a step in Kim's Causal Exclusion Argument.

I attempt to show what these problems imply for the efficacy of a conscious mental state's c-property, given various theories of its nature. With regards to (i), the only type of theory on which consciousness is immune to the externalist problem is one that takes the property to be intrinsic to the state that instantiates it, or at least to the mind/brain. The higher-order approach, though it construes the c-property as a kind of internally directed intentionality, is not immune, I will argue. In addressing (ii), I argue that while *nonstrict* regularities may be held to ground the causal relation for mental states (as some philosophers have suggested in response to Davidson), even regularities of this kind cannot readily be cited for consciousness. Prima facie, then,

consciousness is an anomalistic mental property. Regarding (iii), I claim that even if conscious mental events can be plausibly "included" in the etiology of neural events and behavior, their efficacy *qua being conscious* is not entailed, on virtually all theories of the c-property.

2. The Qua Problem and the Nature of the C-Property

In accord with one convention, I assume that mental and neural *events* are metaphysically suitable as causal relata,[15] although not all properties or constituents of an event may be causally relevant to a given effect. Thus, if event *c* = my bowling ball strikes the last pin and event *e* = the last pin tips over, the ball's being spherical (a constituent of *c*) is causally relevant to *e* but not its having the name of the bowling alley engraved on it.[16] Similarly, if *c* = my wondering what to eat for lunch and *e* = my deciding on a steak sandwich, *c*'s including an inviting mental image of such a sandwich is likely causally relevant to *e*, but perhaps not the color of the dish on which I happened to imagine the sandwich, and perhaps not *c*'s being conscious. Other examples can be drawn from Hume's associative psychological laws: Bringing to mind the image of a tree may cause one to think of shade (ideas of causes bring to mind ideas of their effects) or perhaps a

[15] See Davidson (1967). States can also be causal relata insofar as they are construed as events that exist "entirely" at each point in their duration: e.g., visualizing blue drapes on a window for a few seconds vs. working through a long-division problem. So one's ongoing anxiety about a test, or standing belief that it is "hard," is a mental state that may affect one's thinking throughout the test taking (e.g., causing one to ascertain the answers to the problems less quickly than otherwise). Similar examples can be given for brain states such as inebriation.

[16] On Terry Horgan's view (1989), the causal relevancy factor makes causation into a four-place relation he calls "quausation": *c* qua *F* causes *e* qua *G*. Thus, the engraving might be causally relevant to the effect qua a falling at a very precise speed: since the engraving alters the ball's weight fractionally, it will alter the speed of the ball fractionally, and in turn the precise speed at which the pin falls. Nevertheless, it is plausible that there are constituents of the cause that fail to be causally relevant to *any* property of the effect, such as the particular shape of the engraving (within a range of shapes that make no difference to weight) or the color of the ball.

certain formal proof method (ideas of similar things bring each other to mind).[17] But neither of these events, it would seem, are brought about in virtue of the tree image's green quale (its being an experience *as of* green), which represents a color property that bears no causal association with shade and no similarity-based association with a branching form.

Clearly, if a property is causally relevant to an effect, it must be a property of a cause of that effect. Assume event *c* is *F*. If *F* is causally relevant to event *e*, then *c* is necessarily a cause of *e*. The converse entailment, however, does not hold: if *c* is a cause of *e*, then *F* is not necessarily causally relevant to *e*. Thus, suppose an event with both mental and neural properties – e.g., one that is both a conscious volition and an action potential in the motor cortex – causes a finger movement. Despite the fact that a mental event causes the finger to move, its efficacy is not necessarily in virtue of being mental; for example, if the "causal exclusion" argument is sound, the potential is sufficient for the movement, and thus the event's being a volition, conscious or not, is causally irrelevant. Now suppose the exclusion argument is refuted and the conscious volition is shown to be a legitimate cause of the finger movement. It would not follow from such a demonstration that the volition causes the movement *in virtue of being conscious*. Thus, per an ontological view that parses mental events and their neural substrates as single events with both mental and physical properties, a double "qua" problem must be resolved to secure the relevance of consciousness to producing the finger movement: the event must be a cause qua mental event *and* the mental property must be causally relevant qua conscious mental property. Alternatively, if we individuate two events, a mental one accompanying (and perhaps supervening on) a neural one, we still face a qua issue: Assuming mental event *c* causes event *e*, does *c* cause *e* qua *c*'s being conscious? This, of course, is a specific qua problem concerning the

[17] Here I only sketch how properties of mental events, like those of physical events, might be causally relevant/irrelevant. I am not assuming, much less arguing, that representational or qualitative properties are efficacious.

causation of a particular event, *e*. The general qua problem we are interested in is whether conscious mental events *ever* cause other mental events (or bodily ones) qua being conscious. It seems that arguments for mental causation do not establish that they do: Every conscious mental event could be efficacious without any such event being efficacious qua its c-property. This scenario is possible on all theories of state consciousness except one type of functional theory, as I will explain.

Now, most theories take the c-property to be extrinsic, as opposed to intrinsic,[18] to the instantiating state. But the qua problem can arise on either view. Returning to the case where event *c* = my bowling ball strikes the last pin, and event *e* = the last pin tips over, we can ask whether *c* causes *e* qua the bowling ball's sphericality. If *p* = Taft prosecutes a monopoly, and *d* = the monopoly dissolves, we can ask whether *p* causes *d* in virtue of Taft's metabolic processes. Each of these qua questions applies to an (arguably) intrinsic property of a causal event: For instance, *c* in the first case has the bowling ball as a constituent object, and the ball's sphericality as a constituent property. So we ask whether *c* causes *e* in virtue of *c*'s *being partly constituted by sphericality*, and whether *p* causes *d* in virtue of *being partly constituted by <having metabolic processes>*.[19] But consider *c*'s *occurring while gravity affects the pin* and *p*'s *being recognized as an act of the president by court judgment*. The instantiation of these properties depends ontologically on events external to *c* and *p*: gravity affecting the pin and court judgment, respectively. So they are extrinsic properties of these events. Yet it seems we can still say that *c* causes in virtue of *occurring while gravity affects the pin*, and that *p* causes qua *being recognized*

[18] The notion of intrinsicality is admittedly contentious, and a full discussion of what it means and what kinds of property are intrinsic is beyond the present scope. But roughly, a property of an event is intrinsic to it if that property ontologically depends only on how things are within the event's spatial extent, which may of course vary across the event's duration. As Sidney Shoemaker explains, "If a question about whether a thing has a property at a place and time concerns a genuine nonrelational property, the question is most directly settled by observations and tests in the vicinity of that place and time" (1999, p. 260).

[19] See my discussion of "properties of constitution" in Ch. 1, sect. 3.

as an act of the president by court judgment. The reason is that the causal powers in question depend (at least partly) on *c* and *p* standing in certain relations to those events: *occurring while* and *being recognized as a presidential act by*, respectively. The extrinsic properties thus enable *c* and *p* to have their effects.[20] Similarly, a causal power of a conscious state *m* may depend on *m*'s standing in a certain relation to a cognitive condition external to *m* – an extrinsic property that makes *m* conscious. (One example is *m*'s being represented by another mental state, as will be discussed.) And then *m* could be said to exert that power qua having that extrinsic property, qua being conscious.

In general, then, an event *x* may cause an event *y* qua *x*'s being related in a certain way to another event or state *z*. Let E_1 be this extrinsic property of *x* and let *R* be the relation between *x* and *z*. Note that the existence of *R* also metaphysically entails that *z* possesses its own extrinsic property (E_2) of being related to *x* in a certain way. When *R* is symmetric, that "way" is the same as how *x* is related to *z* (e.g., if *x* is similar to *z*, *z* is similar to *x*). When *R* is asymmetric, *x* and *z* will be related in different ways (e.g., if *x* precedes *z*, *z* succeeds *x*). I will refer to E_1 and E_2 as *complementary* extrinsic properties. Each is instantiated by one relatum; together, E_1 and E_2 constitute *R*, which is instantiated by both relata. With regard to the qua issue, the point I wish to make is: *x* qua E_1 causes *y* iff *z* qua E_2 causes *y*. More fully:

(Q) Where E_1 and E_2 are complementary properties (respectively) of *x* and *z* constituting a relation *R*(*x*, *z*), then *x* qua E_1 causes *y* iff *z* qua E_2 causes *y*.

[20] It follows from this analysis that for a set of events *N* that is necessary for an event *e* to occur, we can say that each member of *N* causes *e* qua accompanying the other members.

Consider: My bowling ball's striking the pin – qua *occurring while gravity affects the pin* – causes the pin's tipping. It follows that gravity's affecting the pin – qua *occurring while the pin is struck* – also causes the tipping. Similarly, if Taft's prosecuting a monopoly causes the monopoly's dissolving qua the prosecution *being recognized as an act of the president by court judgment*, then the judgment also causes the dissolution qua its *recognizing the prosecution as an act of the president.* Essentially, an event's having some extrinsic property metaphysically entails a relation between that event and some other event or condition, and thus the instantiation of the *complement* to that property by that condition. Thus, if the property is causally relevant to an effect, so will its complement be. This analysis, and principle (Q), will be important when I discuss the efficacy of the c-property on higher-order theories, later in this section.

First let us look at how the qua issue arises on first-order theories of state consciousness. Most construe the property as extrinsic in some way, including (arguably) sense-datum theory. Consider the qualitative character of mental states, or "what it is like" to be in them. Such properties are particularly salient in the case of perceptions, bodily sensations, emotions, and mental images: When one consciously sees a green swatch, for instance, there is a certain greenish, square-ish way the experience is like. Sense-datum theory does say that such character is intrinsic to the mental state, but it also claims that the mental state is the direct object of experience: When one sees the swatch, for example, one is actually aware of a mental intermediary that *itself* is green and square in a literal sense. But while the theory claims that qualitative properties belong to the sense-datum, it remains questionable whether the sense-datum's *being conscious* is intrinsic to it. That's because the sense-datum is not the *experiencing of* the physical swatch, it's the (direct) *object of experience*, by which one (indirectly) perceives the physical swatch. Thus, there is a mental act of awareness distinct from the sense-datum, in

virtue of which it is a conscious mental state.[21] The awareness does not constitute the swatch sense-datum, as does its greenness and square shape. Rather, it is a relation between the datum and the subject, and so counts as an extrinsic (though still intramental) property: the datum is *experienced by the subject*.

A different view of qualia holds that they are actually nothing like physical properties: *phenomenal green* and *phenomenal square-ness* do not make for a green, square object in mind. Rather, they make one (directly) aware of the external swatch. Since the experiential properties of my swatch-perception are not properties of an object of experience, the perception's c-property is more arguably intrinsic to it. Furthermore, unlike the literal greenness and square-ness of the swatch sense-datum, phenomenal green and phenomenal square-ness are more plausibly theorized to be instantiated in the brain, a view of qualia that is palatable to the materialist. That is, they are mental properties, but reducible to neural ones.

Now, representationalists about qualia hold that such a reduction can only be accomplished by first showing how phenomenal character is a kind of representation, as the brain does not obviously have experiential properties, but it clearly has representational ones. Contra the sense-datum theorist, the representationalist insists that what our experience is like when we consciously see the swatch – greenish, square-ish – is not constituted by properties of the experience, but by properties of *what the experience is about* (i.e., its content). The experience itself – the qualitative state – has the properties *representing greenness* and *representing square-ness*. And since such properties make us aware of things (like a green swatch), many representationalists, such as Dretske, hold that a perceptual state is conscious in

[21] It might be pointed out that a sense-datum is a mental object, not a state or event. But where there is, say, a bright-red tomato sense-datum, there is surely a state of tomato imaging.

virtue of them.[22] But since it is widely acknowledged that *nonconscious* states can have representational content, representationalists typically add further criteria to distinguish the kind of representation that makes for state consciousness, such as nonconceptuality or fine-grainedness. Once a state's phenomenal properties are equated with representational ones, however, we are led to a view of the c-property as extrinsic, as representational properties are arguably "wide," or determined by the state's relations to the agent's environment. For example, a visual state is about a swatch in virtue of tending to be caused by one under normal conditions.

Yet there is a strong intuition, bolstered by well-known thought experiments like Block's (1997) Inverted Earth, that a state's qualitative character is invariant across changes in its wide representational content, and thus cannot be reduced to it. Block describes a person who is transported to a planet just like Earth except that red objects are green and vice versa. But the person is fitted with lenses that cancel out the switch, so that she doesn't notice anything different and now *misperceives* leaves as green, stop signs as red, etc. Furthermore, Inverted Earthlings use "red" and "green" in the same way we do, so the intentional contents of these words on Inverted Earth are inverted relative to ours. Eventually, Block argues, the newcomer's red/green terms and red/green perceptions would shift in representational content to match those of Inverted Earth's inhabitants: her word "red" would come to be about green, her visual state that was about green objects would come to be about red objects. Yet – and this is Block's point – it's plausible that what it's like for her to be in that visual state would *not* shift during that time, showing that qualia are "narrow," or supervenient on the state of agent's mind/brain.

A representationalist who gives credence to such a thought experiment may instead try to reduce qualitative character to *narrow* representational content. Georges Rey (1998), for

[22] In the case of nonveridical perception, the representationalist maintains that what the state makes us conscious of is not a sense-datum, but rather an "intentional inexistent": uninstantiated universals (Dretske [2003]) or nonexistent objects (William Lycan [2001]).

example, takes this approach, but the content according to him is narrow relative to the agent's mind/brain, not to the qualitative state. Rey argues that qualitative properties reduce to representational ones, and these in turn are constituted by the state's "conceptual role" within the mind/brain. So experiencing green is just representing green, and the state represents green in virtue of its disposition to cause certain other states (e.g., a recollection that one is owed money, if one has that associative link). But the c-property in that case is arguably still a relational one; the (causal) relations are simply "local." Similarly, on "global access" theories, a state is conscious in virtue of being "poised" to impact executive control of action, flexible thinking, and explicit learning and memorization. And such effects, or at least the cognitive conditions that enable the state to have the effects, are clearly beyond the spatial extent of the state. Still, it seems arguable that a state's functional properties supervene on its intrinsic features, as the water-solubility of salt supervenes on its internal molecular structure. Given these considerations, I argue that the issue of whether or not the c-property is intrinsic on a functional theory depends on exactly how the functional property is construed. I distinguish the following three property types:

F1 *Exercised causal power*: A state's being conscious is its causing an event of type *E*. This kind of property is clearly relational, as it is instantiated only if a separate event, one of type *E*, occurs.

F2 *Dispositional causal power*: A state's being conscious is its being poised to cause an event of type *E*. This kind of property is also relational, as the state will be poised in virtue of certain conditions (outside its spatial extent) obtaining, namely those that enable it to cause an event of

type *E*. To give a physical example, the snipping movement of a scissor is poised or disposed to cut string only if the string is in its trajectory.

F3 *Conditional causal power*: A state's being conscious is its having the power, under certain conditions, to cause an event of type *E*. This kind of property is arguably supervenes entirely on the state's internal features. The snipping movement of a scissor is intrinsically such that, were a string in its trajectory, the string would be cut.[23]

In order to qualify as an intrinsic view of the c-property, then, a theory must not reduce it to a wide representational property of the state, or to an F1 or F2 property. That said, the theory may construe the c-property as unreduced qualitative character or as qualitative character that directly reduces to a neural property,[24] or to a representational property that (somehow) supervenes on how things are within the state's spatial extent, perhaps in virtue of supervening on an F3 property.

For all of the theories surveyed thus far, both intrinsic and extrinsic, the issue of the c-property's efficacy can be formulated as a qua problem: Are conscious mental states ever efficacious qua being conscious? The specific question will vary depending on the theory, of course. So an argument for mental causation may establish that all conscious mental states are efficacious, but are any efficacious qua: (i) being constituted by sense data; (ii) being phenomenal (apart from being representational); (iii) being narrowly representational, in the way

[23] Conditional powers can even serve to provide identity conditions for a thing's intrinsic properties, as Shoemaker has argued. See Shoemaker (1999, p. 253).

[24] But if that property is of the F1 or F2 type, then consciousness is *not* intrinsic. Indeed, the view that phenomenality doesn't reduce to mental representation or functionality isn't committed to phenomenality being intrinsic, as it may reduce to a relational property at the neural level. As Block notes, "Many of the theorists who emphasize what-it-is-like-ness [i.e., those who resist reduction to mental representation or function] are also scientific realists, and all the major recent accounts of what consciousness is in the brain have been heavily relational" (2011, p. 420).

that constitutes being phenomenal; (iv) being widely representational in that way; or (v) being functional in that way? Regarding (v), there is one exception to the qua problem, one type of theory on which it would not arise: If the c-property is of the F1 type, then all conscious states are causing something in virtue of it.[25] But if it is of type F2 or F3, then there remains the question of whether the state every causes something in virtue of being poised, or in virtue of having a certain conditional power. A given conscious state may have several cognitive and/or behavioral effects, none of which are due to that functional property. A separate theoretical and/or empirical argument – beyond a general argument for mental causation – would be needed to establish that the dispositions or conditional powers in question are active. The same applies to the other types of theory cited in (i)-(iv): Given the qua issue, establishing that all mental states are efficacious fails to establish that the c-property of any state is efficacious.[26]

Now, there is a view of the c-property on which this point doesn't appear to hold: higher-order representational (HOR) theory. On the HOR view (as it is usually put forth), a mental state is conscious when it is represented by a separate mental state.[27] That higher-order state can be a belief or thought, according to HOT theorists, or a quasi-perception,[28] according to higher-order

[25] It might be insisted that, on this view, the c-property can still turn out inefficacious: if mental state *m*'s being conscious is its causing *e*, the efficacy of *m*'s c-property depends on whether *m's causing e* has any effects, which it may not. But this is to assume that the only way the property counts as efficacious is by *causing* an effect, as opposed to *metaphysically entailing* an effect, which is what it does if it is a kind of efficacy (an F1 property).
[26] In Ch. 4, I deal specifically with conscious volitions, but for now I point out one implication of the qua issue for these states: Even if a conscious volition *was* the cognitive "first cause" of a movement (without the nonconscious neural antecedents cited by B. Libet et al.), the volition may not be efficacious qua being conscious. So consciousness' failing to precede an act's neural initiators is not a *necessary* condition for its being inefficacious.
[27] Not all higher-order views, I should add, posit a distinct state to implement metarepresentation. "Self-representationalists" hold that a mental state is conscious in virtue of representing itself; e.g., a conscious perception of a house is "about" the house as well as itself. (See Kriegel and Williford [2006].) In this case, the consciousness-conferring representational property is not only borne by the conscious state, but directed at the state instead of the external world. It may seem, then, that the c-property counts as intrinsic to the conscious state on this view. But internally directed representation may still supervene on external conditions, as I will argue in sect. 4.
[28] "Quasi" because the way in which one perceives, say, a belief state one is in probably doesn't involve any qualia comparable to the sensory qualities that arise with first-order perceptions.

perception (HOP) theorists.[29] Consider a conscious visual perception of a house. On the first type of view, the seeing is conscious in virtue of being accompanied by the thought that one is seeing a house. On the second, the perception is represented by a hypothesized "internal attention" system. Prima facie, there is no qua problem for conscious causation if the c-property consists in either kind of higher-order representation. On these views, it seems a successful argument for mental causation *would* establish that conscious causation obtains. If every mental event were efficacious, higher-order states themselves would be as well. And this means that the phenomenon of state consciousness is efficacious, since those higher-order states constitute the phenomenon.[30]

The problem with this conclusion is that it is not a HOR, but a specific property of a HOR, that makes for a conscious "target" state. We might say it is the defining property of a higher-order state: representing a first-order state. I'll refer to it as the h-property. On the higher-order view, then, each c-property would have an h-property as a *complement*, in the sense given above. Consider a state *m** that represents another mental state *m*, rendering *m* conscious. The c-property of *m* is *represented by m**, while *m**'s h-property is *represents m*. Together, the h-property and the c-property constitute the representational relation between *m** and *m*. We can then argue as follows:

1. A HOR causes an event qua its h-property iff the target state causes that event qua its c-property (per Q, above).

[29] Theorists who hold this type of view include, respectively, Rosenthal and Peter Carruthers; and David Armstrong and Lycan. The HOT theorist, I should mention, typically stipulates that the HOT must have a particular causal origin, namely, it must be arrived at via what seems to the subject to be a noninferential process. This criterion explains the unmediated way in which we are aware of our mental states. It also staves off certain counterexamples: It seems I can learn about my jealousy via psychoanalysis (thus thinking about it) while that jealousy remains nonconscious. But psychoanalysis would involve inference.

[30] Of course, this leaves open the possibility that the states may have no *useful* effects. I give an account of their cognitive utility in Ch. 5.

2. The efficacy of a HOR does not entail that it causes qua its h-property.
3. So, all HORs could be efficacious without any h-properties being causally relevant.
4. So, all HORs could be efficacious without any c-properties being causally relevant (per 1).

A qua problem for the c-property thus persists on HOR theory. The key premise is 2.[31] HORs have other properties besides the h-property that ground their causal powers. Consider the HOT *I am seeing a blue sky*. It's about my first-order perception, but it's also about blueness. And if, through an associative link, that HOT leads me to think about my blue car, it would be causing another state, but not in virtue of its h-property. And higher-order states may also cause qua their nonpsychological properties. An example provided by Levin is such a state's electric field. Suppose it was strong enough to light up a light bulb and were used for that purpose. The h-property may well be causally irrelevant to this effect.[32]

I should note that 1 may be vulnerable to a reductio if "empty" HORs – i.e., those that lack *actual* target states – are held sufficient for conscious experience. On that view, if, for example, I think that I am seeing red, then I consciously see red, whether or not I am in a first-order visual state of seeing red.[33] I have conscious experiences, then, in virtue of representing myself *as* having those experiences. Given this approach to higher-order theory, the reductio that

[31] Michael Levin brought the point to my attention in correspondence.
[32] If it *were* relevant to the strength of the state's electric field, then the target state's c-property would be relevant to the light bulb's coming on, per 1. While this effect of consciousness would certainly be unusual and prima facie hard to understand, I think that if we knew the precise neural substrate for the HOR and its intentional relation to the target state, presumably it would become clear how the strength of the HOR's electric field is affected as a result. That is, the reductive base of the h-property would provide the explanation for the bulb's lighting up, and we would have to accept that the target state's being conscious can have that unexpected effect.
[33] This view is made more plausible by the point that one can't readily believe oneself to be in any first-order state one chooses, and thus have whatever conscious experiences one chooses. See Rosenthal (2000, p. 233).

1 seems to generate is as follows: Suppose the HOT described above causes another state, *e*, in virtue of its h-property – representing my being in a seeing-red state. Per 1, the seeing-red state causes *e* in virtue of its c-property – being the object of the HOT. But ex hypothesi, one is not in that state; it is a mere intentional object, or what Rosenthal (2000) has described as a "notional state." So we have a causally involved nonexistent state, one that causes *e* qua being represented.

One might avoid this issue entirely by holding that empty HORs, such as false HOTs, do *not* result in consciousness.[34] Thus, it is questionable whether Q holds for the representation relation in the case of "intentional inexistence." And that worry applies to higher-order representation: Where *m** represents *m* but *m* does not occur, does the fact that *m** causes some event *e* qua representing *m* entail that *m* causes *e* qua being represented by *m**? But if we deny that *m** – which represents no actual state – results in consciousness, the c-property's efficacy is not at issue here. Only represented actual states can be conscious, and to those 1 arguably applies.

On the other hand, one can maintain that HORs are sufficient for consciousness, but deny that 1 gives the right criterion for consciousness to be efficacious in the empty case. For in that scenario, the h-property does not literally constitute a *relation* to an intentional object. If that is so, the h-property has no complement: *m**'s representing *m* does not imply there is a state with the property *being represented*. As Rosenthal has put it: "Being in a conscious state is not a mental state's having some special monadic property; rather, it's an aspect of how one's mental life appears to one. So one's being in a conscious state does not imply being in the state one is aware of being in" (2011, p. 432). It just implies being aware of being in that state, which is to represent being in it, which is the HOR's instantiation of the h-property.

[34] Jonah Wilberg (2010) argues for this position.

Now, the grammatical structure of "*m** represents *m*" certainly suggests such a relation. But as adverbial theories of perception have shown, our expression of representation need not imply a relation between representer and represented: If I hallucinate the ringing of a cowbell, I can be said merely to "hear cowbell-wise." Accordingly, where *m* does not occur, *m** can be said to "represent *m*-wise." Of course, we can *speak* as if representation is a relation in the case of intentional inexistence, and thus speak as if Q holds for the relation in that case. So King Lear's madness never existed, since King Lear never did, yet we might say, "Peter O'Toole's performance, qua depicting King Lear's madness, caused a hush in the audience." And, per Q, "King Lear's madness, as depicted by O'Toole's performance, caused a hush in the audience." But this is really to say that the performance had that effect qua being interpreted in a certain way – as being about Lear – not qua standing in a relation to Lear or his madness. Similarly *m** would represent a nonexistent *m*, not in virtue of a relation to *m* (or to some abstract *m*), but rather in virtue of some other properties of *m**. Of course, these would not consist in *m**'s being interpreted as being about *m*, as mental states are supposed to have a *nonderived* intentionality. Presumably, the properties in question would be functional: perhaps *m**'s being a state typically caused by states like *m*, and with the tendency to cause certain other kinds of mental state.[35] We might also adopt a *teleo*-functional semantic theory for *m**: it represents *m* in virtue of being caused by a mechanism whose proper function it is to produce representations of first-order states just when those states occur.[36] While these are extrinsic properties of *m**, none consist in a relation to an occurrent *m*.

When the h-property is reduced in such ways, 1 no longer applies in the empty case: If a HOR causes an event qua its h-property, it does not follow that its nonexistent intentional object

[35] These might constitute a metacognitive process, as will be discussed in Ch. 5.
[36] The notion of proper function as a foundation for naturalizing intentionality was set forth by Ruth Millikan (1984).

causes that event qua its c-property. It only follows that the HOR has caused in virtue of whatever functional properties constitute its *representing m*, which, as noted, need not involve *m*.[37] In fact, on this construal of the h-property, it is arguable that even when *m does* occur, its *being represented* is merely its *accompanying m**'s instantiation of the h-property.[38] That would be *m*'s c-property. Clearly, occurring along with *m** is not a complement to the h-property, which is some (presumably) functional property of *m**. And so on this view, *m*'s *being represented*, i.e., its *being conscious*, does not constitute a relation together with the h-property. It follows that 1 is an improper application of Q, which concerns only relations between events and the complementary properties that make up these relations. Indeed, for *m** to cause an event *e* qua its h-property is clearly not for it to cause *e* qua accompanying *m* (the complement to *m*'s *accompanying m**): to represent *m* is not to accompany it. So *m**'s causing *e* would not entail, contra 1, that *m* causes *e* qua being represented, qua being conscious. In sum, it is arguable that if 1 doesn't give the criterion for consciousness to be efficacious in the empty case, neither does it in the nonempty case.

Assume, then, that a HOR can cause an event qua its h-property without its target causing that event qua its c-property. Are c-properties – mental states' *being conscious* – ever efficacious? I think they can be: Suppose that target state *m* causes some event *e* qua accompanying the state of affairs *m*'s representing m*, which on the present view is just *m**'s instantiating some set of functional properties not involving *m*. That state of affairs is thus enabling *m* to cause *e*. It is a necessary condition for *e* to occur, and so *m**, qua its h-property, is

[37] Note that any effects of *m** would seem to depend on its syntactic or "local" properties, so it seems its representing *m* can't afford it causal powers. This, however, is just the problem externalism about content poses for mental causation, applied to higher-order states. I take up that issue in sect. 4.

[38] In this way, representation can involve a relation to the intentional object contingently, though not constitutively, following Kriegel's suggestion: "Perhaps representation often involves a relation to the represented, but it never does so constitutively. That is, it is never the case that a representation represents in virtue of bearing a relation to the represented" (2007, p. 312).

a cause of *e* along with *m*. We can see that one of the conditionals that comprises 1 still holds: If a target state causes an event qua its c-property, then the HOR causes that event qua its h-property. But this is not because the c-property and the h-property constitute a relation between the states; rather, it is because the causal relevance of *m*'s *accompanying m*'s instantiation of the h-property* (the c-property, on the present view) entails the causal relevance of the h-property to *e*.

We are thus led to the following two candidates for "conscious causation" on the higher-order view:

(i) A HOR causes an event qua its h-property – independently of the c-property of its (actual or nonactual) target.

(ii) An actual target causes an event qua it c-property – entailing that the HOR is a co-cause of that event qua its h-property.

The view that HORs are sufficient for conscious experience would admit (i) as conscious causation. It would also admit (ii), insofar as (ii) is a case of a causally relevant h-property. The contrasting view that actual targets are a necessary condition for consciousness would only admit (ii) as conscious causation: Consciousness occurs only when an actual state is conscious, and is only efficacious when a state causes qua being conscious.

The latter view is motivated by the following objection: How can HORs be sufficient for conscious experience if, in the empty case, the subject is ex hypothesi in no conscious state? In that scenario, there is only, for example, a state with the content *I am seeing red*, but no state of my visual system that represents redness. If the HOR were conscious, then one would consciously think about (or quasi-perceive) that one is in a state of seeing red. But this is not for one to consciously see red, and indeed, HORs are seldom conscious. As Block (2011) argues, the

theory results in incompatible necessary and sufficient conditions in the empty case: The HOR is supposed to be sufficient for a conscious episode, but it is also supposed to be necessary that an episode be represented by the HOR. There is no represented state in the empty case, so there is no conscious episode, contra the sufficient condition.[39]

I think a promising response can be made to this objection by asserting that the HOR is the conscious state, but reconsidering what is meant by "conscious state" or "phenomenal state." It is not necessarily a state whose *content* is conscious, but more broadly one that has a phenomenal property – one that there is something it is like for the subject to be in. So we may consider the HOR a conscious state, though its content isn't conscious: One doesn't *consciously* think one is seeing red. Nonetheless, there is something it's like for one to be in the state, namely, what it's like for one to see red. So in the empty case, there is no state whose content is conscious: there is no first-order state, and the HOR's content isn't conscious. But there is a conscious state in the sense of a state with a phenomenal property: the HOR. On this approach, then, states can be conscious in two different ways: One is by instantiating the c-property, which is for the state to accompany another state's representing it; this entails the target state's content is conscious. The second is by instantiating the h-property, which is for the state to represent oneself as being in a first-order state; this entails the HOR has phenomenality, the phenomenality of being in the first-order state.

For my purposes, the main point that follows from these considerations is that the efficacy of the c-property on higher-order theory depends upon the efficacy of the h-property.

---

[39] Perhaps one may regard the notional state of seeing red as the conscious state. Now, it does seem that notional things have properties, both real and unreal: King Lear and his madness are unreal, but Lear also appears to have the real property of being represented by O'Toole's performance. But this putative property of Lear is quite arguably a property of the actor's performance: *interpreted as being about Lear*. Similarly, a nonexistent red-seeing state's *being represented* plausibly reduces to certain (teleo-)functional properties of a real state: those of the HOR with the content *I am seeing red*. So there is no state – even notional – instantiating the c-property in such a case. The inadequacy of this solution becomes still more acute when we ask about the causal powers of the notional state.

This is important to bear in mind with regard to Chapters 4 and 5, where I will argue that representing first-order states is cognitively useful. The point holds whether we admit (i) or (ii) (or both) as cases of conscious causation. The h-property must be causally relevant in either case. And if we take the h-property and c-property to be complements, then 1 holds and the h-property's efficacy is again critical.

3. The Qua Problem and Mental Event Individuation

The qua problem arises from the fact that any two properties or property-constituents of the same event need not share causal relevance to a given effect of that event. So we might consider Jon's elbow pain and his C-fiber firing to be property-constituents of the same psychophysical event even though only the latter causes Jon's wincing (if the causal exclusionists are correct). Similarly, the circularity of a baseball bat's knob and the bat's hardness can both be considered constituents of the event Jon's-swinging-the-bat while only the latter is causally relevant to the homerun that follows. It is still correct to say of that event that it was a swinging of a bat *with a circular knob* by Jon. What, then, is the metaphysical criterion for properties to count as part of the same event, if not the unity of their causal powers? The answer naturally depends on one's theory of events. According to Kim (1966), an event is a property instantiation at a time. On the view held by Davidson (1980), an event is an occupant of a spatiotemporal region, with its constituents being property instantiations. It follows on Kim's

theory that two properties cannot constitute the same event,[40] though we can surely have a name for a complex of events (e.g., the Civil War), comprised of many property instantiations over durations that all fall between certain times. Under Davidson's theory, in contrast, more than one property instantiation – and thus more than one object – can constitute an event proper.

Kim's theory, then, individuates events more finely than Davidson's, whose only identity condition on any two events is that they occupy the same spatiotemporal region. Consider an example provided by Davidson: a metal ball rotates and heats at the same time. Since both property instantiations (plausibly) take place in the same spatiotemporal region, Davidson's theory would not individuate them as two events. In this case, Kim's theory seems to yield the more intuitive result; but there are cases where his view seems to individuate events at *too* fine a grain. One such case involves predicate modification. Suppose Jon gracefully swings the bat; parsing two concurrent events – a swinging and a graceful swinging – seems counterintuitive. The primary solution Kim has offered is that the graceful swinging should be considered a distinct event, but one that is included in the event of swinging and thus not an independent event. As an alternative solution, he construes the gracefulness as a property of the event itself: Jon's-swinging-the-bat-at-*t* is graceful. On this approach, there is no event of gracefully swinging distinct from the swinging.

Let's consider what the respective theories advocated by Kim and Davidson imply for the individuation of conscious mental events, and the nature of the qua problem for conscious causation on each view. For Kim, Jon's being in pain at *t* and Jon's C-fibers firing at *t* are two property instantiations at times and therefore count as two events. Likewise, Jon's having the conscious volition at *t* to swing the bat and Jon's motor cortex undergoing a certain lateralized

[40] Though an event will have exactly two properties of constitution, corresponding to the object and property that constitute it: a screw's turning will have the property *being constituted by a screw* and *being constituted by a turning*.

readiness potential at *t* are two events. The issue of mental efficacy is then not a qua problem but simply whether the mental events are efficacious along with the neural ones.[41] In contrast, it may not be plausible to parse a mental event's instantiating an intrinsic c-property as a separate event. Doing so makes the theory susceptible to the predicate modification problem: Jon's volition at *t* to swing the bat is an event, but is Jon's *consciously* willing at *t* to swing the bat an independent event? Perhaps it is a distinct event within the willing, per Kim's primary solution to the problem. That solution, however, has been challenged as follows: A graceful swinging cannot, as a matter of course, be included within the swinging as *a mereological part*, as there may well be no part of the swinging that is not graceful. So it is not clear how the graceful swinging is to be included and remain a distinct event. Yet there are still two property instantiations – Jon instantiates swinging as well as graceful swinging – that constitute two events according to the theory. Similarly, there may well be no part of a conscious volition to swing the bat that is nonconscious, and in that case we apparently have just one mental event – despite the fact that Jon instantiates both a conscious volition at *t* to swing the bat and a volition at *t* to swing the bat. So the property instantiation view seems to parse events too finely in the case of conscious mental events: While consciousness is plainly a distinct property instantiation, it should not make for a distinct event. Jon's double instantiation should then constitute a single event at *t*, and we can inquire whether that event causes, for example, Jon's subsequent homerun qua his conscious volition to swing the bat versus his volition simpliciter.[42]

---

[41] We are not precluded from constructing a complex event that subsumes a pair of physical and mental events, and then formulating the qua problem as follows: Is that complex event efficacious qua its mental sub-event? Yet if there is no motivation for compounding an event of this kind apart from posing the qua problem, the exercise becomes ad hoc.

[42] On Kim's second solution to the problem of predicate modification, the event Jon's-willing-at-*t*-to-swing-the-bat instantiates consciousness. Assuming that event causally contributes to Jon's hitting a homerun, the qua problem is whether the volition's being conscious is causally relevant to any property *P* of that subsequent event.

Next we'll formulate the qua problem following the spatiotemporal method of individuating events. On Davidson's monist metaphysics, where any mental event is also a physical event, a mental event must occupy exactly the same spatiotemporal region as a physical event – if they occupied different regions, they would be different events, *ex hypothesi*. So a particular pain occupies exactly the same spatiotemporal region as a particular C-fiber firing, a particular volition to move occupies exactly the same spatiotemporal region as a particular readiness potential, etc. Now, an intrinsic c-property of the volition, for example, can be subsumed spatiotemporally within its respective event, and then we can ask if that event is ever efficacious qua mental event and more particularly qua *conscious* mental event (this is the aforementioned double qua problem). A conscious mental event *m* is, of course, not the same event as nonconscious *m*, and so it follows from Davidson's view that conscious *m* and nonconscious *m* cannot occupy the same spatiotemporal region: presumably conscious *m* will include some additional, concurrent neural activity. Suppose nonconscious *m* and its neural correlate occupy the set of spatiotemporal points *S* and conscious *m* occupies the set *S′*; *m*'s c-property will then occupy *S* – *S′*. So when we ask whether the occupant of *S′* is efficacious to an event *e* qua conscious mental event, we are asking whether it is efficacious qua the mental occupant of *S* – *S′* to any property of *e*.[43]

[43] Alternatively, we can individuate the occupant of *S* – *S′* as a separate psychophysical event, which is natural on higher-order theory: the event would be both a HOR and some kind of neural activation. An independent qua problem would then arise for the occupant of *S* – *S′*: If the event is efficacious, is it so in virtue of being a HOR? I discuss the qua problem Davidson's metaphysics arguably faces in sect. 5.

4. Content Externalism

Externalism about content poses one of the more vexing challenges to a theory of mental causation. It is often held that any qualitative properties a mental state may have are intrinsic to the state or at least to the mind, meaning that a complete description of those properties would involve no reference to anything apart from that state or that mind. But a mental state's intentional properties are arguably relational. For example, a perception's being *of* a stoplight seems to depend on the existence of certain conditions in the world (i.e., a stoplight) standing in certain relations with the perception (e.g., causing it under normal conditions). Similarly, what a thought is about seems to depend on the mind-independent nature of the thing it picks out, a view motivated by thought experiments such as Putnam's (1975) Twin Earth case. And if intentional properties are constituted by external conditions or relations that extend beyond the mind/brain, how can they be efficacious within the mind/brain? Any properties that are efficacious in this regard, it seems, must be "local": Suppose mental event *c* causes event *e*, and *c*'s semantic identity (property *S*) depends on its relation to external condition *E*. If that relation to *E* did not obtain or were altered, *S* would be different. But because *c* would be the same intrinsically, *e* would still result (insofar as an event's causal powers are exclusively a function of its intrinsic properties[44]). So only *c*'s *syntactic* properties – the "vehicle" of representation – have causal relevance, or so the argument goes. (To cite a common illustration, a quarter represents a certain monetary amount in virtue of conditions external to that object that include everything from

[44] As argued in sect. 2, the exercise of a given causal power may well depend on certain conditions apart from the event, and thus will depend on the event's *relational* property of occurring along with those conditions. But for a mental event's power to cause another mental event, likely these enabling conditions would be local to the system, i.e., the mind. And *E*, the semantically relevant condition being discussed, presumably obtains *outside* the mind. In any case, *c*'s property of occurring along with *E* – and thus bearing semantic value *S* – *need* not make any difference to its ability to cause *e*.

beliefs to government documents, but its effects on the machinery of a washing machine depend on its shape and weight, properties that could easily have "meant" a very different monetary amount.)

Several solutions to this problem have been proposed,[45] but I will not be assessing them here. Instead I will focus on what the problem entails for the c-property, specifically given the higher-order view that I favor. Assuming a first-order view of the property, much has been written on the plausibility of reducing it to wide representational content. If that view, known as *phenomenal externalism*, is true,[46] the property would be subject to an externalist challenge to its efficacy. If, on the other hand, one of the intrinsic views discussed in sect. 2 is true, the c-property would be immune to the problem. How do things stand if the property is a kind of *internally directed* intentionality, per the higher-order approach? Prima facie, its content is then also unproblematically efficacious within the mind/brain. As discussed, the c-property of a state *m* on the higher-order view is efficacious only if a state *m** is efficacious qua its h-property – qua representing *m*. Whether *m** represents conceptually or quasi-perceptually, it seems exempt from the problem of externalism since it targets an internal condition, i.e., *m*. So if a first-order mental state's intentional properties are causally problematic because they supervene on a relation(s) that extends beyond the subject, that relation seems entirely *inside* the subject in the case of a HOR, entailing that it can affect the mind/brain.

---

[45] In sketch, some of these approaches include: (1) *c*'s intrinsic nature (what an Earthian and Twin-Earthian in *c* share even though *E* differs in their respective worlds) is plausibly "narrow content," and so some aspect of *S* is causally relevant to *e*. (2) Provided *S* supervenes *in part* on causally efficacious properties, *S* is causally relevant. Insofar as *S* is constituted by the correlation between *c*'s syntactic properties and *E*, *S* supervenes partly on local properties and is thus causally relevant to *e*. (See Segal and Sober [1991].) (3) Assuming that *c*'s representing *E* consists in some causal correlation between *c*-type events and *E*-type conditions, it follows that *E* is part of *c*'s causal history and is thus distally causally relevant to *e* (the causal chain runs $E \rightarrow c \rightarrow e$). The fact that *c* could cause *e* while having been otherwise brought about – entailing a difference in *S* – does not negate the causal relevance of actual-world *S*. (See Heil and Mele [1991].) (4) According to Dretske (1988), behavior is construed as an intention causing an action; thus *c* causing *e* is a unit of behavior. So, for *S* to be relevant to behavior means for *S* to establish the causal connection between *c* and *e*, which is to connect *c*-type events and *e*-type events.

[46] See Lycan (2001) for a discussion and defense, although Lycan, as a higher-order theorist, does not identify consciousness with qualia.

Upon closer inspection, however, the externalist challenge to mental causation besets HORs whether they are thoughts or experiences. On the first theoretical option, *m** picks out *m* in virtue of describing it, e.g., "I am feeling pain." Due to the self-reference, this HOR arguably fails to target an exclusively internal condition, as the self seems constituted in part by relations to the environment. If so, the intentional content of "I" is wide. For example, suppose two molecular duplicates, Bill and Twin Bill, each think "I am feeling pain." They will be referring to different individuals in virtue of the fact that they differ in their extrinsic properties: Bill and Twin Bill stand in different physical relations to their environment. Since this fact makes a difference to the content of "I am feeling pain" despite Bill and Twin Bill being internally the same, the HOT's content is at least partly wide.[47] What about the description of the mental state itself, "pain"? Tyler Burge (1979) has argued that belief content depends on the speech practices of the thinker's community, and it's arguable that HOT content is likewise dependent with regard to the mental-state terms it deploys. Burge's example concerns the term "arthritis": We consider a situation where Fred, who has pain and swelling in his thigh, complains to his doctor "I have arthritis in my thigh." But his belief is false, as "arthritis" is used to refer only to pain in the joints, as his doctor informs him. Consider then a counterfactual situation where Fred also believes "I have arthritis in my thigh," but "arthritis" refers to swelling in *either* the bones or the joints. Here, Fred holds a true belief. Since Fred is intrinsically the same in actual and counterfactual scenarios, Burge claims that his belief content must supervene on facts about his linguistic community, which explains why that content alters in the counterfactual case. Would this kind of argument also show the wideness of a HOT Fred may have, say, that he is feeling depressed? Suppose he is feeling guilty, and misuses "depressed" such that his HOT is false, thinking "I'm feeling depressed." Should we accept the counterintuitive consequences of this

[47] Kim (1998b, p. 203) makes this point (outside the context of HOT theory).

kind of belief content – an introspective one – being determined by speech practices? Those consequences would be that, first, Fred does not know the nature of his own mental state – simply in virtue of misusing words. And second, on HOT theory, Fred becomes consciously depressed, instead of consciously guilty, in virtue of that misuse.[48] The more plausible state of affairs is that Fred is consciously guilty in virtue of believing he feels guilty, but misdescribes his mental state. Despite thinking the *word* "depressed" he deploys a *concept* of feeling guilt. The false belief he has is that "depressed" is the word expressing that concept. So perhaps HOT content does not supervene on the practices of one's speech community, as world-directed intentional content arguably does.[49]

Even so, a HOT's mental-state descriptions can be wide in another way: by deploying concepts whose content depends on the nature of an external referent. Suppose that internal duplicates Oscar and Twin Oscar (from Putnam's thought experiment) each consciously believes that beer is 90 percent water. Per Putnam's point, they would have different belief contents due to the different substances picked out in their respective environments: Oscar would believe that beer is 90 percent H_2O, Twin Oscar that beer is 90 percent XYZ. Since the beliefs are conscious, they each are targeted by the HOT: *I believe that beer is 90 percent water*. Clearly, the content of this belief would vary for the same reason when held by Oscar and when held by Twin Oscar. Oscar would think he has a belief about water, Twin Oscar that he has a belief about XYZ. Thus, when a HOT picks out a mental state in terms of its content – what a belief is about, what a perception is of, etc. – it may refer to an extra-mental object or state of affairs, whose nature then

[48] Assuming that a suitable HOT is sufficient for consciousness. If an actual state of depression is required along with the HOT about being depressed, then Fred, having only the latter, would presumably not be in a conscious state of depression or guilt – which may also be a counterintuitive consequence.

[49] Although we might argue, regarding the arthritis case, that Fred in the actual world does not have a false belief about his thigh: he believes truly that he has pain and swelling in his thigh. His false belief is merely that "arthritis" is the word for such a condition.

(partially) determines the HOT content. If so, the HOT's representational content is not wholly within the mind/brain, and the externalist challenge to the internal efficacy of that content applies.

On HOP theory, *m** picks out *m* in virtue of representing it in a quasi-perceptual way. The causal-correlational approach to representation is perhaps most suited to this theory, and it may seem to avoid the semantic externalism that affects some of the concepts deployed in HOTs. But it turns out, I argue, that if the target state's content is externally determined, that externalism transfers to the HOP, calling into question the efficacy of its representational content as well. Suppose perception-like state *m** picks out *m* in virtue of being the state that *m* normally causes, or the state that is normally correlated with *m*, under certain cognitive conditions (including, e.g., the normal operation of the "internal scanner" and *m*'s accessibility to that system). Suppose *m* is a visual perception of rain. What makes *m** have the content *I am seeing rain* is its being correlated in a certain way to *m*, namely, to a state that is correlated in a certain way to rain (e.g., "tracking" that state of affairs) The externally determined semantic identity of *m* thus partly comprises the semantic identity of *m**. Without *m*'s external correlation, *m* wouldn't be the intentional state that it is (a rain-perception), *m** wouldn't be correlated with a rain-perception, and so *m** wouldn't be *about* a rain-perception. And if *m*'s intentional content is inefficacious in the mind/brain because of its external nature, then the intentional content of *m** will be partly inefficacious as well: "partly" because it is constituted by both its correlation to *m* (internal) and *m*'s correlation to rain (external). Alternatively, suppose *m* is a propositional attitude, such as the belief that beer is 90 percent water. The point is the same, given the externalist view: What makes *m** have the content that one believes beer is 90 percent water is its being causally correlated to *m*, namely, to a belief state whose content is determined by the

mind-independent nature of water.[50] So again, the FOR's externalized content permeates the HOR's content.

In sum, the c-property is exempt from the problem of externalism only if it is intrinsic to the mind/brain, as for example, unreduced phenomenal character, or narrow representational content. The higher-order approach, I have argued, is not a way to reduce consciousness to representation while avoiding the causal problem posed by the externalist.

5. Anomalism

Davidson saw anomalism – or the lack of law-like character – as a prima facie problem for mental causation. An event sequence *c*/*e* is causal, for Davidson, only if it is covered by a (strict) law that says *c*, under certain conditions, is sufficient for *e*. But sequences in which mental events figure, whether purely psychological or psychophysical, at best instantiate nonstrict regularities of the sort given by folk psychology: For example, under the condition that an agent *S* wants it to be the case that *P*, if *S* realizes that doing *x* will make it the case that *P*, then *S* will want to do *x*. While such a conditional generally holds, it clearly can have exceptions: Suppose *S* wants to avoid the effort involved in doing *x* more than *S* wants it to be the case that *P*; *S* will then not likely want to do *x*. Psychophysical anomalism in particular derives additional support from the rationality of the mental: Whether *S* believes that *P*, for example, will be determined (non-strictly) by *S*'s other propositional attitudes and the rational principles *S* observes. So if *S* believes that $P \vee Q$ and that $\neg Q$, then *S* will tend to believe that *P*. But if we

[50] Of course, if the target state's content is narrow (as, perhaps, in the case of a mood or emotion), then *m**'s semantic identity would be entirely constituted by a relation to an internal condition.

assume that neural event *n* causes *S*'s belief under a strict psychophysical law, we have potentially conflicting principles (rational and electrochemical) governing the occurrence of *S*'s belief. And since only the physical principles are true laws that substantiate causal relations, the rationality of the mind would be compromised, Davidson argues.

Davidson is concerned with propositional attitudes, not qualitative mental states like perceptions. But clearly, beliefs and desires can be conscious, and instantiations of that property may also fail to exhibit a nomological character. That would create a problem for the property's causal relevance, given Davidson's assumption about the law-governed nature of causality.[51] In what follows I argue that the c-property also appears anomalistic, though in a different way than propositional attitudes. The issue is not that instantiations of the property cannot be subsumed under strict laws, or that such instantiations exemplify rational principles that physical causation would compromise. Rather, the issue is that regularities involving the c-property (exceptionless or not) are difficult to isolate, due to its prevalence and uniform introspective appearance across instantiations. In this case, an introspective approach to the property – as opposed to a theoretical one like higher-order representation – is justified since it's our primary empirical source for ascertaining the instantiation patterns of consciousness as a *mental* property, not a neural one.[52] Similarly, the regularities involving propositional attitudes that Davidson claims can't conform to strict laws are ascertained by introspecting our own thought patterns, e.g., that we tend to become disappointed if we want it to be the case that *P* and we learn that not-*P*.

My argument for the anomalistic character of the c-property is premised on two claims, which I argue for below.

---

[51] It's fair to note that Davidson did not favor the notion of event properties having causal relevance, i.e., relations of the form *c* qua *F* caused *e* qua *G*. Indeed, a major objection to his Anomalous Monism as a model of mental causation is that it secures the efficacy of mental events, but not of mental properties.

[52] Of course, we may also ascertain mental regularities through what others communicate about their thought patterns. But that information will be based on their own introspection.

(i) *A mental state's c-property is introspectively distinguishable from the state's content.* In other words, we can distinguish what we are aware of from our awareness of it. When I consciously gaze at a painting, I am aware of shapes, colors, degrees of light and shade (and perhaps, if the painting is not terribly abstract, I am aware of these qualities *as* recognizable objects: clouds, hillsides, etc.). What happens when I introspect this conscious perception? Surely, I retain awareness of the same visual properties of the painting. This follows from what is known in the literature as the Transparency Thesis, according to which introspection "sees right through" any phenomenal mental state to properties of the represented object.[53] But while the visual properties are not experienced as mental, I am still able to judge, upon introspecting, that I am in a certain perceptual state. My claim is that there is a phenomenological datum *apart* from the shapes, colors, etc., that grounds this judgment. The datum that emerges upon introspection is a sense that these qualities are objects of my experience, a sense of being presented with them, of mental directedness toward them. For the visual properties could in principle comprise the content of *anyone's* visual experience; I introspect them as the content of *my* experience. This claim is thus at odds with the strong construal of the Transparency Thesis, which says there are *no* introspectible aspects of an experience apart from properties of the represented object. So when one tries to become aware of perceptual content as *mental* content, e.g., to focus on a stop sign's redness as mental red, there *is* something phenomenologically new that surfaces, namely, the for-me component of what-it's-like-for-me-to-see-that-redness. This sense of mental directedness toward the content is, I think, the closest we can get to introspectively isolating the c-property of

[53] See, for example, Gilbert Harman (1997): "When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience" (p. 667).

a mental state. That's how the property introspectively appears, even if it is actually constituted by higher-order representation, for example.

Moreover, our ability to do this explains how we are able to recognize the same property (i.e., consciousness) across a great variety of mental states, from perceptions of flashing lights to itches. Our ability to introspectively abstract a phenomenon of *being appeared to* from a conscious mental state also allows us to conceive of that state occurring nonconsciously. Consider one's consciously hearing the noise of the car engine while driving: To think of one's hearing that noise nonconsciously is to think of one's cognitively processing that very aural content *minus some aspect of the experience* – which from the first-person perspective can best be described as a sense of being presented with that content.[54] Now as explained, Davidson focuses his argument for anomalism on propositional attitudes, and we can't plausibly introspect a phenomenon of *being appeared to* from, say, a conscious belief in the law of noncontradiction. But we can introspectively distinguish a sense of subjectivity: *being in* that mental state.

Authors such as Thomas Metzinger and Kriegel have countenanced this kind of phenomenon. Metzinger's phenomenal model of the intentionality relation (PMIR) asserts that the content of consciousness is more than the object of consciousness; it is the "subject-object relation." "The content of consciousness never is a mere object," he writes, "it is always a *relation*. Phenomenologically, a PMIR typically creates the experience of a self in the act of knowing, of a self in the act of perceiving …" (2006, pp. 23-24). For Kriegel (2009), phenomenal character is also a relation – or rather a combination – of awareness of the external

[54] Thus, the nonconscious version is understood to contain *the same* content without apperception. Admittedly, this scenario can be difficult to conceive when it comes to qualitative content like a noise – difficult, but not impossible. As Norton Nelkin (1995) argues, since the only qualitative states we introspect are conscious ones, the c-property and the qualitative content can seem inseparable. He writes, "We are tempted to ask, 'What would it be to "feel" an experience that is not apperceived?' But the answer to that question is '*Exactly* what it feels like to experience one that *is* apperceived'" (p. 375).

object and awareness of experiencing the object: Thus, what-it-is-like-for-me (e.g., to see a waterfall) combines qualitative character (what-it-is-like: shimmering, whitish, etc.) with subjective character (for-me-ness). But he adds that one is only *peripherally* aware of subjective character. Unlike Metzinger and Kriegel, I don't think that subjective character need always be part of a conscious experience: perhaps in cases of "immersed" experience, conscious content becomes fully transparent. The watch repairman at work on an expensive piece might be consciously aware of the mechanism's intricacies and the position of his pincer, without any sense (even peripheral) of *his experiencing* these properties. But for the ensuing argument, I only need that the c-property of a conscious experience be introspectively distinguishable, which is a weaker claim.

(ii) *The c-property, from the first-person point of view, is informationally impoverished compared to the content of the state that instantiates it.* By "informationally impoverished" I simply mean there is relatively little about the property to describe. Introspection does not reveal that property as, for instance, a relationship between one's mental state and an internal monitoring system, or as the "global availability" of the state. What it does reveal is much about the *content* of the conscious state, per the Transparency Thesis. If I am consciously viewing an apple while holding it, I can describe a variety of features of the apple based on the detailed representational content of my perceptual state, from the roughness of the stem to the greenish splotch near the base. I find it relatively difficult, however, to elaborate on *my experiencing* those features: At least when I introspect my state, this property seems to attend the roughness and thinness of the stem, as well as the greenness and rough circularity of the splotch, but is structurally opaque by comparison. This, I submit, explains the general ineffability of a

conscious qualitative state: it's not that it can't be described, even in great detail, using public-language words like "rough" and "green" (the described qualities would be *mental* roughness and *mental* green, of course). Rather, it's that any such description must also refer to the subjective mode of presentation, the concept of which is nearly primitive. More important for the present argument, however, is the fact that there is no phenomenological variance in the property across the many kinds of mental state that may instantiate it, whether perceptions of flashing lights, itches, or propositional attitudes like mathematical beliefs. For every conscious state that is introspected, its being conscious manifests in the same informationally impoverished way. Thus, instances of the c-property not only appear prevalent to introspection, but also uniform in character.

With (i) and (ii) in mind, we next observe that the relatively well-defined content of mental states falls into regularities more saliently than the c-property of those states. And since regularities comprise our evidence of causal relations, the causal efficacy of a state qua its content is better supported than qua its being conscious. My conscious volition to run is followed by my running, my conscious perception of a red stoplight is followed by my desire to hit the brakes, my conscious thought that it's tax time is followed by thoughts about deductions. The subsequent mental events here pair up with the antecedent events qua their respective volitional, perceptual, and propositional contents, not their being conscious, which appears to be too prevalent and uniform a property to carry any distinctive effects.

Now, mental-event regularities are not exceptionless, and so the above examples should be qualified with "usually," "under normal conditions," or "ceteris paribus." Nor do mental events fall into strict regularities with neural events: If my perception of a red stoplight is usually

followed by my desire to hit the brakes, it will only usually be followed by an instance of whatever neural event type is correlated with that desire. And even if the perception were *invariably* followed by the desire, and thus by instances of a certain neural event type, the instances will not likely be duplicates; surely there are slight variations in how the same desire is subvened in the brain. Yet we *can* expect these variations to correlate strictly with variations in other neural events, such as the one subvening the perception.

In short, mental events are anomalous, as Davidson has argued. And since all bona fide causal relations have a strict "nomological character," it would seem that there can be no psychophysical causation. However, Davidson's monist metaphysics (putatively) allows a causal role for the mental to be preserved in spite of this problem. On that view, any event that can be mentally described can also be (in theory) neurologically described,[55] meaning that the mental event and the neural one are token identical. And it is the extension of those descriptions that enters into causal relations. "Causality and identity are relations between individual events no matter how described," Davidson writes. Laws, on the other hand, are opaque contexts: "[E]vents can instantiate laws, and hence be explained or predicted in the light of laws, only as those events are described in one way or another" (1991, p. 250). So assume an event *c*, as neurologically described, precedes another event *e*, as neurologically described, according to a strict law that supports counterfactuals like "If *e* were slightly different neural event *e´*, then *c* would be slightly different neural event *c´*." As mentally described, however, *c* does not precede *e* (as neurologically described) according to a strict law: If *e´* occurred instead of *e*, *c* would be *the same* mental event (i.e., fall under the same mental description). Yet in virtue of *c*'s neurological description, the *c*/*e* sequence is nomologically strict and qualifies as causal. And

[55] The converse clearly does not hold: an event that can be described as a prefrontal action potential has no description in psychological terms.

since the causal relation is extensional, *c*, as mentally described, causes *e*, as neurologically described. Essentially, a mental event causes *e* even though the mental description is of no use in formulating a strict law involving *e*. Presumably, consciousness can be causally efficacious via the same reasoning: If an event that enters into an exceptionless regularity as neurologically described can also be correctly described as a conscious event, then a conscious event enters into a causal relation.

The typical counterargument to Davidson is that the extensions of mental and neural descriptions of an event are actually distinct *properties* of that event. So if that event can be shown to participate in a strict regularity only as neurologically described, that's because only its neural properties are entering into strict regularities and doing the causal work. In turn, the event's mental properties become causally irrelevant, i.e., the event is not efficacious qua its mental properties, including any c-property. This result follows from the view that causal relations are grounded in strict laws, an assumption that Jerry Fodor (1989), for example, has contested. We arguably need only nonstrict or ceteris paribus laws, of the form *F*-events cause *G*-events when a certain set of conditions *C* holds. The $F \rightarrow G$ law is thus one that can have exceptions, namely when a member of *C* doesn't obtain. If such laws are sufficient to ground causal relations, my perception of a red stoplight can cause my desire to hit the brakes even if the regularity between the two event types has exceptions (e.g., if I happen to know the brakes don't work). Similarly, *c* can cause neural event *e* in virtue of *c*'s mental properties even if it is not the case that the instantiation of those properties by a *c*-event *always* results in an *e*-event; on occasion, the slightly different *e´*-event may follow. The sequence from those mental property types to *e*-events is still a rough or nonstrict psychophysical regularity that can substantiate a causal relation.

We might try to extend this approach to consciousness by seeking a nonstrict correlation between a mental state's being conscious and some type of subsequent mental or neural event. But since consciousness ranges across so many kinds of event sequence – from desires to brake succeeding red-stoplight perceptions, to feelings of annoyance succeeding itches – it is more difficult to isolate a regularity, imperfect or not, that subsumes the phenomenon and thereby supports a causal role for it. Of course, if mental events of type *A* regularly precede desires to drink, and mental events of type *A* are always or usually conscious, then conscious *A*s regularly precede desires to drink. But it does not follow that conscious mental events *simpliciter* regularly precede desires to drink; they in fact precede a great variety of mental events, neural events, and behaviors, while retaining a uniform phenomenal character (i.e., the subjective mode of presentation) throughout. So any type(s) of mental event or behavior consciousness regularly precedes must be significantly broader, ranging over lower-order event types such as thoughts, visualizations, and movements. One such type may be complex mental or physical acts, especially nonhabitual or creative ones. Of course, we also introspect consciousness preceding and accompanying simple, habitual acts. But we often discover that these have occurred nonconsciously, which is seldom the case with nonhabitual acts. So the introspective evidence seems to suggest that consciousness is a necessary condition for the latter, based on regularities such as: Most nonhabitual acts are preceded and accompanied by conscious states (e.g., perceptions and volitions relevant to the act). I do not deny that this kind of rough regularity can be introspectively confirmed and that it suggests a causal role for the c-property. But I think that due to the informationally impoverished appearance of that property, its causal contribution, if any, to such acts is inscrutable from the first-person point of view: It's the volitional content, not the sense of subjectivity, that seems to control the ensuing action.

Two even broader introspectible regularities subsuming consciousness are as follows: (i) Nearly every conscious mental event is followed by another conscious mental event of some kind; indeed, without this regularity there would be no *stream* of consciousness. And (ii), nearly every instance of consciousness is followed by some act of the subject, whether mental or behavioral. Both of these general regularities seem to indicate nonstrict laws: sometimes a conscious mental event *isn't* succeeded by another conscious mental event, e.g., when the subject is suddenly knocked unconscious, and sometimes a conscious mental event isn't succeeded by some act of the subject, e.g., when the subject is electrocuted. So on the (rough) regularity theory of causation, the distinctive effect of consciousness seems to be, in case (i), the instantiation of the c-property by subsequent states, meaning that the phenomenon's causal role would be to propagate its own type; and, in case (ii), agentive acts in general, perhaps as a necessary condition for them. But in case (i), consider a sequence of two conscious mental events *a* and *b*, say where *a* is the desire for a drink and *b* is the volition to reach for one. Phenomenally, there is plausibly a sense of the *a*-content causing the *b*-content, but in contrast, there is *merely* a sense of two instances of the c-property succeeding each other, not of the *a*-content's subjective mode of presentation bringing about the *b*-content qua its subjectivity. And regarding case (ii), I think there is no sense of *b*, as a mental act of the subject, being produced by *a*'s being conscious. That's because the causally relevant information contained in *a* seems to be the propositional attitude (the desire that I have a drink), as opposed to the subjective mode in which that content is presented, which is informationally vacuous by comparison.

The point, then, is that (nonstrict) regularity-based evidence is more readily available for the efficacy of mental states as content-bearing vehicles than as instances of consciousness. Again, the "availability" I am discussing is introspective: From the first-person stance, the

ubiquity and uniformity of the c-property inhibit one from discerning a distinctive regularity it falls into. And for certain general regularities it does appear to participate in, its informational impoverishment hinders an understanding of its causal participation. Thus, we can't easily fit the c-property into the causal workings of the mind/brain from the first-person stance. This result undermines the folk psychological assumption that consciousness *does* appear causally active, and underscores the need to investigate its efficacy on the basis of reductive accounts. As a kind of representational content, for example, the c-property loses much uniformity, since the content would be different for each conscious state, and acquires informational content that can be causally effective. Similarly, as a kind of neural activity, such as a spiking pattern in a certain area, the property would have a causally viable structure. And as a kind of functional property, such as the one postulated by "global access" views, consciousness is constituted by a causal role. But prior to the issue of causal role is, of course, the question whether these are adequate reductions of phenomenal consciousness in the first place.

6. Causal Exclusion

Perhaps the most serious challenge to a theory of mental causation is Kim's (1998) Causal Exclusion Argument (CEA). If sound, CEA proves that all mental phenomena, including consciousness, are causally inert. In this section I will review certain counterarguments to CEA and show that they can secure the efficacy of mental events without establishing the causal relevance of the c-property that some of those events instantiate. First, I will briefly explicate Nonreductive Physicalism, a theory that according to Kim entails the causal exclusion of the

mental by the neural. Nonreductive Physicalism claims that mental events supervene on neural events but remain ontologically distinct from them. Although the notion of supervenience has been variously construed by theorists, one plausible definition of the relation is as follows: Mental event-type *M* supervenes on neural event-type *N* if and only if, for any time *t*, the occurrence of an instance of *N* at *t* is sufficient (but not necessary) for the occurrence of an instance of *M* at *t*. (We can then also say that a mental event token *m* of *M* supervenes on the neural event token *n* of *N* .) So if pain supervenes on C-fiber firing, a particular case of C-fiber firing entails that a particular experience of pain concurrently obtains. Thus the mental is "fixed" by the physical, and physical duplicates will be mental duplicates. Yet the fact that *N*'s instantiation is not necessary for *M*'s means that *M* can be instantiated along with some other (presumably neural) substrate: pain in an octopus, for example, may well be subvened by a different kind of neural activity entirely. Hence a mental event cannot be identical, or reducible, to a neural one.

Apart from preserving the reality of the mental, three theoretical advantages are held to result from this feature: First, it justifies calling the theory physicalism, as the mental is "fixed" by the physical[56]; in other words, physical duplicates are mental duplicates, and mental differences entail physical differences. So for mental event *m* and neural event *n*, if *m* supervenes on *n*, *n* is both nomologically sufficient for *m* and concurrent with *m*.[57] Second, the feature

[56] Parallelism, on the other hand, does not make this claim. Whether the theory is cast in terms of event or property dualism, mental causal chains and neural causal chains run in parallel, and the regular correlation of certain mental event types with certain neural event types is left as a brute fact. The problem for this approach is to give an account of how metaphysically independent mental events can ultimately affect behavior, which seems to have a sufficient cause in the neural. Supervenience theory, by postulating the dependence of mental events on neural ones, both explains the correlations and makes mental events' causal relation to behavior more theoretically viable, as will be discussed.

[57] The theory can be cast in terms of properties instead of events: If *M*, a mental property, supervenes on *N*, a neural property, a subject's instantiating *N* at time *t* nomologically necessitates his instantiating *M* at *t*. This relation is typically referred to in the literature as "same-subject necessitation." Alternatively, the necessitation can be

allows for the multiple realizability of mental states. While *n* is sufficient for *m*, a computer state or alien brain state might also be sufficient. Thus, *n*'s occurrence is not necessary for *m*'s (contra identity theory). Third, and most relevant to the topic of this chapter, the supervenience of *m* on *n*, unlike the causation of *m* by *n*, secures *m*'s causal efficacy – at least prima facie. The reason is found in the metaphysical relation between *m* and *n*, which, although weaker than identity, is stronger than causation.[58] Depending on the version of the theory, *n* is held to *realize m*, where *m* is a functional role; *constitute m*, where *m* is a macrostructural event; or *determine m*, where *m* is a determinable. When *n* bears relations such as these to *m*, it can be argued that *m* causes neural events and behavior along with *n*. For example, Frank Jackson (1996) argues that the constitution relation enables *m* to inherit *n*'s causal powers. "If mental state tokens are constituted by brain state tokens rather than being identical with them, it remains true that mental state tokens are in the brain and that their causal powers are those of the relevant brain state or states" (p. 389).

Yet according to CEA, *m* is preempted from causing any of *n*'s neural or behavioral effects – its supervenience on *n* notwithstanding. The reason is that any such effect – call it *e* – is physical and therefore its sufficient cause must be physical. The more general premise here is the causal closure of the physical domain (CCP): For any physical event *x*, if *y* is part of the sufficient cause of *x*, then *y* is a physical event. So if *m* is to be part of the sufficient cause of *e* along with *n*, *m* must be physical – but according to Nonreductive Physicalism *m* is irreducibly mental. One objection to this argument may go as follows: Granted, *m* cannot be reduced to *n*; yet in virtue of its supervenience on *n*, it is, at a more fundamental ontological level, a nonmental phenomenon. And assuming the thesis that all fundamentally nonmental phenomena are

---

metaphysical, assuming that laws governing properties are essential to them. On that view, for any possible world *w*, *N* obtains in *w* only if *N* nomologically entails *M* in *w* (ceteris paribus).

[58] Indeed, Stephen Yablo has characterized supervenience as "a kind of 'supercausation' which improves on the original in that supercauses act immediately and metaphysically guarantee their supereffects" (1992, p. 257).

physical, *m* counts as physical, which means that CCP is *not* breached should *m* be part of *e*'s sufficient cause along with *n*. In response, we can deploy the following stronger version of CCP, which does entail *m*'s exclusion. CCP*: For any physical event *x*, if *y* is part of the sufficient cause of *x*, then *y* is a thoroughly nonmental event. To say *y* is "thoroughly nonmental" is to say that any phenomenon that *y* supervenes on is nonmental *and y itself is nonmental*. CCP* is plausible in that, presumably, every neural event and every behavioral event can be given a complete causal explanation in terms of phenomena that are thoroughly nonmental, such as neural events, sensory stimuli, etc. So on CCP*, the sufficient cause of *e* can include *n*, since *n* is nonmental and supervenes on molecular events that are also nonmental. But it cannot include *m*, which is only *fundamentally* nonmental, in virtue of supervening on *n*. Though *m* is arguably physical according to Nonreductive Physicalism, it is clearly not nonmental on that theory, if the view is to remain distinct from Reductive Physicalism. Thus, CCP* seems to entail that *m* is preempted from causing *e* by the set of thoroughly nonmental phenomena that are causally sufficient for *e*.

Now it might be that *m* can cause another mental event – *m´* – a scenario that would not breach CCP*. However, this claim is problematic in at least two respects: (i) According to a sub-argument of CEA, the supervenience feature results in *m* and *n* competing for the causation of *m´*. Per Nonreductive Physicalism, *m´* must have a supervenience base, say *n´*, that is metaphysically sufficient for its occurrence. And per CCP*, *n* (or *n* plus other thoroughly nonmental events) is causally sufficient for *n´*. So by the transitivity of the sufficiency relation, *n* is sufficient for *m´*, which seems to exclude *m* as a cause of *m´*. (ii) Even if *m* can be a legitimate cause of *m´* in spite of (i), if the causal chain of mental events that includes *m*→*m´* never includes neural events (i.e., there is no "downward causation," as Kim puts it), then how can

mental events ultimately make a difference to behavior? Behavior is physical and its most proximate sufficient cause is physiological, so in order to affect behavior a mental event must be part of the etiology of neural events, which transgresses CCP*.

Note that problem (ii) also besets Parallelism, a theory that postulates only the regular correlation of certain event types in the mental causal chain with certain event types in the neural one, and it is the supervenience feature that is supposed to make downward causation more theoretically viable. Kim himself has suggested that mental events can satisfy a notion of supervenient or epiphenomenal causation: "When mental event M causes a physical event P, this is so because M is supervenient upon a physical event P* and P* causes P. … Similarly, when mental event M causes another mental event M*, this is so because M supervenes on a physical state P, and similarly M* on P*, and P causes P*." We can expand on this account to claim that a mental event's causing a neural event consists in the fact that the former has a realizer, constitutor, or determiner that causes that the latter. Kim (1991) concedes, however, that insofar as supervenient causation depends upon subvenient causation, it is a lesser grade of causation: "It would be foolish to pretend that the proposed account accords to the mental the full causal potency we accord to fundamental physical processes," he adds (p. 264).

If we do accept supervenient causation as a means to secure the efficacy of mental events, presumably the efficacy of conscious mental events is also secured. Yet it would not follow that said events are efficacious *qua their being conscious*. Let us assume that *n* is the sufficient cause of both *n´* and *m´* (insofar as it causes *n´*, and *n´* is the supervenience base of *m´*). The fact that *m*, a conscious mental event, supervenes on *n* entails that *m* superveniently causes *n´* and *m´*. For *m*'s c-property to be (superveniently) causally relevant to these subsequent events, it must supervene on one of *n*'s properties. More than that, it must supervene on one of *n*'s *causally*

relevant properties. As discussed in sect. 2, not all properties of a cause need be causally relevant to a given effect. Thus, *n* can have properties that are causally irrelevant to *n´*, and hence to *m´*. The c-property of *m* may supervene on one of those properties. In "Mental Causes," Heil and Mele suggest this possibility: Assuming that "the causal clout of a supervenient characteristic resides in whatever realizes that characteristic," they note that even if a mental state M supervenes on a biological condition C that produces behavior B, C will have "a range of features that have no bearing on its behavioral effects," and M's phenomenal features may supervene on those features of C. They write: "M may have characteristics – phenomenal characteristics, for instance – unrelated to its causal role. These might depend on characteristics of C that are themselves causally irrelevant to the production of B" (1991, n. 18). Thus, if the notion of supervenient causation provides a valid solution to the exclusion problem, the solution would be the same for a mental event and its c-property: both would need to supervene on causally efficacious neural phenomena if they are to be efficacious. But a mental event's satisfaction of this criterion would not entail that its c-property does, due to the qua issue outlined in sect. 2. Essentially, where *C* is *m*'s c-property and *m* supervenes on *n*, *m* superveniently causes *n´* qua *C* iff there is some property *F* of *n* such that (i) *n* causes *n´* qua *F* and (ii) *C* supervenes on *F*. On higher-order theories, where the efficacy of a higher-order state *m**'s h-property is essential to the efficacy of the target state *m*'s c-property, the same kind of qua issue arises: Where *H* is *m**'s h-property and *m** supervenes on neural state *n* (distinct from *m*'s subvenor), *m** superveniently causes *n´* qua *H* iff there is some property *F* of *n* such that (i) *n* causes *n´* qua *F* and (ii) *H* supervenes on *F*.

The notion of supervenient causation is, however, a questionable one: why exactly should a mental event cause (even with less "potency," as Kim contends) the effects of its supervenience

base? To be sure, a realizer, constitutor, or determiner necessarily entails that which it realizes, constitutes, or determines, and we might think that the supervening phenomenon is entitled to a causal claim on the effects of its base in virtue of being a necessary condition for that base to obtain. Yet from the fact that an event *c* is causally sufficient for an event *e* and *N* is a necessary condition for *c*, it does not follow that *N* is a plausible cause of *e*.[59] So instead of positing a distinct species of causation for mental events based on the supervenience feature, let us return to Jackson's claim that the phenomenon's causal powers simply *are* those of its base. Suppose *e* is a physical effect of *n*; *n* then has the power to cause *e* (under certain circumstances). For *m*, which supervenes on *n*, to "inherit" *n*'s power to cause *e* means either that *m*'s power to cause *e* is *numerically the same as n*'s power to cause *e*, or that it *duplicates n*'s power to cause *e*. The first construal, I argue, results in *m* being causally irrelevant: presumably *m* is distinct from *n* (as it must be if *m* can exist with a different supervenience base), and for *m* to be causally efficacious as such is for it to have numerically distinct causal powers. Without its own power to cause *e*, *m* is causally irrelevant to the *n*→*e* causal process. The second construal is thus to be preferred if we are arguing that *m* has a causal role in this process: *m*'s power to cause *e* is numerically distinct from *n*'s. And since (following CCP*) *n* is sufficient for *e*, any further causes of *e* entail that *e* is overdetermined. As an additional cause, *m* clearly could not be

[59] To give a common example, a fire is causally sufficient to melt wax, and the fire could not occur without smoke, but the smoke is (quite arguably) not causally relevant to the wax's melting. But consider the version of supervenience theory that is based on functionalism: A mental event is defined by a certain functional role (e.g., an event is a worry just when it tends to distract me, leads to problem-solving efforts, etc.). A mental event *m* will then supervene on a neural event *n* that plays the relevant functional role *f*; *n* will be sufficient, but not necessary, for *f* to be played and thus for *m* to occur. Further, *n* will not be able to have certain effects (call one *e*) without playing *f* and so without *m* occurring. Here Jackson argues that while only *n* is "causally efficacious" in producing *e*, *f* – and thus *m* – should still be accorded a distinct "causal relevance" (1996, p. 398) to *e* because *f* is an "ineliminable part" (p. 395) of the neural causal story. But so is the smoke with regard to the fire's melting the wax, one might say. The difference, Jackson would respond, is that *f* is also "integral" to the neural causation, which according to Jackson means that it is like a program implemented by the neural events. The smoke is plainly nothing of this sort: It is merely nomologically relevant to the fire's effect, not metaphysically relevant in the way Jackson describes. Yet it is ad hoc, I argue, to call this metaphysical relevance *causal* relevance, if *m* does not have a distinct causal power over *e* apart from that of *n*.

necessary for *e* to occur, or *n* wouldn't be sufficient. It would instead be an additional sufficient cause.

Accepting overdetermination is thus another way of preventing *m*'s exclusion from causing *e*, albeit the view may be no more plausible than supervenient causation, for reasons such as the following: (i) If *e* has a sufficient cause in both *m* and *n*, the standard counterfactual analysis of $n \rightarrow e$ would become problematic: if *n* did not occur, *e* presumably still would, in virtue of *m*. (ii) Overdetermination within the mind/brain would not only be widespread and systematic, but also arguably a result of evolutionary design. The latter, notably, is not a feature of other cases where overdetermination seems to occur (e.g., a firing squad's multiple shots simultaneously killing a person), and makes overdetermination in the mind/brain seem especially counterintuitive: Why would the mind/brain have been "engineered" with such causal redundancy?

A case can be made for overdetermination in spite of such objections,[60] but there is another theoretical option that seeks to avoid the causal competition between mental and neural events altogether. The basic idea is that these events are causally active at distinct ontological "levels," just as macrophysical and microphysical events are. For example, if we don't want to be saddled with the view that a macrophysical event like sipping very hot tea is inefficacious (i.e., that its causal powers "drain" to those of the atomic properties of the tea), we hold, first, that the macro-event is a real entity; and second, that it causes a macro-effect: a scalding of the mouth, as opposed to certain atomic-level chemical alterations in the tissue. Similarly, a mental event like fearing a rabid dog would be a cause of one's running away, while the electrochemical activity in one's amygdala would be a cause of the "raw" behavior, or the corresponding events at the neuromuscular level. Yablo's (1992) theory of mental causation, according to which a

[60] See, for example, Ted Sider (2003).

mental event supervenes on a neural one as determinable to determinate, allows for this kind of solution. In regards to the present examples, his point would be that just as the tea's atomic properties are a particular determination of the property *being very hot*, the amygdala activity is a particular determination of the fear: In both cases, the determinate is sufficient but not necessary for the determinable to obtain. Should the amygdala activity occur slightly differently, the fear would still obtain – and so would the running, ceteris paribus. Thus, neither is that specific neural event necessary for the running, meaning that it is not a *cause* of the running, Yablo argues. However, it *is* necessary for – and thus a cause of – the neuromuscular event that subvenes the running, for that event depends upon an equally specific neural cause. The fear, on the other hand, is only commensurate with the running: it lacks the right structure to be a cause of the neuromuscular event. So the fear and the amygdala activity are both efficacious; they just have different effects.

Let us examine what the overdetermination and levels-of-causation approaches imply, respectively, for the efficacy of the c-property. As discussed in sect. 2, on several theories of state consciousness, establishing the efficacy of mental events still leaves us with a qua problem for the c-property, the one exception being a functional theory that says the property is an exercised causal power (what I call an F1 property). Putting that type of theory aside for the moment, we can see that the qua issue on the overdetermination approach arises as follows (let the predicate '*C*' stand for the c-property): conscious mental event *m* causes event *e*, while *m*'s supervenience base, *n*, also causes *e*. But does *m* qua *C* cause *e*, or any event for that matter? If *m* qua *C does* cause some event *x*, that event will be overdetermined by *C* and the property of *n* that *C* supervenes on (call it *F*): Thus, it will be the case that *m* qua *C* and *n* qua *F* each is a sufficient cause of *x*. Alternatively, following Yablo's approach: *m* causes *e´* while *n* causes *e*, which

subvenes *e´* as a determinate. The outstanding question is then whether *m* qua *C* causes *e´*, or any (macro-level) event *x´*. If *m* qua *C* does cause an event *x´*, then the supervenience base of *x´*, micro-level event *x,* will be caused by *n* qua *F*. Mutatis mutandis, the qua issue for state consciousness would be formulated in the same ways when the overdetermination and levels-of-causation approaches are combined, respectively, with higher-order theory: Instead of *m* and its c-property, the above points would apply to a HOR *m** and its h-property.

Moving to the theory of consciousness where the qua issue doesn't arise, suppose *C* is an F1 property. Then, once *m*'s efficacy is established via one of the approaches under consideration (assume), there is no remaining issue of its efficacy qua *C,* as *C is m*'s causing some event *x*. If we allow overdetermination, *m*'s being *C* is an overdetermination of *x,* as *n* would also cause *x*. On Yablo's approach, *m*'s being *C* is the causation of a macro-level event *x´*. Indeed, as an F1 property, *C can't* be instantiated if mental events are causally excluded: The view that consciousness is a mental state's causing some other event is incompatible with the view that neural events preempt mental ones from causing.

7. Conclusion

The focus of this chapter has been whether and how three traditional problems for the causal efficacy of mental events – content externalism, anomalism, and causal exclusion – apply to a property that some mental events instantiate: consciousness. Below I summarize my results.

Content Externalism: Whether this is a problem for the efficacy of consciousness depends on whether the c-property bears any representational content, specifically content that supervenes on

how things are outside the mind/brain. The effort to naturalize consciousness has led to various programs to reduce it to a kind of mental representation. But with the exception of narrow representationalism in the manner of Rey's theory, these programs invite an externalist challenge to the property's efficacy: If its content supervenes on external conditions, how can it affect cognition? Higher-order theories, despite reducing the property to internally directed representation, do not avoid the problem, I've argued. A HOT, due to its use of a self-concept and world-directed concepts to characterize their target mental states, fails to be exclusively about an internal (mental) condition. And if a HOR represents via causal relations to its target state as opposed to description, its semantic identity would depend on that of the state it is related to: If the target's content is externally determined, then the HOR's will (partially) be.

Anomalism: I've argued that while strict laws don't seem to apply to our mental states, even nonstrict laws subsuming the c-property aren't readily apparent due to the variety of mental event sequences that the phenomenon attends. The introspective appearance of the property (I've argued there is such a thing) is thus highly anomalous. It may well not have that character from the third-person perspective – as, for example, a functional property like global access. From that perspective, it would presumably be evident that only certain mental events instantiate consciousness (whereas through introspection all mental states appear conscious), and thereby patterns become discernable: We might see that certain types of mental activity precede, succeed, or accompany the particular mental events that are conscious. Indeed, I will argue in Chapter 5 that the c-property, as higher-order representation, falls into a (nonstrict) correlation with reasoning about one's occurrent mental states.

Causal Exclusion: If all mental events are preempted from causing by their neural subvenors, then no c-property of any mental event is efficacious. But accepting, say, overdetermination or "levels of causation" as a solution to the exclusion problem does not entail that any mental event is efficacious qua being conscious. I've argued that this qua problem arises on various theories of state consciousness, with the exception of those that reduce the property to an exercised causal power of a mental state. On that sort of view, the instantiation of the c-property is actually inconsistent with causal exclusion, insofar as the property *is* the instantiating state's having some effect. Of course, we can still ask regarding a given effect of a conscious mental state, whether it is the type of effect that defines the state's c-property. But conscious mental states necessarily have effects qua being conscious if this view holds.

III. EPIPHENOMENAL CONSCIOUSNESS

1. Introduction

Thomas Henry Huxley's "conscious automaton theory" – the original term for epiphenomenalism – holds that states of consciousness play no causal role in the mechanistic system in which they take place, i.e., the brain. And since an automaton is typically regarded as a nonmental being, those states presumably confer mentality. Thus, the phrase "conscious automaton" implies a conflation of consciousness with mentality. For example, Huxley (1874) writes, "The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working. … Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes." Such a passage implies that volition is a form of consciousness, which in turn means that there cannot be nonconscious volitions. But it is now held plausible that mentality can be, and often is, nonconscious.[61] More specifically, mental *events* (or states) often occur nonconsciously – consciousness being a property that a mental event may or may not instantiate. Consequently, it becomes theoretically possible that the "causally closed system" that excludes consciousness *includes* mental events, as discussed in Chapter 2, section 6. This would not be epiphenomenalism as it is usually understood philosophically: the mental in toto is not an inefficacious byproduct of the neural, but rather

[61] Probably since Freud. The idea of the nonconscious mental, however, arguably originated with Leibniz, who, unlike Descartes and Locke, held that consciousness (or apperception) is not essential to mentality. There can be "petite," or nonconscious, perceptions.

consciousness specifically is. On this scenario, a conscious mental event's c-property could not causally impact the cognitive system in which the event occurs.[62]

While this characterization of epiphenomenal state consciousness seems correct, it should be qualified in view of Shoemaker's (1999) argument about the nature of properties. Properties are distinct from their causal powers, Shoemaker contends, but they metaphysically depend on those powers for their identity. These powers are conditional: To cite one of his examples, property *P* = *being knife-shaped* iff *P* affords certain powers to the object that instantiates it (e.g., cutting bread) when combined with other properties, namely, *being knife-sized* and *being made of steel*. I suggest that the same theory can be applied to properties of states and events, including properties of constitution. A knife-wielding event has the property *being partly constituted by being made of steel*. That property affords the event the power to cause a splitting of bread conditionally, namely, when combined with *being partly constituted by being knife-shaped* and *being partly constituted by being knife-sized.* With regard to consciousness, the c-property counts as a property of constitution if it is intrinsic to the conscious state,[63] and so, per Shoemaker's point, it must be *able* to exercise certain causal powers, presumably when combined with other psychological properties of the conscious state. Thus, its inability to causally impact the mind/brain (if that is so) must be due only to the absence of the relevant circumstances, not to the property's being fundamentally noncausal. An epiphenomenal c-property should thus be

[62] Consciousness would then be epiphenomenal in the "psychological" sense: it would be a byproduct of a process with no causal role in that process. The scenario is not consistent with "philosophical" epiphenomenalism, on which mental phenomena, consciousness included, are byproducts of physical processes. (For more on this distinction, see Daniel Dennett [1991].) Thus, in addressing the question whether consciousness – as a property of a mental state – is epiphenomenal, it is natural to assume that philosophical epiphenomenalism is false: mental states have causal roles, and it is the efficacy of the c-property that is at issue.

[63] As argued in the Ch. 1, sect. 3.

understood as one that is causally irrelevant *as a matter of fact*, since the right conditions never obtain.[64]

Now, as discussed in Chapter 2, sect. 2, the c-property is usually construed as extrinsic in some way, and it may seem that extrinsic or "mere Cambridge" properties[65] can't afford a state causal powers, even conditional ones. Causation seems to be a "local" matter, with a state's causal powers supervening exclusively on its intrinsic properties. If that is so, a c-property that consists in higher-order representation – *being represented by a distinct mental state* – would appear not to be definable in terms of the conditional powers it affords its instantiating state. But it is arguable that the extrinsic properties of a state or event *can* afford it causal powers. Consider the knife-wielding event: qua *being constituted by being made of steel*, … *being knife-sized*, and … *being knife-shaped*, it causes the bread's splitting. Such intrinsic properties not only afford the knife-wielding causal powers or abilities, but they also, we may say, exert the causal force, being in causal contact with the bread. On the other hand, an extrinsic property of the knife-wielding such as *being within range of the bread* (call it *R*)[66] would not plausibly exert causal force upon the bread. But it *is* partly responsible for the event's ability to cause the bread to split, along with the intrinsic properties cited. *R* can thus be defined in terms of conditional powers: it is such that, when combined with certain intrinsic properties of its instantiating event, affords that event the power to cause the bread to split. Accordingly, it is plausible that an extrinsic property of a

[64] The same point applies to a physical epiphenomenon. For example, sewing machines, especially industrial ones, can produce significant vibrations (to the point that vibration-control devices are sometimes used), but these vibrations do not affect the accuracy of the stitching due to the stability of needle, feed, and looper mechanisms. But that is not to say the vibrations couldn't causally affect the stitching under *some* circumstances, e.g., where said mechanisms did not have the proper stability.

[65] Shoemaker uses this phrase (originated by Peter Geach) to designate properties whose change involves no genuine change in the instantiating object, such as "having been slept in by George Washington … being fifty miles south of a burning barn… being such that Jimmy Carter is President of the United States" (1999, p. 254). A mental state's *being represented by a distinct mental state* seems to fall into this category: A change in the HOR, or in the target state's intentional relation to it, would not change the target state intrinsically.

[66] That is, the wielding's spatial relation to the bread is such that the knife comes into contact with it.

mental state is definable in terms of the conditional powers it affords that state. Consider a state's *being represented by a distinct mental state*. That property's identity would depend on the effects that the HOR (qua its h-property) enables the state to have under certain psychological conditions.[67] Whether these conditional powers are ever exercised is another matter, relevant to whether the c-property is epiphenomenal on this theory. The basic premise, derived from a well-known dictum put forth by Samuel Alexander, is that for a property to be real is for it to afford its instantiator causal powers. Indeed, the notion of a noncausal property, one that is causally inert in all possible scenarios, is certainly suspect,[68] even when the phenomenon in question is state consciousness, whose nature is still controversial.

Inasmuch as the mind/brain functions as an integrated collection of processes – from an act of associative thinking to an active neuronal circuit in the early visual system – to say that state consciousness never causally affects the mind/brain is to say that it is causally irrelevant to every mental and neural process in the system. The epiphenomenalist about consciousness would make this claim, but of course the non-epiphenomenalist need not hold the contrary claim, i.e., consciousness is causally relevant to *all* processes. She need only hold the (far more plausible) contradictory claim: There are some processes to which consciousness is causally relevant. Thus, it is theoretically possible that consciousness is epiphenomenal to some processes and not others, meaning that there are processes in which instances of the c-property are causally relevant, and

---

[67] As argued in Ch. 2, sect. 2, the efficacy of HORs qua their h-properties is necessary to the efficacy of the target state's c-property whether we consider the c-property and the h-property *complements* (i.e., constituting a relation between the states), or whether the c-property is merely the target's *accompanying* the HOR's instantiation of the h-property.

[68] I have in mind here nomological possibility. Under natural laws, every actual-world phenomenon in *some* circumstances would afford certain causal powers to its bearer (in the case of properties) or has certain causal powers itself (in the case of events). This claim, a metaphysical truth on Shoemaker's view, certainly seems plausible with regard to physical phenomena, and the physicality of state consciousness is a credible hypothesis. The claim does not entail, however, that for a phenomenon *x* and a given object or system *y*, the power to affect *y* under certain circumstances must be among *x*'s conditional powers. And so while consciousness must have certain conditional powers, the potential to affect the mind/brain need not be among them. But what other kind of thing *could* it affect, if not the mind/brain?

processes in which they fail to be (and, of course, processes in which the property is not instantiated at all). We should be clear, then, on what it means for the c-property to be epiphenomenal to a cognitive process. Here is a first approximation of a definition: Let m_c be a conscious mental event causally involved in process *T* and *m* the nonconscious version of that same event. The c-property of m_c is epiphenomenal to *T* iff that property's being instantiated is not necessary for *T* to occur exactly as it does. We can generalize this definition as follows:

(ε) A property *P* of an event *e* in a causal process *T* is an epiphenomenon of *T* iff *e*'s being *P* is not necessary for *T*'s occurring exactly as it does.

While (ε) may be prima facie reasonable, it proves inadequate upon closer inspection. First, the RHS of the biconditional is *insufficient* for the LHS: *T* can obtain independently of *e*'s being *P* and yet *P* can fail to be an epiphenomenon of *T*, for the simple reason that *P* may not be an *effect* of *T*. Outside the mental sphere, consider a brick's flight as the "process" by which a window is shattered. *Being red*, suppose, is a property of the brick and hence a constituent of the process, but clearly the window would shatter exactly the same way were the brick any other color. So the brick's being red is unnecessary to the occurrence of the process. Yet the color is not an epiphenomenon, because it is not an effect of the brick's flight. In contrast, bodily trembling *is* an effect of the abnormal brain waves constituting an epileptic fit.

The second problem with (ε) is perhaps more interesting. The RHS is arguably *not required* for the LHS: *T*'s occurrence can depend on *e*'s being *P* while *P* is nevertheless epiphenomenal. Suppose that *e*'s being *P* is a nomologically necessary effect of some prior event *a* in *T*; then, if *e* were not *P* (ceteris paribus), *a* would not occur and *T* would not occur as it does

(or at all). Yet it is doubtful that *P*'s mere nomological necessity to *T* establishes its causal role in *T*. To give just one example: Physical laws entail that a train can't run without vibrating to some degree, increasing in temperature, emitting sound waves, etc.; but it is counterintuitive to say these effects play causal roles in the train's running in virtue of being nomologically indispensable. They are plausibly still considered epiphenomena. Thus, it appears (ε) should revised as follows: A property *P* of an event *e* in a causal process *T* is an epiphenomenon of *T* iff *e*'s being *P* is an effect of *T* that plays no causal role in *T*.

My objective in this chapter is, first, to elucidate and further refine (ε). In doing so, I explain the notion of a causal process and what it means, respectively, for an event and an event property to play a causal role in a process. With regard to the c-property specifically, since it is held to be relational by some theorists and intrinsic by others,[69] I also examine how the property on either type of view would fit the description of an epiphenomenon of a cognitive process. Understanding what it is for state consciousness to be epiphenomenal thus comprises the first part of my discussion.

Next, I examine *why* a phenomenon that is nomologically necessary to the process that gives rise to it (i.e., its "base process") does not thereby play a causal role in that process. What is the metaphysical justification for our intuition that the nomological tie is insufficient? With regard to state consciousness, the inquiry is important because we can reasonably suppose that a given instance of the property is produced by mental or neural events according to psychological or physical laws, and so its failure to be instantiated entails, under those same laws, the prevention of its base process. Yet several theorists would resist attributing a causal role to consciousness merely on this basis. Velmans, for example, writes, "Conscious contents that *follow* given forms of information processing cannot be thought of as entering into that

[69] As discussed in Ch. 2, sect. 2.

processing" (1991, sect. 9.1). Block makes a similar claim: "[I]f P-consciousness is … a byproduct of and supervenient on certain kinds of information processing, then P-consciousness in that respect will appear to have no function" (1997a, p. 403). Presumably, these assertions would be maintained even if consciousness *necessarily* followed or arose as a byproduct of certain kinds of information processing. I do not dispute the general claim that nomological necessity is insufficient for causal relevance; indeed, I intend to show why it is insufficient, based on causal theory. But I will also argue that one type of side-effect presents an exception to this tenet, and furthermore that state consciousness may fall into that category.

2. The Nature of Epiphenomena

I will assume the following two properties are mutually exclusive: *being epiphenomenal to process T* and *having a causal role in T*. I take a process to be an event or causal chain/network of events. The concept is clearly integral to epiphenomenalism, for we consider an event *x* epiphenomenal only insofar as it is caused by an event *y* that is also part of a causal chain distinct from $y \rightarrow x$; say $y \rightarrow z$. We can then say that *x* is an epiphenomenon of $y \rightarrow z$ if *x* is not causally relevant to *z*. Thus, while *x* is caused by *y*, *x* is not an epiphenomenon of *y*, but rather of the *process* that subsumes *y*, which becomes the base process of *x*.[70]

[70] To say *x* is an epiphenomenon of *y* is to say it is caused by *y* and lacks a causal role in *y*, which it does. But then *y* must be the sort of thing *x could* coherently play a causal role in; otherwise *x*'s epiphenomenality to *y* becomes trivial. And *x*'s playing a causal role in *y*, as opposed to causing *y*, would require that *y* itself be a causal chain/network. It follows that only a process is ontologically suited to have epiphenomena.

Most cases of epiphenomenality are forking causal chains of this kind,[71] yet I think that more is implied by the notion of epiphenomenality; namely, a disparity in significance between the epiphenomenon and the last event(s) in the base process, which we may call its outcome(s). Specifically, the epiphenomenon is secondary to the outcome with regard to our interests and/or teleological intuitions. This means that, relative to the outcome, we are less interested in the epiphenomenon and/or we feel it is not the result the forgoing events in the process are tending to. For example, while we do not consider a cough or other symptoms of the common cold insignificant, we are generally more interested in the infection of cells in the upper respiratory tract by the rhinovirus, as that informs our attempts at curing and preventing the affliction. And teleologically, we feel that the rhinoviral activity is tending toward that infection, not toward producing coughing or congestion. Note that the Oxford English Dictionary (1993) defines "epiphenomenon" as "a secondary symptom occurring with a disease but not necessarily regarded as its result." While this is given as a medical definition, I think the implication of secondary status is generally carried by the definiendum and its synonyms: a side-effect implies there is a "central effect," a byproduct implies a "main product." Without this disparity in significance we are less inclined to call one of the causal chains epiphenomenal. Consider a chess clock: pressing the button on one's side causes one's clock to stop, the opponent's clock to start, and the button on his side to rise. The button's rising is an effect of the causal chain beginning with the pressing and ending with the clock's starting, and is causally irrelevant to it. But it's counterintuitive to call it "epiphenomenal" since (i) we're just as interested in that result as we are in the starting clock: the opponent needs to be able to do the same to one's clock, and the rising button gives him the means to do that; and (ii) we have no sense that the pressing is

[71] If an epiphenomenon is caused by the last event, or outcome, of a process, no causal fork is entailed. Also, a modification to the schema must be made in the case of an epiphenomenal event *property*, as will be discussed.

tending toward starting the clock as opposed to raising the button. We simply have a forking process with two outcomes.

Such a disparity in significance between outcome and epiphenomenon is (quite arguably) not "given" in nature. So while there are mind-independent features of the world we can cite in defining an epiphenomenon – causal forks, processes, and inefficacy – it also seems there is a subjective aspect to the definition: the lesser status we attribute to an epiphenomenon. Of course, admitting this anthropocentric criterion means that there can be no *objectively determined* answer to the question whether an event is epiphenomenal.[72] And with regard to consciousness, the criterion is not only subjective, but difficult to apply: If consciousness is a causally irrelevant effect of some mental or neural process, is it less significant (in the senses explained) than any behavioral or cognitive outcome(s) of that process? We might, for instance, insist that consciousness has a high intrinsic value to us and is just as significant as those outcomes. Nonetheless, there *is* an objectively determined answer to the question whether an event is causally irrelevant to its base process. So a tractable problem remains: does consciousness play a causal role in the mind/brain?

Let's look at causal-role playing more closely. Consider a basic case where process T = $e_1 \rightarrow e_2 \rightarrow o$ (that is, event e_1 causes event e_2, and e_2 causes outcome event o). For a phenomenon X to *play a causal role in T*, it must be (i) identical with e_1 or e_2; (ii) a property of e_1 that is causally relevant to e_2; or (iii) a property of e_2 that is causally relevant to o. X can also *play a causal role with regard to T*, which is to say X is (iv) an event that is not e_1, e_2, or o, but is a cause of any of

[72] This is not to say there aren't scenarios where we *typically* apply the criterion in a certain way. Consider, for example, the process of men engaging in a building project: One effect of that causal network of events is the building's completion; another is, say, the shadows that the moving men and equipment cast on a wall. We hold the former to be the outcome and the latter the epiphenomenon, since we are typically more interested in the building's completion. But there can always be atypical judgments of significance: Suppose a theater director observes the construction process because she wants to create a backdrop of silhouetted construction activity for a scene in a play. Relative to her interests, the outcome of the process is the shadows cast on the wall, while work's leading to a completed building is the epiphenomenon.

these events; or (v) a property of such an event that is relevant to causing e_1, e_2, or o. An example: Let e_1 = the rotating of a bicycle's pedals, e_2 = the chain's circulating, and o = the back wheel's turning. If X is a pushing of the pedals, X plays a causal role with regard to T, since it is an event that causes e_1; if X is *being constituted by the flexibility of a pair of links*,[73] it plays a causal role in T, since it is a property of e_2 that is causally relevant to o; and so on. Now, assuming X does not fit any of the descriptions (i)-(v), it is not necessarily an epiphenomenon of T: it also must be caused by e_1, e_2, or o. So in the present example, if X is a squeaking noise, it is an epiphenomenon of T if it is an effect of the pedals' rotating, the chain's circulating, or the wheel's turning. It follows from these definitions that "being epiphenomenal to process T" and "having a causal role in/with regard to T" are contrary predicates: They cannot both be true of a phenomenon X, but they can both be false. X falls into this "middle ground" if it is causally unrelated to T. Then, it has no causal role in/with regard to T, nor is it an epiphenomenon of T. To be an epiphenomenon of T, X must be an effect of e_1, e_2, or o, which would mean that it is causally *related* to T.

We can also define these properties relative to a cognitive system S as opposed to a causal process. Let us say a cognitive system is, inter alia, a collection of causal processes. If mental phenomenon X *has a causal role in* S, then there is some process T in S such that X meets one of the criteria (i)-(v) qua T. If X *lacks a causal role in* S, then every process T in S is such that X is either causally unrelated to T or epiphenomenal to T. And if X *is an epiphenomenon of* S, then some set of processes Z in S produces X and (1) X is causally irrelevant to all events in Z; and (2) X is causally unrelated to any process T in S except Z.

Phenomena that have *utility* I define as a subclass of the causal role-players, in the following way: Assume that X plays a causal role in T. If o is *beneficial* to the greater system that

[73] See Ch. 1, sect. 3, for an explanation of properties of constitution.

T is a part of, we can say that *X*'s use is to help produce *o*. Suppose that state consciousness has the effect of interfering with habitual volitional processes: the more states in those processes are conscious, the slower and more inefficiently the processes run. State consciousness would then have a causal role in that regard but not utility, since these effects are not beneficial to the creature. Given this distinction, an anti-epiphenomenalist about consciousness need not argue that it has a utility, but only that it plays a causal role. Accordingly, the epiphenomenalist needs to argue that consciousness fails to play a causal role in or with regard to any process in the mind/brain, not merely that it fails to benefit the creature.

Still more specifically, the epiphenomenalist must argue that consciousness meets criteria (ii) and (iii) below for at least one process *T* in the mind/brain:

(ε´) A phenomenon *X* is an epiphenomenon of a process *T* iff *X* is (i) neither a causally relevant property nor a causally efficient event in *T*; (ii) caused by some event in *T*; and (iii) secondary (in the senses explained) to the outcome(s) of *T*.

Note that (ε´) refines (ε) by accommodating epiphenomenal events as well as event properties, and by adding a criterion (iii) that serves to distinguish a base process with *y* number of outcomes and *x* number of epiphenomena, from a process with *y+x* outcomes. Observe that if we don't include (iii), it follows that for any forking process *F* where no causal chain "feeds back" into *F*, each last event *e* of a chain is both an outcome of *F* and an epiphenomenon of whatever chain *e* branches from. In other words, every outcome, no matter its significance, can be considered epiphenomenal relative to some other branch of the process. So for example, we are committed to the following: Relative to a brain process whose last event is a miniscule rise in

temperature in the motor cortex, a walking movement caused by that process is epiphenomenal, since the movement is caused by the process and has no causal role in it. As discussed, I think such a usage would contravene our semantic intuitions about "epiphenomenon": the importance of the movement and our sense that it is the purpose of the motor potential would disincline us to call it epiphenomenal relative to the heating process. But if (iii) is stipulated away (in order to avoid any subjectivity in the definition), we still have the issue I am primarily concerned with; namely, whether consciousness meets (i) and (ii), and plays any causal role in any other process in the mind/brain that includes *T*.

Furthermore, note that condition (ii) is the source of the claim that an epiphenomenon is indispensable from its base process. For if the events in *T* are causally sufficient, under certain natural laws and circumstances, for *T*'s outcome as well as for *X*, then, if *X* (ceteris paribus) did not occur, neither would that outcome (assuming the events in *T* are also necessary for the outcome). Where *X* = epiphenomenal event *e*, and *T* = event *c* causing event *f*, David Lewis (1993) expresses the scenario as follows:

> Suppose that *e* is an epiphenomenal effect of a genuine cause *c* of an effect *f*. That is, *c* causes first *e* and then *f*, but *e* does not cause *f*. Suppose further that, given the laws and some of the actual circumstances, *c* could not have failed to cause *e*; and that, given the laws and others of the circumstances, *f* could not have been caused otherwise than by *c*. It seems to follow that if the epiphenomenon *e* had not occurred, then its cause *c* would not have occurred and the further effect *f* of that same cause would not have occurred either. (p. 203)

Thus, the base process $c \rightarrow f$ is dependent on the epiphenomenon, and this dependence is grounded in laws and circumstances. We might say *c* is *nomo-circumstantially* sufficient for *e*, and thus *f* is nomo-circumstantially dependent on *e*. Lewis' "problem of epiphenomena," then, is that this dependence seems to give *e* a role in causing *f* under his counterfactual account of causation. I discuss his solution in sect. 4. For now, I simply point out that the scenario seems to entail that the epiphenomenon is *indispensable* from its base process.

But is it really? In particular, why should we observe the ceteris paribus constraint when assessing *e*'s dispensability? It may be, for example, nomo-circumstantially necessary that a moving train cause a moving shadow (the circumstances including daytime, clear skies, etc.). It is implausible, though, to say the shadow's occurrence is thereby indispensable to the train's running, for we could in principle alter the circumstances (e.g., run the train through a tunnel or at night) in such a way that the same process occurs without the epiphenomenon under the same laws.[74] Now, it remains the case under determinism that if the circumstances had been different a change to the laws would be entailed. But the fact that another instance of the process type could occur in different circumstances and not yield the epiphenomenon is arguably enough to establish its dispensability.

A certain kind of case, however, poses a difficulty for this approach. Suppose the circumstances under which the epiphenomenon is generated are necessary to the base process, perhaps insofar as the "circumstances" simply are the intrinsic, causally relevant properties of the events in the process. Natural laws entail that the train can't run without producing a variety of effects in virtue of its physical nature, such as vibrating to some degree, increasing in

[74] We might individuate the process so as to include the circumstances: e.g., the train moving during the daytime, under clear skies etc. The identity of the process would then depend on the circumstances being in place. But without any independent motivation for admitting such a process into our ontology, the inclusion of these relational properties of the moving train as constituents becomes ad hoc.

temperature, emitting sound waves, displacing air molecules, etc. Some of these can be construed as epiphenomenal events (the sound waves), with others being epiphenomenal properties (the heating), but the case's relevant feature is the same: The train's physical nature, which nomologically entails the epiphenomena during the process, cannot be altered without abolishing the process. So any other instance of the process type would still yield instances of the epiphenomenon types, insofar as the same laws hold. Let's say that such epiphenomena, those that arise from conditions essential to their base processes, are *nomological byproducts*, as opposed to mere nomo-circumstantial byproducts such as the train's casting a shadow or reflecting in a car window. Only the former, then, are indispensable to their base processes – and in the nomological sense, not the logical or metaphysical one.

With regard to state consciousness in particular, might it be a nomological byproduct of certain kinds of neural processes? Let's examine a couple of processes often associated with consciousness:

(i) The Willed-Action System (WAS). Any conscious volition to make a certain bodily movement is preceded by a certain readiness potential (RP) and lateralized readiness potential (LRP) occurring in the motor cortex.[75] Call a particular RP e_1 and a particular LRP e_2. The $e_1 \rightarrow e_2$ process then has a certain movement as its outcome and, suppose, also engenders a conscious volition to so move (occurring after e_2 but before the movement) as its epiphenomenon. Now, RPs and LRPs also precede nonconsciously initiated movements, so these kinds of potential need not generate a conscious volition. Perhaps, then, it is the nature of $e_1 \rightarrow e_2$ *in particular* that,

[75] The LRP was found by P. Haggard and M. Eimer (1999) to precede the reported time of the conscious will to move by 50-100 ms. Prior to lateralizing to the hemisphere that will control the movement, the potential occurs on both left and right sides of the motor cortex as the readiness potential (RP) that Libet et al. (1983) showed to occur 350-400 ms prior to the reported time of the conscious will to move. Haggard and Eimer hypothesized that the LRP, unlike the RP, is the cause of a specific movement, not just *a* movement.

together with psycho-neural laws, causally necessitates the conscious volition. This would mean the conscious volition is a nomological byproduct of that process, and indispensable for it. Alternatively, the conscious volition is a *nomo-circumstantial* byproduct of $e_1 \rightarrow e_2$, meaning that, first, conditions obtain in other parts of the WAS – e.g., prefrontal cortex, cingular cortex, basal ganglia – that enable the potentials to produce the conscious volition; and, second, that were these conditions *not* in place, the potentials would yield exactly the same movement nonconsciously. The conscious volition is then a dispensable epiphenomenon, for (in principle) those conditions could be eliminated and $e_1 \rightarrow e_2$ would run in the same way.

(ii) Focal-Attentive Processing (FAP). Distinguished as relatively slow, serial, and limited in informational capacity, FAP is believed to be causally required for a variety of outcomes, including the identification of novel stimuli, optimal learning and memorization, and flexible action planning, which are subserved by various specialized systems in the brain. FAP seems to depend on some mental events – whether perceptions or thoughts – being conscious: "When consciousness is absent, focal-attentive processing is usually absent," writes Velmans (1991, sect. 7). Yet, as Velmans points out, FAP begins before the onset of consciousness, which is in fact an *effect* of that processing: "The processes which enable information to be integrated into a particular conscious state *also* enable that information to be broadcast to other parts of the system. Consciousness *results* from such focal-attentive processing but does not *enter into* [i.e., causally influence] it." This claim presumably means that consciousness is an epiphenomenon of FAP.[76] However, given the way the brain is built, "the disruption of consciousness is also likely

[76] And of the cognitive system as a whole, given his claim that neither is consciousness causally involved in any *subsequent* information processing Velmans emphasizes, I should note, that consciousness is epiphenomenal only from the third-person perspective. "[F]rom a first-person perspective, things look very different … consciousness

to interfere with at least some aspects of (normal) focal-attentive processing." For example, blindsighted patients, who lack phenomenal consciousness of a part of their visual field, can identify properties of stimuli in that field on forced-guess trials, but cannot, it seems, voluntarily initiate responses toward such stimuli[77] – suggesting that phenomenal consciousness is nomologically linked to the transmission of visual data to modules subserving voluntary control.[78] Now, if consciousness is a nomological byproduct of (normal) FAP, it is causally necessitated by FAP plus psycho-neural laws. On the other hand, if it is a nomo-circumstantial byproduct of FAP, then it is causally necessitated by FAP, psycho-neural laws, and conditions unessential to FAP, perhaps other processes running in parallel to it. In the latter case (but not in the former), consciousness is dispensable from FAP, as the elimination of the relevant circumstances would prevent the states from becoming conscious and FAP would still run normally.

Until we know more about the neural conditions sufficient to generate conscious states, we cannot determine whether they are nomological or nomo-circumstantial byproducts of the base processes cited in (i) and (ii). So we cannot at present rule out the former, i.e., the possibility that consciousness is an indispensable epiphenomenon of FAP or certain kinds of motor potential, in the way that, say, the production of steam is ineliminable from cooking processes that include boiling water. Which means that *defining* epiphenomenal consciousness as an eliminable phenomenon is questionable. Furthermore, there is an important difference

---

appears to exert a *central* influence on human affairs," he writes. And on his "complementarity" view, both perspectives are equally valid.

[77] See, for example, A.J. Marcel (1986).

[78] Velmans' thesis, as he sometimes states it, is that consciousness is not necessary to *any* form of information processing – probably because he is arguing for epiphenomenality, and equates dispensability with epiphenomenality. But, as I have been arguing, epiphenomena can be necessary to their base processes. So consciousness can be an indispensable effect of normal focal-attentive processing and still lack a causal role in that processing.

between a nomo-circumstantial byproduct in the case of a complex, integrated system like the brain, and one that occurs in a nonintegrated system or situation. Consider the reflection of the moving train in the car window: this epiphenomenon is dispensable since we can simply remove a circumstance that is a necessary condition for it – i.e., the car – and the train would run as before. Similarly, it may be that the brain can be re-engineered so that a neural circumstance necessary for a particular FAP to yield consciousness does not obtain, and that FAP runs just the same. But this re-engineering may well disrupt some other process that *does* depend on that neural circumstance. If that is so, the instance of state consciousness that is dispensable from that FAP will not be dispensable *from the system*. And the theorist who casts epiphenomenal consciousness in terms of dispensability will of course hold that it is eliminable while preserving the functioning of *all* cognitive processes. This scenario, first, may entail the transgression of natural laws; and, second, is an unnecessarily strong claim for the epiphenomenalist. As discussed, the epiphenomenalist need only argue that state consciousness *fails to play a causal role* in any process in the mind/brain, and that it meet the more specific criteria given by (ε′) for at least one cognitive process. And the causal inefficacy of a side-effect is arguably consistent with its indispensability. In sect. 4, I examine why this is so from the perspective of causal theory; at this point, I refer to the cases of nomological byproducts as intuitive support: e.g., the sound waves that are nomologically ineliminable the moving train do not cause it to move.

Thus, dispensability is an improper constraint upon a theory of epiphenomenality: A property or event *X* epiphenomenal to a process *T* is not necessarily dispensable. Yet certain accounts of what makes mentality or consciousness epiphenomenal suggest that it *is* a proper constraint. Let us first note that Huxley (1874) gives empirical support for the epiphenomenality of consciousness (i.e., mentality, on his understanding) by citing cases where it is apparently

unnecessary to behavior: A frog with a partial leucotomy that, Huxley supposed, rendered the animal unconscious was still able to walk, hop, swim and perform other reflex-like actions. In another example, Dr. Mesnet observes a French soldier with brain damage to occasionally fall into a trance-like state during which he could still carry out many complex actions. We can conclude from this data[79] that consciousness is epiphenomenal, but to offer these cases as representative of epiphenomenal consciousness also suggests that dispensability (with no behavioral or cognitive deficits) *defines*, or is necessary to, epiphenomenality – when in fact there are indispensable epiphenomena. The suggestion becomes explicit in Flanagan's (1997) remark that an epiphenomenal consciousness would be a "dispensable cog in the machine." The machine could very well require an epiphenomenal consciousness for the operation of certain processes, just as a car engine requires fumes to be exhausted for ideal levels of horsepower.

Indispensable epiphenomena, I would add, can also be said to "make a difference" to their base processes, depending on what we take the locution to mean. Note, for example, William Robinson's (2012) claim that "the leading anti-epiphenomenalist intuition" is that "the mental *makes a difference* to the physical, i.e., that it leads to behavior that would not have happened in absence of the mental." On Robinson's construal, then, to say a phenomenon makes a difference to behavior is just to say it is a necessary condition of a certain behavior. But again, this criterion of non-epiphenomenality alone does not secure a causal role for the mental, as a mental event can be a necessary condition of certain behavior *and* an epiphenomenon of the neural process that did cause the behavior – in virtue of being a nomological byproduct of that process. Given these considerations, we should require a non-epiphenomenal consciousness to have a *causal role* in the system rather than merely to be ineliminable from cognitive processing.

[79] The premises aren't entirely empirical, of course. It is theorized that the leucotomized animal and the soldier in the "trance" are not conscious of their actions.

3. The C-Property and Epiphenomenal Events

As discussed, definition (ε′) accommodates both events and event properties as epiphenomena. So an epiphenomenon of a causal process can be (i) an event produced by the process that is not a cause of the outcome; or (ii) a property of one of the events in the process (whose instantiation was caused by earlier events in the process) that is not causally relevant to the outcome. Naturally, an epiphenomenal c-property would conform to scenario (ii). But where the property is the state's relation to some other mental event, the property's causal relevance to the outcome will depend on whether that event is a cause of the outcome. Following principle Q,[80] if an event *x* has an effect *o* qua being related to another event *y*, then *y* must be a cause of *o* (qua being related to *x* in a complementary way). So if *y* is inefficacious – if it is an epiphenomenon of the process that yields *o* – then *x* does not cause *o* qua being related to *y*. Scenario (i), which concerns epiphenomenal events, can therefore be relevant to whether the c-property is epiphenomenal.

Suppose the c-property is what I called an F1 property in Chapter 2: a mental event *m*'s causing some other mental event(s) *e*. Suppose further that *m* is part of a causal process that yields outcome *o*. Here, an event *e* that is an effect of *m* and causes *o* will cause *o* qua *being an effect of m* iff *m* causes *o* qua *being the cause of e*. Perhaps we have a causal chain *m*→*e*→*o*, or perhaps *e* is a cause of other events that enable *m* to cause *o*. (Clearly, *e*'s occurrence depends on *m*, and in that respect it causes *o* partly in virtue of *being the effect of m*.) In any case, if *e* were not a cause of *o*, then *m*'s c-property – its being the cause of *e* – would make no difference to its

[80] See Ch. 2, sect. 2.

ability to cause *o*. Scenario (i) is relevant in a similar way if *m*'s c-property is its *being represented by a HOR*, which may be merely *m*'s accompanying the HOR's instantiation of the h-property, as discussed in Chapter 2, sect. 2. If that c-property is causally relevant to *o*, then the HOR (qua its h-property) must cause *o* in virtue of accompanying *m*. This is to say both *m* and the HOR are necessary to *o*'s occurrence. It follows that *m*'s c-property will not be causally relevant to the outcome if the HOR is not a cause of *o*: If the HOR is brought about by the process[81] but is not a cause of *o*, then it cannot be a cause of *o* qua its h-property. And so *m*'s accompanying that HOR will make no difference to *m*'s ability to cause *o*.

I should add that there is a third description of an epiphenomenon to which the c-property can conform: In scenario (i), any property *F* of epiphenomenal event *e* is surely also epiphenomenal to the process in question, assuming that the process caused *e* qua *F*. A causally relevant property is a property of a cause, so if *F* is causally relevant to the outcome, then *e* must be a cause of the outcome; but this result contravenes the assumption that *e* is epiphenomenal. So *F* must be epiphenomenal along with *e*. One situation where the c-property fits this third description is as follows: An epiphenomenon of the mental process of answering a test question might be anxiety (assume a low-level anxiety that does not affect the correctness or timing of the answer). The anxiety is naturally construed as an ongoing mental event (perhaps mental state) distinct from the series of cogitations that gives rise to it.[82] So the c-property of that event (if it is *conscious* anxiety) would also be epiphenomenal to the process. A related scenario is based on

[81] Oftentimes, the HOR may be brought about by nonconscious states in the process. The empirical literature on voluntary movement suggests one example of this possibility. Based on Haggard and Eimer's (1999) result, we can plausibly suppose that (i) a nonconscious volition to move in a certain way supervenes on the LRP for that movement; (ii) the nonconscious volition has a causal role in producing the movement; (iii) the nonconscious volition is a cause of a suitable HOR coming to accompany it, rendering it conscious. As Rosenthal (2008) notes, "The best interpretation of the Libet-Haggard results involves a lag between the initial occurrence of the volition and its becoming conscious" (p. 833).

[82] Under certain cognitive conditions including the belief that the test is difficult, the desire to do well, doubts about one's abilities, etc.

the theoretical assumption that mental processes are epiphenomena of neural processes, entailing that each thought, mental image, etc. seemingly involved in arriving at the answer to the test question is epiphenomenal to the neural process that does the causal work. It follows that the c-property of any such thought, mental image, etc. would also be epiphenomenal to that neural process.

4. The Causal Status of Side-Effects

The definitions of causal-role playing and epiphenomenality I have given clearly trade on the notions of causal efficacy and relevance. For example, an epiphenomenon is an event produced by a process that is not *efficacious* qua the outcome, or a property so produced that is not *causally relevant* to the outcome. In this section I consider how various views about the nature of causation might justify a contention implicit in these definitions: A side-effect lacks a role in its base process, despite the fact that the side-effect's occurrence is presumably governed by physical or psychological laws, and is thus nomologically tied to the process's outcome. Now, terms like "side-effect" and "byproduct" do *imply* causal inefficacy/irrelevance, and "epiphenomenon" is even partly *defined* in terms of causal inertness. But since these events and properties can fulfill some of the traditional criteria for efficacy/relevance, it is worth assessing how they come up short. I will argue that David Lewis' counterfactual solution to the problem of epiphenomena – i.e., the problem of distinguishing epiphenomena from genuine causes – is the right approach.

Let us consider schema for two kinds of epiphenomenalist scenario discussed in the last section, spelling out the relations of necessity and sufficiency involved:

(i) Event *e* is causally necessary for outcome *o* and causally sufficient for epiphenomenal event *b*.

(ii) Event e_1 is causally necessary for e_2, which in turn is causally necessary for *o*. In addition, e_1 is causally sufficient for e_2 qua epiphenomenal property *B* of e_2.

Regarding (i), under the laws and circumstances, *b* is necessary for *e* to occur, and hence for *o* (via the transitivity of necessity).[83] Regarding (ii), under the laws and circumstances, e_2's instantiating *B* is necessary for e_1 to occur, and hence for *o*. Such relations ground certain counterfactuals: For example, if *b* did not occur, then *e* would not and neither would *o*.

However, the necessity of *b* and *B* to the outcomes of their respective base processes, it is widely held, is not enough to confer them roles in generating those outcomes. For example, Gabriel Segal and Elliott Sober (1991) assert that "properties of causes that are necessary for their effects need not be efficacious"; so in (ii), *B* is a property of a cause of *o* (e_2) and necessary for *o*, yet this need not entail *B*'s causal relevancy to *o*. Why not? Prima facie, on both intuitive and pragmatic grounds. Intuitively, the symptoms of a disease, though nomologically tied to its progress, do not "make" those stages happen; similarly, a fired gun's heating up doesn't propel the bullet through the barrel. Pragmatically, for most any process, we greatly multiply the number of causes of the outcome by admitting the process's many nomologically necessary

[83] If the event leading to the outcome is not necessary but only sufficient (or sufficient together with other events/conditions), a necessary effect of that event will of course not be necessary for the outcome via transitivity. But in that case, the epiphenomenon functions like an INUS condition (see n. 84) for the outcome: Even if it is not a necessary *part* of the event(s) sufficient for the outcome, as it occurs later, it is still necessary *to* the sufficient condition's occurrence.

effects as causes: Is it useful to have such a profusion of "causes"? But the most important reasons for the inefficacy of side-effects, those having to do strictly with the metaphysics of causation, are perhaps not immediately apparent, since these events fulfill major criteria for causal potency: First, they are at least INUS conditions for the outcomes of their base processes,[84] since they are necessary for those outcomes to obtain. Second, they may meet the criterion of temporal priority to the outcome that is often added to accounts based on necessary and sufficient conditions. Third, side-effects fall into law-like regularities with the outcomes.[85] Consider scenario (i): the necessity of *e* to *o* may be explained by a law on which all *o*-type events are preceded by *e*-type events, and the sufficiency of *e* for *b* may hold in virtue of a law that all *e*-type events are followed by *b*-type events. And if *b*-type events precede *o*-type events, we have *b*-type events regularly preceding *o*-type events. For example, during certain historical periods, deaths by musket balls were regularly preceded by musket ball firings; as a result, many epiphenomenal properties and events necessitated by those firings regularly preceded the deaths: musket barrels becoming hot, smoke, loud cracks, etc. And fourth, as will be discussed, the outcomes are (prima facie) counterfactually dependent upon side-effects, which for Lewis entails causal dependence.

One criterion for causal efficacy that byproducts may fail is spatial contiguity with the outcomes of their base processes. On Curt John Ducasse's (1993) view, an event *x* causes an

---

[84] An INUS (insufficient but necessary part of an unnecessary but sufficient) condition is minimally what is meant by "cause," according to John L. Mackie (1965).

[85] As discussed in Ch. 2, sect. 5, finding regularities that associate the c-property with specific mental or behavioral event types is difficult. But even if they *were* found, we would have evidence for nomological relations subsuming consciousness, which is arguably insufficient for the property to play a causal role. For the c-property may fall into a regularity with mental or behavioral event type *E* as a result of both the c-property and *E* issuing from the same cause, with the former being a side-effect. Rosenthal (2008) makes this point with regard to voluntary behaviors: There surely are "some distinctive types of behavior that occur … only when the relevant volitions or desires are conscious. … But that by itself does not show that the consciousness of those motivational states plays any role in making those behaviors possible. … [T]he very factors that result in those behaviors may also cause the volitions and desires to be conscious" (p. 6). Whether such a causal fork entails the inefficacy of the side-effect is, of course, the very issue I am taking up in this section.

event *y* iff *x* is sufficient for *y*; the temporal endpoint of *x* and temporal starting point of *y* are the same; and the spatial cut where *x* ends and the spatial cut where *y* begins are the same. So returning to scenario (i), assume *e* is also sufficient for *o*, and that it shares a temporal point and spatial cut with *o* per Ducasse's definition. Now, byproduct event *b* may share the same temporal point (its endpoint in time may be the same as *o*'s beginning point), and thus be temporally contiguous with *o*. But it seems that if the spatial cut at which *o* begins is shared with *e*, it cannot also be shared with *b*, as *b* is (ex hypothesi) not an event subsumed in *e*.

Yet it is not entirely clear that *b* must fail the spatial contiguity criterion qua *o*, i.e., that it not share the spatial nexus with *o* along with *o*'s sufficient cause, *e*. For example, let *e* = a large rock's rolling downhill occupying region of space *S* (at its furthest downward limit) at time *t*; *o* = a sapling's bending starting within *S* at *t*; and *b* = an unsettling of dust ending at *t*. If the spatial nexus between *e* and *o* consists in a *geometric correspondence* between certain spatial points *e* occupies at *t* and certain ones that *o* occupies at the next instant, then surely the dust particles are precluded from standing in *the same* geometric relation to the points that constitute the start of the sapling's bending. But Ducasse's notion of spatial contiguity is based on the *shared* set of spatial points *S* between *e* and *o*. He writes, "One identical space-time *cut* marks both the end of the cause process and the beginning of the effect process … the cut itself … having no space-time dimension at all" (129). On this construal, it seems several preceding events can share the same spatial nexus with the outcome: after all, if both *e* and *o* can overlap on a dimensionless *S*, why not *e*, *b*, and *o*? Moreover, consider scenario (ii), where the epiphenomenality of an intrinsic property of a cause of *o* is at issue. Suppose the rolling rock has not only unsettled dust, but also slightly increased in temperature as a result of its motion. We would like to cast the warming as causally irrelevant to the sapling's bending, despite its nomological ties to that event. But it

seems the property satisfies Ducasse's spatial contiguity requirement[86]: If the rolling-rock event takes up *S* at *t*, surely the rising of rock's temperature does as well. The main problem with Ducasse's spatiotemporal nexus criterion, however, is not that it (arguably) fails to exclude byproducts as causes. Rather, it is that the requirement excludes less proximate events in a causal chain leading to *o*. Intuitively, we would like to call a brief tremor that caused the rock's descent a cause of the sapling's bending, the tremor being part of the causal history of the latter event. But there is no temporal or spatial nexus between the tremor and the outcome, as the rock's descent intervenes.

Perhaps, then, the fact that the base process is sufficient for the outcome blocks a side-effect from being causally efficacious/relevant. On scenario (i), *b* is necessary for *e* to occur, and hence for *o*. But if *e* is necessary *and sufficient* for *o*,[87] it seems nothing else is needed to produce *o*, including any event that supervenes on *e* or results from *e*. As Yablo formulates the exclusion thesis for events: "If an event *x* is causally sufficient for an event *y*, then no event *x** distinct from *x* is causally relevant to *y*" (1992, p. 247). Alternatively, *e* may be a member of some set of events sufficient for *o* (and for *b*, ex hypothesi). Now, *b* is necessary for that set to obtain, since it is necessary for *e* to obtain. But since the set is sufficient for *o*, it seems *b* must be excluded as a cause of *o*, despite being a necessary condition for *o*. However, the same problem besets this approach as the one based on the causal nexus criterion, namely, unwanted exclusion. Consider the causal chain $e_1 \rightarrow e_2 \rightarrow o$, where e_1 and e_2 are each necessary and sufficient for their respective succeeding event. Since e_2 is sufficient for *o*, e_1 is excluded as a cause of *o*. Yet each is plausibly

[86] And the temporal one, if we suppose the particular temperature increase only lasts until the temporal starting point of the sapling's bending.

[87] Ducasse of course *requires e* to be sufficient if it is to be a cause, and thus if *b* is only a necessary or an INUS condition of the base process it can't be a cause of *o*, either. But this requirement is questionable, as Mackie has argued. And in any case, it may be that the nomological relations between *e*, *b*, and *o* satisfy biconditionals (i.e., "*e* occurs iff *b* occurs," "*o* occurs iff *e* occurs"), entailing that *b* is sufficient for *e*, and hence for *o*, by transitivity of sufficiency.

entitled to be a cause of the outcome; the less proximate e_1 simply "works through" another event to produce *o*.[88] Similarly, a side-effect of either event, say a property *B* of e_2 that e_1 necessarily yields, works through e_1 to produce *o*: It is a necessary condition of e_1's occurrence, and by transitivity of necessity ends up being necessary for *o*. Here, we might claim that what disqualifies *B* as causally relevant is that, unlike e_1, it works through a *prior* event (an epiphenomenal event *b* would be disqualified for the same reason). Thus, the putative causal chain from the byproduct to the outcome contains a "link" that isn't causal, because the effect precedes the cause. This objection is plausible, but there are reasons to resist the claim that a cause *by definition* cannot follow its effect. Lewis gives these: "It rejects *a priori* certain legitimate physical hypotheses that posit backward or simultaneous causation"; and "It trivializes any theory that seeks to define the forward direction of time as the predominant direction of causation" (1993, p. 203).

Another problem with the exclusion approach is that it seems to beg the question: If *b* is necessary for *e* and hence for *o*, to claim that *e* is sufficient for *o*, or part of some set of events and conditions sufficient for *o* that excludes *b*, is just to assume that *b* is a kind of inefficacious necessary condition. For on the usual view, a necessary condition for an event is part of every sufficient condition for it. Perhaps the reason *b* shouldn't be considered part of *o*'s sufficient condition along with *e* is that if there were some way *e* could occur without producing *b*, *e* would still be sufficient for *o* under certain natural laws, which we can designate as set *L*. That "some way" will of course involve changing the laws and/or circumstances that make *b*'s occurrence necessary to *e*'s; call these sets *L** and *C*, respectively. Thus, a certain counterfactual must hold:

[88] Distal causes must be allowed for; indeed, sometimes they are held to be more etiologically fundamental than proximate causes. For example, the causal impotence of the conscious will that seems to follow from (one interpretation of) Libet's result is due to there being a *cause* of the finger movement, a nonconscious volition, that precedes the conscious volition in the chain of events leading to the movement.

If *L** and/or *C* were so different that *e* fails to cause *b*, *e* is still sufficient for *o* under *L*. Following a possible-worlds semantics, we can interpret the counterfactual as follows: There is a possible world whose circumstances and/or laws are different such that *e* fails to cause *b*, and *e* is still sufficient for *o* under actual-world laws. If there is such a world, let us say *b* can be *screened* from the base process. In general, a necessary condition is screenable from a process iff it can be counterfactually eliminated with some adjustment to actual-world laws and circumstances such that the remaining conditions or events are sufficient for the outcome under actual-world laws. The idea is that no screenable necessary conditions for an event's occurrence are part of that event's causal history. Note that many necessary conditions can't be screened: Returning to a previous example, consider that the rate at which the stone rotates in rolling down the hill is a necessary condition for the sapling bending in the way that it did.[89] In each possible world where the stone does not rotate at that rate, the stone's rolling against the sapling is *not* sufficient for that outcome under actual-world laws. In contrast, there is a logically possible world in which the stone does not displace dust particles, or does not produce sound waves, and the stone's rolling against the sapling *is* sufficient for the outcome under actual-world laws.[90] The displacement of dust particles, of course, is a nomo-circumstantial byproduct, while the rock's producing sound waves is a nomological byproduct and properly called "indispensable," as discussed in sect. 2. But since screenability allows counterfactual adjustment to both

[89] Here I mean to describe the effect with enough specificity that the cause is necessary to it. As Davidson (1967) observes, "The fuller we make the description of the effect, the better our chances of demonstrating that the cause (as described) was necessary." Extensionally, on Davidson's view, all causes are necessary for their effects: A short-circuit at a particular place, for example, seems to be merely sufficient for the fire to occur, but "a short-circuit elsewhere could not have caused *this* fire, nor could the overturning of a lighted oil stove," Davidson argues (p. 77).

[90] When explicating causal notions in counterfactual terms, it may seem that we can't coherently talk of what is sufficient in any one world. But I think we can. On the counterfactual approach, the truth conditions for "*e* is causally sufficient for *o* in world *w*" are not given by how things are in *w* exclusively: There must be some possible world where both *e* and *o* occur that is closer to *w* than any world where *e* occurs and *o* does not. Suppose that is so. Then, that relation between possibilia and *w* makes it the case that, *in w*, *e* is causally sufficient for *o*. So we can talk about the sufficiency holding in *w*.

circumstances and laws, either side-effect is screenable from its base process. And this justifies denying them causal roles.

A difficulty may arise for this approach when the laws that govern $e \rightarrow b$ are the same as those that govern $e \rightarrow o$. For then there is no possible world that is nomologically different such that *e* fails to cause *b* but still causes *o* under actual-world laws. For example, suppose consciousness is identical to some neural byproduct event *c* of neural process $n_1 \rightarrow n_2$; specifically, it is caused by n_1, which is necessary and sufficient for n_2. The conjunction *E* of electrochemical laws of synaptic transmission that govern the production of *c* are the same as those that govern $n_1 \rightarrow n_2$. So if *E* is altered or absent in a possible world *w* so that *c* does not arise from n_1 in *w*, then clearly, neither will n_2 arise from n_1 under *E* in *w*. Thus, it seems *b* could only be screened by altering the *circumstances*, which in this case means that there is some possible world where *E* holds but the neural connectivity, synaptic weights, etc., are so different that n_1 and n_2 occur just as they do in the actual world, but *c* fails to occur.

Let us assume, then, that side-effects lack causal roles because they are screenable. A familiar sort of problem emerges: In the causal chain $e_1 \rightarrow e_2 \rightarrow o$, where e_1 and e_2 are each necessary and sufficient for their respective succeeding event, e_1 may well be screenable, yet it is intuitively also a cause of *o*. There is a possible world where e_1 does not occur, but some other preceding event is nomologically necessary and sufficient for e_2. Once e_1 is screened, it is clear that e_2 is still sufficient for *o* under actual-world laws. Here, Lewis' counterfactual approach does justice to distal causes while affording a reason to deny causal efficacy to side-effects.

First, a synopsis of that approach is in order. Lewis analyzes causal dependence between events in terms of counterfactual dependence as follows: Letting "$O(x)$" designate the proposition that event *x* occurs, *o* causally depends on *e* if a certain counterfactual dependence

relation holds between the propositions asserting their occurrence: $\neg O(e) \mathrel{\square\!\!\rightarrow} \neg O(o)$. So in general, proposition *B* counterfactually depends on *A* if $\neg A \mathrel{\square\!\!\rightarrow} \neg B$. (Lewis defines the counterfactual implication operator, "$\mathrel{\square\!\!\rightarrow}$", truth-functionally on the basis of a possible-worlds metaphysics: $A \mathrel{\square\!\!\rightarrow} B$ is true in the actual world if there is a world where proposition *A* holds and *B* holds that is more similar to the actual world than any world in which *A* holds and *B* does not hold.) Now, if *b* is a side-effect of the process $e \rightarrow o$, per scenario (i), we would have, in Lewis' words, "a spurious causal dependence" of *o* on *b*. The solution, he argues, is to deny the counterfactual dependence that grounds that causal dependence, i.e., to deny that $\neg O(b) \mathrel{\square\!\!\rightarrow} \neg O(e)$. Some possible world *w* where $\neg O(b) \wedge O(e)$, Lewis claims, is more similar to the actual world than any world *w** in which $\neg O(b) \wedge \neg O(e)$, for the reason that in *w*, divergence from the actual world occurs later, after *e* has occurred. So "the spatiotemporal region of perfect match" with the actual world is greater in *w* than in *w**. Of course, *w* violates the laws under which *e* is sufficient for *b*, and this is certainly a point of dissimilarity from the actual world. But, Lewis argues, *w** also violates laws of nature insofar as "under determinism *any* divergence, soon or late, requires some violation of the actual laws" (1993, p. 203). So why not prolong the spatiotemporal match? And even if our intuitions tell us that *w** is a better nomological match to the actual world than *w*, it is unclear how much better; what is clear is that *w* is a better factual match.

Thus, the notion of closeness of worlds based on particular facts, in addition to laws of nature, motivates the claim that the closer possible world is the one in which the base process occurs sans the side-effect. And the outcome's counterfactual independence from that side-effect entails, on Lewis' view, its *causal* independence, which means that the side-effect lacks a causal role in the process. The question, then, is whether the efficacy of the distal event e_1 in producing

o in the causal chain $e_1 \rightarrow e_2 \rightarrow o$ is preserved on this approach. Essentially, there must be a possible world where $\neg O(e_1) \wedge \neg O(e_2)$ that is *more* similar to the actual world than any world where $\neg O(e_1) \wedge O(e_2)$. Now, on the epiphenomenalist scenario, the reason for preserving *e* while giving up *b* (i.e., selecting *w* as the closer world) is that *e* precedes *b* and therefore *e*'s occurrence means that *w* matches the actual world in terms of particular facts over a longer time. But on the causal chain scenario, e_2 does not precede e_1, so preserving e_2 while giving up e_1 does not entail that greater similarity: A world where $\neg O(e_1) \wedge O(e_2)$ contrasts with the actual world on particular facts at the time e_1 occurs, as does a world where $\neg O(e_1) \wedge \neg O(e_2)$. And while the factual differences in both worlds entail that they nomologically contrast to the actual world (per Lewis' point about determinism), intuitively the latter is the better nomological match, conforming to a law that we presumably have evidence for: all e_2-type events are preceded by e_1-type events. So on the balance, the closer world is one in which $\neg O(e_1) \wedge \neg O(e_2)$ is true.

Having established that e_2 counterfactually depends on e_1, we have, on Lewis' view, established that e_2 causally depends on e_1; via the same reasoning, we can establish the next link in the chain, that *o* causally depends on e_2. The two causal dependencies entail that e_1 is a cause of *o* on Lewis' account, causation being a higher-order notion than causal dependence: "One event is a *cause* of another iff there exists a causal chain [i.e., a series of causal dependencies] leading from the first to the second" (1993, p. 200). It follows that the distal cause e_1, unlike a side-effect of e_1 or e_2, has a causal role in producing the outcome. Lewis' analysis, then, yields the intuitively correct result. In the next section, I argue that there is one type of side-effect that nonetheless qualifies as causally potent on that analysis.

5. Side-Effects as Controls

Consider these three side-effects: (a) the erosion of a riverbank, (b) the shifting of a wind vane, (c) the whistling of a steam locomotive. The difference I would like to highlight between (a) and (b) is that the latter is an example of a *designed* side-effect, while (c) is a kind of designed side-effect I will call *systemic*, as both it and its base process are part of a system and designed to work together as they do. Side-effects are designed for various reasons: They may serve as indicators of processes we are concerned with, or they may improve the functioning of their base process, as in the case of an exhaust pipe. But sometimes, albeit more rarely, effects that are necessary to the operation of a process are built in as a means of controlling it. If the process must yield the effect in order to occur, and the effect is blocked, so is the process. Here are two examples of systemic effects that function as controls:

Check valves: Also called “non-return” valves, these mechanisms are installed in a pipeline to allow “forward flow” of water but automatically prevent “reverse flow” into the pipe. Depending on the variety, some type of stopper – a plug, disk, or ball – is lifted off its seat by forward flow pressure, and returned to its seat with pressure from a spring or reverse flow itself, thus closing the valve. The disengaged stopper, then, is a side-effect of the flow used as means of controlling it, i.e., blocking the stream when it becomes too weak to overcome spring pressure or when reverse flow obtains.

Disk brakes: This device slows or stops the rotation of a wheel by applying friction with a brake caliper to a disk connected to the wheel’s axle. Similarly in this case, the spinning disk is a

systemic side-effect of the spinning wheel, and one that is used to control the wheel's rotation and the greater process that is the vehicle's moving.

I argue that this subclass of systemic side-effects qualify as causal-role players in their base processes – in virtue of their nomo-circumstantial necessity to their base process *and the fact that they function as controls*.[91] To see why, let us return to Lewis' grounds for denying a causal role to side-effects: Where *b* is a side-effect of the process $e \rightarrow o$ (specifically, an effect of *e*), *o* counterfactually depends on *e*, but *e* does not counterfactually depend on *b*—the reason being, some possible world *w* where $\neg O(b) \wedge O(e)$ is more similar to the actual world than any world *w** in which $\neg O(b) \wedge \neg O(e)$, as the spatiotemporal match with the actual world is greater in *w* than in *w**, whereas some world where $\neg O(e) \wedge \neg O(o)$ is closer than any world where $\neg O(e)$ and $O(o)$. Now assume that *b* is a controlling effect, e.g., a rising plug in a check valve (with $e \rightarrow o$ being stages in the flow of water), or a disk spinning (with $e \rightarrow o$ being stages in the operation of the moving vehicle). In these cases, I argue, *w* is *not* more similar to the actual world than *w**, because the actual world contains many instances of *b*-type events failing to occur along with *e*-type events failing to occur.[92] These instances, a consequence of the fact that the side-effect is often blocked in order to block the base process, support the claim that some *w**, in which $\neg O(b) \wedge \neg O(e)$, is more similar to the actual world than any *w* in which $\neg O(b) \wedge O(e)$. True, the latter are event-identical to the actual world until the time when $\neg O(b)$. But this minor prolongation in factual match is offset by the uniformity with previous instances that *w** exhibits. Our intuition that *w** is a better nomological match to the actual world than *w* then

[91] Thus I refer to them as a type of side-effect or byproduct, as opposed to an epiphenomenon, which by definition plays no causal role. With regard to (ε′), then, a side-effect would only need to satisfy criteria (ii) and (iii).

[92] These being cases where the *e*-type event would have been produced if the *b*-type event had obtained, i.e., cases of $\neg O(b\text{-type}) \wedge \neg O(e\text{-type})$ within similar systems. They're not merely instances of neither event type being instantiated, which of course are profuse.

becomes the difference-maker in favor of *w** being the closer world. And this result entails that *e* counterfactually depends on controlling effect *b*, which supports a causal dependence.

There is a further argument for the causal efficacy of side-effects that function as controls, based on Georg Henrik von Wright's "manipulative" or "experimentalist" theory of causation. According to von Wright, "What makes *p* a cause-factor relative to the effect-factor *q* is … the fact that by manipulating *p*, i.e., by producing changes in it 'at will' as we say, we could bring about changes in *q*" (1993, p. 118). Manipulability, he adds, applies to cases in which *p* is a sufficient condition for *q* as well as cases where it is necessary for *q*. Thus, preventing *p* to prevent *q* – e.g., pressuring the rotating disk to stop the vehicle – counts as manipulating *p*. The controlling effect, then, is a kind of systemic side-effect that conforms to von Wright's criterion of causal efficacy. Furthermore, its manipulability means that when the side-effect does occur it is "allowed" to occur, so as to allow the base process to occur. This "letting the process happen" isn't what a side-effect like a car's vibration does qua the car's forward motion, and affords intuitive grounds for calling the controlling effect a *cause* of its base process. And from a pragmatic point of view, we do not greatly multiply the number of causes of a process by admitting controlling effects as causes, insofar as they are relatively rare. Usually processes are not controlled via their effects, but rather directly. Most valves, for example, control a flow of water by simply closing the pipe, not by inhibiting some effect of the flow.[93]

In sect. 2, I argued that state consciousness may be a nomological byproduct of certain kinds of neural process, such as focal attention or motor events in the willed-action system, and

[93] Observe that a fuse, i.e., an electrical circuit breaker, is *not* an example of "controlling effect" in the sense I am describing. The base process that sets off a fuse is overcurrent; the device then breaks the circuit to halt that process. This is clearly a different means by which an effect serves to control its base process. Here, the effect is actually a causal chain that "feeds back" into the process, as opposed to an event that can be prevented in order to affect the process. (Note that the actuator mechanism that breaks the circuit could be disabled, but that wouldn't stop the overcurrent. So the mechanism's operation could not serve as a controlling effect.)

that this status is consistent with its epiphenomenality. But if the c-property is a controlling effect, it would be causally relevant to its base process. I shall now try to make the case that state consciousness may be an effect of this kind. Ultimately, the question whether it is must be settled empirically, but here are some general points in favor of the possibility: First, note that state consciousness would qualify as a *systemic* side-effect of its neural base process, insofar as it is an event or property in the brain, which is "designed" in the sense of being the product of natural selection. Of course, the brain is not *literally* designed, and the forgoing examples of controlling effects are based on systems that are so designed: check valves and breaks are conceived and created by sentient beings. But the feature of controlling effects that makes them causally relevant to their base processes is not their occurring in designed systems. Rather, it's that they enable a certain regularity: preventions of instances of the effect type followed by preventions of instances of the base process type. As argued, that controlling role entails the effects count as causes on Lewis' counterfactual theory. It just so happens that the most salient examples of side-effects that have this feature are in mechanisms we design. But naturally occurring systemic effects might also be cited, if they exhibit the feature: The opening of the flaps of the pulmonary valve, for instance, is a side-effect of blood flow from the right ventricle to the pulmonary artery – one that is often prevented by reverse pressure and, in turn, prevents forward flow.[94]

Second, note that brain processes work via complex networks of excitatory and inhibitory connections between neurons and layers of neurons, and preventing a process from occurring involves inhibiting the population of neurons that subserves it (which in turn requires exciting the neurons that have inhibitory connections to that population, and so on). With this picture in

[94] Of course, the side-effect here is not so much used to control forward flow as it is to prevent reverse flow when pressure falls in the right ventricle and forward flow weakens. Nevertheless, the effect *happens* to serve as a means for preventing its base process (forward flow), and thus illustrates how that feature – and not a systemic effect's being designed by us – is what is important to the thesis that there are causally relevant side-effects.

mind, let *p* be an instance of a type of brain process that generates the neural correlate of consciousness (NCC); let event *n* be an instance of the NCC; and assume *n* is a side-effect of *p*.[95] If *p* is to be inhibited (as may benefit the organism in the particular case), this will be accomplished by inhibiting the population of neurons that subserves *p*. And if that population cannot spike without yielding *n*, the neural architecture may be such that control of *p* is achieved by inhibiting *n*. This scenario would make *n* not merely a side-effect of *p*, but a controlling effect that can properly be said to bear causal relevance to *p*. If this seems a roundabout way to control *p*, consider that biological neural networks are not perfectly streamlined systems; for example, there is a great deal of functional redundancy in neural elements.

It may be objected that if *n* is the activity of some separate population of neurons that is excited by *p*, it's unlikely that *p* is a *sufficient* condition for *n*, and that inhibiting *n* prevents *p*. After all, inhibition of a post-synaptic neuron does not prevent the pre-synaptic neuron from firing; the pre-synaptic neuron still releases neurotransmitter, just without the same effect. But here we assume that *n* is the activity of a set of neurons apart from those subserving *p*, as opposed to ancillary activity within the very neurons that subserve *p*. This assumption, I think, is unwarranted. For example, the NCC for the visual experience of motion may be a certain kind of activation in MT/V5,[96] which, on the present hypothesis, is just a byproduct of information-processing events in those areas. But this ancillary activation need not be the spiking of a certain population of "consciousness neurons" in MT/V5. The substrate of consciousness in that region may be a kind of activity within the information-processing neurons themselves. Dendritic activity is a good candidate for an intracellular side-effect of a spiking neuron, one that may

[95] The causal nature of the NCC, I should add, carries implications for the causal nature of state consciousness on at least two defensible views about their relationship: (i) state consciousness is identical to its neural correlate; and (ii) state consciousness supervenes on its neural correlate, in such a way that its causal powers are determined by those of the neural base. See Ch. 2, sect. 6 for a discussion of the latter view.

[96] Or "in and around" this region, including perhaps also V1. See Block (2005).

function as a control. The arbors are the sites of a variety of events beyond the transferring of signals from synapses to axon hillock, such as sub-threshold increases in membrane potential, localized signal integration, and even backpropagation – impulses from the axon hillock back to the dendrites, signaling the output state.[97] If an action potential, or a particular spiking frequency, is sufficient for events of this kind under electrochemical laws, their inhibition would presumably block that process.

So let us suppose the NCC turns out to be an effect of a particular kind of process with no apparent further effects of its own. Flanagan (1997) gives an example with regard to the conscious perception of a red square: After the brain "gets on with whatever it intends to do with the information that a red square is there, it goes into this funny oscillatory state that persists for a few seconds and subserves the experience of seeing a red square. The oscillatory pattern does nothing useful" (p. 358). The fundamental points in such a case are that (i) the NCC – and hence consciousness – is not necessarily dispensable, for it may be nomologically entailed by that process or by the proper functioning of a broader neural system; and (ii) the NCC is not necessarily an epiphenomenon of that process, for it may be a controlling effect and thereby satisfy certain theoretical constraints for causal relevance.

I think that (ii) is a useful point for functional theories that equate the c-property of a state with that state's playing a certain information-processing role. Such theories usually trade on the correlation between conscious states, particularly perceptions, and certain access relations to higher-level systems, arguing that consciousness *is* that access. But empirical investigation may reveal that the NCC is merely an *effect* of that access (hence the correlation) with no apparent further effects of its own. The function of consciousness in that case may still be to further access, if it serves as a control for the process. For example, on Prinz's AIR (Attended

[97] For an overview, see "Information Processing in Dendrites," in M. Zigmond et al. (1991), p. 363.

Intermediate-Level Representation) theory, conscious perceptions are intermediate-level representations available to working memory via attention. Such representations are not so fine-grained and disintegrated as those generated by peripheral receptors (e.g., retinal irradiation changes), nor do they abstract from features like object size and egocentric location, as higher-level representations do. Prinz rightly argues that consciousness is "extremely valuable" on AIR theory, as it broadcasts "viewpoint specific information into working memory. Viewpoint specific representations are important for making certain kinds of decisions. If we encounter a predator, for example, it is useful to know whether it is facing us or facing in another direction." And working memory in turn enables flexible (non-reflexive) responses to our environment. But suppose it turns out that working memory's accessing such a representation is the base process for consciousness, which is therefore a distinct (though causally related) property. In that case, we should further investigate whether that property serves as a control for its base process – whether preventing it from being instantiated is the way the brain inhibits access to working memory. For then consciousness would still be causally relevant to the crucial function given by AIR theory.

Now, we would naturally wonder why phenomenal consciousness should be involved in a control mechanism as described. This sort of concern, however, is not specific to the proposal that a phenomenal property serves as a controlling effect. The functionalist faces a similar challenge with regard to her reductivist proposal: Why should a certain functional property constitute phenomenality? But most functional theories, such as those of Prinz and Tye, reduce consciousness to a kind of functional property *along with* a representational one. For Prinz, a perceptual state is conscious (inter alia) in virtue of representing at the intermediate level; for Tye it is conscious (inter alia) in virtue of representing in a nonconceptual way; that is,

representing features for which the subject need not have matching concepts. The representational criterion is added to account for the content of consciousness: We are aware of things in a viewpoint-specific way, and/or in a way that does not necessarily recruit our concepts for those things. And the posited functional role that (also) makes for consciousness causally involves the representation in a plausible way: The representation is attended and transferred to working memory (on Prinz's theory) or becomes "poised" so that it can impact beliefs and desires (on Tye's theory). So we do have the puzzle of why phenomenal consciousness should equate to a certain representation-plus-functional role, but beyond that, there is no additional question of why the representation should play that role. In that regard, the scenario of consciousness as controlling effect incurs a second theoretical puzzle: Assuming we do reduce phenomenal consciousness to a kind of representation (a promising step toward naturalizing the property), why should a representation – whether intermediate level, nonconceptual, higher order, etc. – be recruited as a controlling effect? A mere subpersonal neural effect with no representational properties could equally serve to block the base process, provided it is a necessary effect. Our theory must therefore attribute a plausible function to the *representation* we take consciousness to be, if that theory is to be commensurate with the explanandum. I put forth my theory of the function of consciousness as representation in Chapter 5.

6. Conclusion

In this chapter I have sought to define the conditions for a state's c-property to count as epiphenomenal. I first defined epiphenomenality relative to a mental process: Essentially, a c-

property is epiphenomenal to a mental process *T* iff it is causally irrelevant to the *T*'s outcome, *T* causes its instantiation, and the property's instantiation is "secondary" to that outcome, relative to our interests and teleological intuitions. I then used this definition to give conditions for a c-property to count as epiphenomenal relative to the mind/brain; namely, it must meet the outlined conditions for some process *T* in the mind/brain, and be causally irrelevant to the outcome of every other mental/neural process. Of course, what we are primarily interested in is not merely whether the c-property is causally relevant, but whether it has utility, which I've defined as being causally relevant to a cognitive outcome that is beneficial to the organism. In Chapter 5, I will argue it has the function of enabling a certain useful metacognitive ability.

Importantly, I've also argued that the c-property's being *necessarily* caused by its base process, and thus being necessary to the outcome, is consistent with its causal irrelevance to that outcome. So its indispensability is consistent with its epiphenomenality, contra what several formulations of epiphenomenality have implied. The property is most properly considered indispensable if it is a nomological byproduct of its base process, as opposed to a mere nomo-circumstantial one. But if the c-property is an indispensable effect of some mental/neural process, what are the grounds for arguing that it lacks a causal role in that process? My position is that genuine causes can be distinguished from such necessary concomitants via Lewis' counterfactual approach: Prima facie, the epiphenomenalist scenario entails that the outcome of the base process counterfactually depends on the instantiation of the c-property by one of the states, and thus causally depends on that property. But it can be justifiably denied, as Lewis shows, that a possible world in which (a) that property were not instantiated and its cause in the process did not occur is more similar to the actual world than a world in which (b) that property were not instantiated and its cause in the process *did* occur. And if that is so, the outcome does

not causally depend on the property's instantiation. The exception, I've argued, would be where the c-property's instantiation is a controlling effect of its base process. In that case, a world in which (a) w*ould* be closer to the actual world than a world in which (b), based on the actual world's (presumably) numerous cases of the property's instantiation being systematically prevented in order to halt the base process.

But the hypothesis that the c-property can be causally relevant as a controlling effect, while a viable theoretical resource, is not an ideal one. The role of controlling effect might be filled, as I've argued, by a subpersonal neural property. And in view of the naturalist program to reduce consciousness to a kind of representation, a theory should posit a function that is more uniquely suited to a representational vehicle.

IV. THE SOURCE AND FUNCTION OF CONTROL PHENOMENOLOGY

1. Introduction

In Hume's search for the empirical source of our idea of a necessary connection between cause and effect, he discredits volition and its mental and behavioral consequences as a viable candidate:

> [I]n contemplating the operations of mind on body – where we observe the motion of the latter to follow upon the volition of the former, [we] are not able to observe or conceive the tie which binds together the motion and volition, or the energy by which the mind produces this effect. The authority of the will over its own faculties and ideas is not a whit more comprehensible. (1988, p. 69)

Thus, Hume allows that we can introspect volitions, which I understand as intentions to act – bodily or mentally – at the present moment.[98] Following our introspection of the intention to produce bodily act *b* now, we might perceive *b*. We might also introspect the volition to produce thought *m* and then introspect *m*'s occurrence. In both cases, we observe the contiguity of volition and act. But, Hume claims, that observation is not accompanied by a sense of the volition's power over its goal state, a psychological "tie" between the two.

[98] For example, my standing intention is to strike the tennis ball when it flies over the net; my volition is to strike the tennis ball *now*. A volition is thus the most proximate intentional cause of an act – or at least the most proximate that is accessible to consciousness. By "volition," then, I mean what John Searle (1983) has described as a "prior intention," except that I take that intention to immediately precede a movement, and specify that the movement is to be performed immediately.

Is Hume's phenomenology on target here? It's questionable. For when a volition-act sequence is observed from the first-person stance, one normally gets a sense of *producing* the act – a qualitative experience of causation over and above the mere observation of the act's coming about after one has willed it. Descartes countenances this experience when he claims that "Everyone *feels* that he is a single person with both body and thought so related by nature that the thought can move the body …" (1991, p. 228; emphasis added).[99] In modern times, Mathis Synofzik et al. (2008) have described the qualitative feature arising while performing a voluntary movement and not while making a reflex movement, an "anarchic" movement, a "Penfield action,"[100] etc. as "a rather diffuse feeling of a coherent, harmonious ongoing flow of action processing" (p. 415), which is just to say a feeling of *controlled* action processing.

That feeling of agency, the phenomenon of control, is my focus in this chapter. Specifically, I examine whether it implies the causal efficacy of consciousness. Now, I regard felt agency as a conscious feeling by definition; the questions I ask are: (i) Is consciousness causally relevant to the feeling arising? That is, are the mental states that ground the feeling necessarily conscious? And (ii) what purpose does felt agency serve, and is its c-property causally relevant to that purpose? If so, how?

As to the first issue, I take it as a premise that felt agency is (partly) grounded in the fact that a volition *represents* an act: my volition to do *a* – call it *v* – is about *a*, insofar as *v* deploys some representation of *a* or its properties. So unlike cases of purely physical causation, here there is a sense in which the effect can be "found" in its cause. Following Searle (1980), *v* can be

---

[99] Berkeley also thought the causal force of the will was introspectible, at least in regard to producing thoughts: The "making and unmaking of ideas doth very properly denominate the mind active. This much is certain and grounded in experience" (Armstrong [1965], p. 72). Indeed, for Berkeley, volitions are the *only* things that cause, phenomenally and actually: "When we talk of unthinking agents, or of exciting ideas exclusive of volition, we only amuse ourselves with words."

[100] In the 1950s, the American neurosurgeon Wilder Penfield elicited movements from patients via direct neural stimulation; the patients reported experiencing the movements as outside of their control.

further analyzed as a propositional attitude *I will that* P, where *P* is a proposition expressing a certain act (*a*) occurring immediately (e.g., *I lift the cup now*). To use Daniel Wegner's term, an act is "consistent"[101] with *v* if and only if the act is the truth condition of *P*. We can also say that such an act meets the "satisfaction conditions" given by *v*'s representational content.

Wegner takes perceived will/act consistency to *explain* the feeling of agency, along with the other principles given in his Theory of Apparent Mental Causation: priority (that the agent perceive the volition to occur just before the act) and exclusivity (that she perceive no other cause of the act).[102] Once these conditions are met, the person will tend to infer that she caused the act, and in turn *feel* she caused it, Wegner argues. "When we think that our conscious intention has caused the voluntary action that we find ourselves doing, we feel a sense of will," he writes. The causal judgment thus requires consciousness of two data: the intention (which Wegner sometimes calls "the thought") and the action. The ensuing feeling is "based on the causal inference that one makes about data that do become available to consciousness – the thought and the observed act." In other words, "the apparent link [is] between the *conscious* thoughts that appear in association with action" (2002, pp. 65-68; emphasis added).

I do not dispute that an experience of agency would likely result from Wegner's conditions being met. In particular, if the consistency criterion is satisfied, the person's volition effectively prefigures the act. As a result, she will tend to think she has "designed" the act, which naturally fosters a feeling of control over it. It is the consistency criterion, then, that does most of the explanatory work with regard to felt agency. It also most often explains the *lack* of felt agency, as we typically feel nonagentive during action when our act fails to satisfy our will; less

[101] See Wegner (2002), p. 78; or (2003), p. 67.
[102] See Wegner (2002), Ch. 3.

often do we perceive another source of control over our act (a breach of exclusivity), and far less often would we experience our will to act following our acting (a breach of priority).

The drawback to Wegner's conditions, however, is that they constitute a *phenomenological* ground for felt agency: mental causation will be apparent to us (i.e., we will feel we cause a given act) only if certain *other* states of affairs are apparent to us, namely that the relations of priority, consistency, and exclusivity hold between our volition and our action. And "apparent," for Wegner, implies "conscious": "It is only when a thought is conscious prior to action that it can enter into the person's interpretation of personal agency and so influence the person's experience of will." But as I will argue, in the course of everyday action these relations – and volitions themselves – are seldom apparent: we are simply aware of acting, and feel agentive in so doing. So while Wegner's three conditions may be sufficient for felt agency, they cannot be necessary – at least insofar as they require *conscious* judgments by the agent. In other words, since consciousness is included in the set of conditions, the set as a whole isn't necessary.[103]

My first aim is to adapt Wegner's account to address this problem, focusing on volition/act consistency. If a person is to feel agentive in the course of action (as is typical), she must judge her act to be consistent with her volition during that time, as Wegner's theory requires. But volitions, as I argue in sect. 3, are usually nonconscious. And even when they are conscious, they tend to be phenomenologically "indeterminate" or "thin," as Metzinger (2006) has observed; that is, their phenomenal properties tend to be few and slight. Prima facie, these

[103] See Wegner (2002), p. 164. Other passages (pp. 164-165) suggest Wegner doesn't hold consciousness to be a *necessary* condition for "the process of assessing apparent mental causation." For example, he writes that "the experience of will is not *likely* to occur when action is caused by unconscious thoughts," and that in such a case, the person "will normally experience a *reduced* sense of conscious will" (emphases added). My view is still at odds with these weaker claims, insofar as I think the unawareness of volition entails neither an improbability that one will feel agentive nor a reduction of that feeling. Nonconscious volitions are able to support a robust sense of agency, as I will argue.

factors would prevent the person from assessing volition/act agreement. In view of these considerations, I argue that consistency judgments are made on the basis of a volition's representational content, which need not be conscious. Positing a *nonconscious* consistency judgment, one based on nonconscious volitional content, does justice to Wegner's intuition that will/act consistency supports the feeling of causing a movement. As to the role of consciousness in *causing* felt agency, then, I will argue that it plays none: nonconscious states are sufficient to generate the feeling. In sects. 6 and 7, I extend Wegner's explanation of the feeling's source to *mental* acts like imaging and thinking, arguing against Galen Strawson's claim that such acts are phenomenologically passive. I also take issue with a different account of felt agency proposed by Rosenthal, in sect. 8. On that view, the sense of control over a given act arises from the fact that one is conscious of one's volition causing the act and *not* conscious of the causes of that volition.

My second aim is to determine what function felt agency may play in the mind/brain. We might posit that the feeling plays the information-processing role of signaling action-control to the brain's executive system (responsible for reasoning, planning, etc.). But for reasons I will discuss in sect. 9, the natural occupant for that role is the *belief in agency*, the immediate result of Wegner's conditions being met. The functional role of the belief in agency, then, is as a representational state – it represents that one is controlling a given act. In contrast, the role of the sense of agency, I will argue, is as a qualitative state: the positive version (a feeling of efficacy and harmony during action) draws one to seek to satisfy one's volitions, while the negative version (a feeling of impotency and frustration) moves one to correct and avoid making poorly controlled acts. Essentially, feelings of agency and non-agency act as cognitive drivers like pleasures and pains, and the evolutionary explanation I offer for that role is that it is generally beneficial for a creature to seek successful voluntary action.

My last aim is to isolate the causal role of the sense of agency as a conscious qualitative state, which it regularly is.[104] Following a higher-order theory of consciousness, this requires answering the question, what is the causal relevance of higher-order states with contents such as *I am feeling power over act a*, *I am feeling frustrated in trying to produce a*, etc., qua their respective h-properties? For if the sense of agency is efficacious qua its c-property, such HORs must be efficacious qua their h-properties. What doesn't necessarily hold is the converse entailment, namely: If such HORs are efficacious qua representing their target qualitative states, then the targets cause qua being represented – qua their c-properties. As discussed in Chapter 2, sect. 2, that entailment only holds if the c-property and the h-property are complements (i.e., constitute a relation between the states). It does not hold if the c-property is merely the target's accompanying the HOR's instantiation of the h-property, for then that HOR can cause qua its h-property without the accompaniment being causally relevant.

Now, in sect. 9, I will argue that a HOR about a control quale plays an information-processing role: its occurrence constitutes the executive system's being informed about the occurrence of that quale. That "informing" is clearly a function of the HOR's h-property specifically. The property is thus necessary to deliberative and reasoning processes that take into account that quale. So it seems that my theory, in giving this condition for the efficacy of the h-property, entails that the first-order state causes such metacognitive processes qua being conscious *only* if the c-property and the h-property are complements. However, there is another theoretical option: The efficacy of the h-property does not entail that the first-order state causes

[104] Higher-order theories of consciousness do allow for the possibility of nonconscious qualitative states, and thus would admit nonconscious qualia associated with act control. But even if such qualia are logically possible, they may not exist in fact. In any case, my concern here is with the ubiquitous *conscious* qualia that I take to define felt agency. This is not to say that felt agency need be *focally* conscious: If I am thinking while typing, my sense of controlling the keystrokes may be reduced to the fringe of my awareness. Several researchers have discussed the phenomenon known as fringe consciousness, or the periphery of awareness. See, for example, Bruce Mangan (1999).

qua being conscious, but it does entail that *consciousness of control* is efficacious – given the view that HORs are sufficient for consciousness. A third option I will discuss allows that the efficacy of the h-property is insufficient for the efficacy of the c-property of the target state, or for the efficacy of state consciousness understood as a phenomenal property of the HOR. One argues, instead, that the first-order quale's c-property is just its *accompanying a HOR's instantiation of the h-property,* and claims that the quale is also a cause of the metacognition in virtue of that property. That is, it brings about deliberation and reasoning concerning itself, with the HOR's h-property as a co-cause.

2. Mental Causation: Not Merely Apparent

While I agree with Wegner that felt agency is grounded in certain judgments concerning volition and act, I do not subscribe to his view that mental causation is *purely* phenomenological, or illusory. I assume that volitions can, and typically do, cause acts. As discussed in Chapter 2, there are several challenges to mental causation, but there are also plausible solutions. In particular, there are theoretical resources to establish the efficacy of volitions if they are taken to supervene on the preparatory motor stages that Wegner takes to be the real cause of action.

Given that mental efficacy, however, an alternate explanation of the sense of agency should be explored; namely, a volition's actually causing a movement is what produces the feeling of causing it. Prima facie this explanation is problematic, since the neural and physiological mechanisms by which a volition causes a movement aren't accessible to introspection. But that's not a problem: nonintrospectible neural and bodily processes cause

conscious qualitative states all the time, and the feeling of control is such a state. Rather, the drawback to this proposal is that it would only account for our sense of controlling willed bodily movements, as opposed to the willed *effects* of such movements. Following Wolfgang Prinz (2003), let's call the former "resident goals" and the latter "remote goals." Oftentimes remote-goal agency is stronger phenomenologically than resident-goal agency. For example, compared to our sense of making the kitchen light come on, our felt agency over the switch-flicking motion is typically reduced to fringe consciousness. Now, suppose the latter sense arises from the causal chain that leads from the volition to flick to the flick itself; that is, it arises from the successful operation of certain motor mechanisms. As the felt agency qua the light is the *same* phenomenon, only with a different (remote) object, we would expect the same kind of explanation for its occurrence. But the causal chain between the volition to flick one's finger (or to turn on the light) and the light's coming on runs outside the body, through the electrical wiring and so forth. Clearly, no sense of agency over the light's coming on could arise from *that* causal chain.

What this means is that something more[105] is needed to explain felt agency qua remote effects, and that something is naturally the consistency of that effect with our intentions (along with Wegner's other criteria). And I argue that, in fact, the same explanation applies to felt agency qua resident effects. Consider the following thought experiment: My volition to wave my hand causes the waving, but unbeknownst to me, my motor system is unable to effect the waving, which is actually produced by a computer that reads a chip registering my volition, and then causes another chip, in my spinal column, to produce the waving in the exact manner and with the exact timing that it would have been produced had my motor system been operative.

[105] I do not deny that *if* motor causation explained felt agency qua resident effects, it arguably would be required for felt agency qua remote effects. For perhaps one couldn't feel that one has caused the light coming on if one feels one has not caused the flicking motion, and one knows that the former is caused by the latter.

Here, it seems likely that the felt agency over the waving would remain, despite the absence of motor causation. If that is so, it follows that motor causation is unnecessary for felt agency to occur. While this hypothetical case is merely an intuition pump, there is some empirical grounding for the idea that felt agency for bodily movements does not depend on motor causation. Consider experiments with brain-computer interfaces (BCIs) that relate to motor control. The subject (often a paralyzed person) has electrode implants in her motor cortex that transfer signals to a "decoder" that converts them to computer commands. In 2005, tetraplegic Matthew Nagle became the first person to control a prosthetic hand via a BCI, specifically the BrainGate interface. He reportedly expressed delight and a "transformed sense of independence" (Henderson [2006]) regarding the successful trial, suggesting he felt agentive about the hand's movements despite his impaired motor process. Now, it might be objected that since the hand is not actually part of Nagle's body, its movements qualify as *remote* effects, and so the case at most would show that felt agency for remote effects can obtain despite motor impairment. Felt agency for resident effects may still require that the normal process of motor causation be intact. BrainGate's ongoing work may eventually reveal whether this is so, as the company seeks to build an electrical stimulation device that would receive electrode input and move paralyzed limbs directly.[106] If such movements obeyed motor commands as precisely as movements caused via the spinal column, the BrainGate-assisted action would meet the conditions of consistency, priority, and exclusivity as well as the normal case. Patients could then judge whether they feel fully agentive qua these resident effects. Wegner's theory predicts they would, insofar as it holds that the agent's awareness of those conditions being met is necessary and sufficient for the sense of control. But as will be argued, that awareness need not be *conscious* awareness, contra Wegner's proposal.

[106] See http://www.braingate.com/action.html

3. The Phenomenology of Volition

A volition represents an action, and when one is conscious of volition one is conscious of willing an act *as represented* in some way. That representation can exhibit varying degrees of specificity, depending on the number of action components that make up the conscious content. When playing pool, for example, one can consciously will to *strike the 6-ball low*, or simply to *strike the 6-ball*. One's ensuing movement can agree with both contents, and thus promote felt agency in both cases, following Wegner's account. But it may also agree only with the latter, more general volition. Thus, if one wills to *strike the 6-ball low* and then strikes centrally, one's act has only partially agreed with the volitional content. Accordingly, a *diminished* feeling of agency should result. One would not sense an utter lack of control over the motion, as if one's stroke had gone in a random direction; nor would one experience total control. So the consistency criterion, it seems, can even account for degrees of felt agency. Note also that the account is not committed to the *linguistic* representation of movements in volition, as in the forgoing example. Plausibly, a conscious volition may "pick out" the desired act via nonconceptual representations of the sensory stimuli and proprioceptions expected to obtain with the movement, or via a mental image of the movement.

But while one *can* consciously will in these ways, a conscious volition tends to be phenomenally "thin" as compared to conscious perception, Metzinger notes. It may be experienced as no more than a mild urge prior to movement, for example. Volitions are also phenomenally "evasive," meaning that they tend to recede when focused on, again in contrast to perceptions, whose phenomenal detail tends to become more salient when brought to focus. So

even though a volition *allows* for a consistency assessment by being introspectively distinguishable from action,[107] there may not be much *to* distinguish, in order to facilitate a comparison. Still, it may be that the urge – or sense of being about to act – is always relative to some particular act, however vaguely represented. And if there is some conscious representational content that goes along with the quale of being about to act, a consistency assessment can be based on that conscious content. A more serious problem for Wegner's account, I argue, is that volitions are usually *entirely* nonconscious, meaning that their representational content is typically nonconscious and any accompanying quale, the "feeling of being about to act," typically does not occur. Based on my own introspection, I find that for my typical act *a* consciously performed at time *t*, I am *not* conscious of a volition to do *a* or feeling about to do *a*, etc., just prior to *t*. I am simply conscious of *a*, while concurrently sensing control over *a* (to whatever degree).[108] To see why this phenomenological picture of voluntary action is accurate, consider the sheer number of acts that one may consciously perform, even within the space of a minute, whether pressing a key, folding one's arms, reaching for a cup of tea, etc. Presumably, there is a volition driving each of these acts, but how often is volition conscious? Relatively seldom, I think. One is aware of lifting a cup of tea, and feels agentive in so doing, usually without being conscious of a volition to do so, even in a phenomenally thin sort of way.

A couple of objections are plausible here. First, it might be held that careful introspection reveals that conscious volition routinely occurs. For example, in Libet's (1983) seminal study of voluntary action, subjects were asked to keep a close watch on a dot that revolved around an

[107] This is a roundabout way of saying that a volition is phenomenally *opaque*: it presents itself as a mental state, distinct from any action it causes. One is thus able to assess, from the first-person stance, how well the volition represents the action. A perception, on the other hand, is phenomenally *transparent*: it simply presents (in some way) the external object that causes it, so one can't judge its representational accuracy, at least not in the act of perceiving.

[108] This, I think, is the result of introspection that is not "theory laden": if one believes that every voluntary act is consciously willed, one must be careful not to let that belief inform the introspective report.

analog clock and perform a spontaneous wrist flick at some point. They were then to report the position of the dot at the time they felt the urge to move (what Libet et al called the "W judgment") and its position at the time they were aware of moving (the "M judgment"). Subjects were always able to make a W judgment, which of course they could not do if their volitional activity were completely nonconscious. Moreover, a wrist flick is neither a challenging nor novel movement for the subjects; so there are no grounds for writing off the act as a special case where conscious volition *would* be the norm. But while the act itself is commonplace, the psychological scenario is not. One is not typically disposed to introspect one's volitional activity prior to action (at least not for rote actions like a wrist flick), as the subjects were in Libet's experiment. So the subjects did not simply happen to focus on their volition and judge its time of occurrence, entailing that it occurred consciously. In addition, prior to that introspective act, they were *disposed* to assess the time of volition, per the experimenters' instructions. That, I argue, likely caused the volition to occur consciously in the first place. A disposition to attend to one's mental activity at an upcoming time – e.g., a volitional state prior to movement, an emotional state while listening to a song, a deliberation when faced with a buying choice – is of course a desire to be fully conscious of that activity. Naturally, such a disposition could cause states to occur consciously that otherwise would not, or to become conscious in more phenomenal detail.

Second, it might be objected that volition *is* typically conscious, but not saliently so, since the phenomenal stages of voluntary action are "temporally compressed." As Haggard notes: "We do not normally have several distinct conscious experiences corresponding to each element in the [voluntary action] chain, such as desire, plan, intention, movement, feedback, and consequence. Rather, we tend to have a rapid, condensed experience of the whole sequence of events" (2006, p. 73). The idea, then, would be that conscious volition is fleeting, which leads

one to forget that it occurred for most acts – even though it was reportable (as all conscious states are) at the moment it occurred[109] But surely one could usually report contents that occurred consciously a few seconds ago, even if they were fleeting: I find I can report plenty of acts I've just performed, but typically not volitions, urges, and the like preceding them. Thus, it's doubtful they were conscious.

To claim that *conscious* volition seldom precedes felt agency is to claim that "minimal actions" are the norm. As Tim Bayne and Neil Levy express the notion, minimal actions "might be caused by intentions, but that is not how they are experienced. It seems possible to experience oneself as performing an action without experiencing that action as the result (or implementation) of an intention." That is, minimal actions feel agentive without being preceded by conscious intentions (2006, pp. 50-51).[110] Searle has acknowledged such actions, which he describes as "spontaneous":

> Many of the actions one performs, one performs quite spontaneously, without forming, consciously or unconsciously, any prior intention to do these things. For example, suppose I am sitting in a chair reflecting on a philosophical problem and I suddenly get up and start pacing about the room. My getting up and pacing about are clearly

[109] Perhaps it can even be argued that since the experience of volition is fleeting, one may be unable to introspect (and report) it at the time it occurred. But barring any psychological blocks or distractions, why *wouldn't* one be able to introspect a presently conscious state? No matter how fleeting the state is, if conscious, it's introspectible and reportable under normal circumstances.

[110] Other authors more resolutely hold that the sense of agency can be divorced from *conscious* volition: Metzinger, for instance, gives this example: "Imagine snatching a child away from a fastly [sic] approaching car. We have a full-blown experience of agency, but no subjective experience of any preceding volitional process" (2006, p. 27). I would add that this phenomenology of agency applies more broadly to "everyday," non-heroic actions, as well as thoughts. As Shaun Gallagher has observed: "[W]e should say that most cases of normal thinking are neither prefaced by conscious intentions to think, nor followed by an introspective awareness of that intention. In normal phenomenology, at least in the large majority of cases, there is not first an intention and then a thinking, nor thinking plus a concurrent but separate awareness of intention to think" (2004, p. 12).

> intentional actions, but in order to do them I do not need to form an intention to do them prior to doing them. (1980, p. 52)

I agree with Searle that many, if not most, of our acts do not *appear* to be preceded by intentions; that is, they seem spontaneous. But I see no reason to deny that they are *unconsciously* willed prior to action. One of Searle's motivations for positing an intention-in-action – namely, an intention that occurs along with the movement and has the content *I perform movement x as a result of this intention-in-action* – is to explain how spontaneous movements can be intentional actions.[111] They are intentional actions, minimally, when they satisfy intentions-in-action, meaning that they are represented by those intentions. That in turn entails they are *caused* by those intentions, as the representational content specifies. And since intentions-in-action occur *along with the action*, they imply no preceding awareness of volition. Yet we may alternatively posit nonconscious volitions to explain why spontaneous movements count as intentional actions. Moreover, as a *prior* intention, a nonconscious volition is metaphysically suited to cause the movement, as opposed to a concurrent intention-in-action.

Anthony Marcel, though he allows that volitions often fail to be experienced (particularly in cases of "immersed" action), questions how they can ground felt agency when nonconscious: "Unawareness of intention appears to be more common than supposed. … [I]t is hard to see how

[111] Another motivation for positing intentions-in-action is to rule out the satisfaction of prior intentions by behaviors they *obliquely* cause. Consider the case posed by R. Chisholm that Searle cites (1980, p. 51): Bill intends to kill his uncle and does so as a result of that intention, but in the following way: "Suppose he is out driving thinking about how he is going to kill his uncle, and suppose his intention to kill his uncle makes him so nervous and excited that he accidentally runs over and kills a pedestrian who happens to be his uncle. Now in this case it is … not true to say he carried out his intention to kill his uncle." It is only *actions* that "carry out" or satisfy prior intentions, and these are intentions-in-action causing behaviors. Thus, the causal link between a prior intention and an intentional behavior is an intention-in-action (and not, say, nervousness). Bill lacked an intention-in-action; so his killing his uncle was not intentional. I'm claiming that a volition can play the role Searle posits here for the intention-in-action: Bill had the standing intention to kill his uncle, but he (presumably) lacked an intention to kill his uncle *now* (i.e., a volition), just prior to running over his uncle.

intention, in those cases where we are unaware of it, can … play a significant part in the sense of our causality of an action" (2003, p. 61). My reply to Marcel's concern is essentially that nonconscious volitional content can figure in a consistency assessment, which in turn grounds felt agency. This position entails that consistency assessments are also typically nonconscious. For if they were typically conscious, that would entail – implausibly, as I have argued – our routine consciousness of volitions' satisfaction conditions. Therefore, we must posit nonconscious consistency judgments, based on nonconscious volitional content, as the psychological norm during action if we are to adhere to a Wegnerian account of felt agency.[112] Note that nonconscious volitions (though not phenomenally thin ones) would also inhibit conscious *priority* judgments. So we must posit nonconscious versions of those judgments as well.

Now, since these judgments involve volitional content as well as a representation of one's act, their being nonconscious entails that one also nonconsciously represents the act prior to feeling agentive over it. The subsequent feeling, however, entails act consciousness at that time: How can one consciously feel in control of an act one isn't conscious of performing? The sense of agency is thus always relative to some particular act. So one never feels agentive simpliciter, but over some act *a*, however vaguely one is aware of *a*. But the various representations leading to that feeling – a volition, an act perception, and Wegnerian judgments about volition and act – are usually nonconscious.

Wegner seems to acknowledge the phenomenology of volition – or general *lack* of phenomenology – I have argued for: "For some amount of what we do everyday, our conscious intentions are vague, inchoate, unstudied or just plain absent" (2002, p. 145). His means of

[112] Haggard notes, "[I]n everyday life, this matching process [between intention and action] seems to operate in the background, without focal directed attention, and in phenomenal silence" (2006, p. 71). Contra Haggard, I don't think the process is even fringe conscious during everyday action.

preserving his account of felt agency in such cases is to posit that we confabulate "prior consistent thoughts" after the act, to *then* feel agentive about it. But even if we engage in this confabulation as often as we feel agentive for acts that lack conscious volitions (which is questionable), the account would not explain the agency we often feel *during* acts that are not preceded by conscious volitions. Thus, a better approach is to posit that consistency judgments are usually made nonconsciously, based on volitions' nonconscious representational content.

Prior to the feeling of agency (or of lack of control), there is another intentional state that I also claim is typically nonconscious: the *belief* that one is (or isn't) in control of a given act. That's the result of the "inference" about mental causation Wegner posits, which leads in turn to felt agency.[113] Like a volition, an agency belief represents the self, causation/control, and the act in question. But the contents of the two states differ: an agency belief is about an act one *is* controlling or causing, not an act one *wants* to bring about. Neither type of state is typically conscious. As to the agency belief: we surely believe that we cause each act that we feel control over, though without consciously thinking so in each case. Thus, what we may call the intentional precursors of felt agency – volition, Wegnerian judgments, agency belief – are all typically nonconscious. It follows that *conscious* involvement with the generation of felt agency is usually quite limited; that is to say, there tend to be few conscious states, if any, that cause it. The feeling thus stems from, and in effect indicates, typically nonconscious processes of action-generation and evaluation. More fundamentally, the feeling would – in the case of felt agency over bodily movement – indicate the comparator mechanism's matching a forward model to

[113] Synofzik et al. (2008) have distinguished this belief as the "reflective" (vs. pre-reflective) sense of agency. But contra my proposal that the agency belief ("reflective level" of agency) causes felt agency, they argue it's the other way around: felt agency, once "conceptually processed," yields the belief that one is in control. That's why they call felt agency *pre*-reflective. But Synofzik et al. may be overlooking the possibility of a nonconscious agency-belief. It seems to me that once the system nonconsciously judges volition/act agreement, the most immediate cognitive effect would be a nonconscious *belief* that one has caused the act, typically followed by the conscious feeling of control. Indeed, nonconscious processing is generally held to be faster than conscious processing. That's not to say the feeling could not, in turn, bring that belief to consciousness, upon reflection.

sensory feedback. That neural process, I propose in sect. 5, subvenes a nonconscious consistency judgment qua one's movement. But first, I address a problem that Searle's view of volitional content poses for Wegner's account of how we come to believe in our agency.

4. Consistency Assessments and Searlean Volitional Content

The notion of a volition's satisfaction conditions has been discussed by several writers. Metzinger defines them as "those conditions that would make the action count as *successfully terminated*" (2006, p. 21, 28), while Searle offers a more explicit account: "[M]y intention is satisfied iff the action represented by the content of the intention is actually performed" (1980, p. 49). As noted in sect. 1, a volition *v* represents an act propositionally: *I run now, I turn on the kitchen light now*, etc.[114] If this intentional event is actual, *v* is satisfied, which is to say there is consistency between *v* and the act. Following Searle's approach, we can hold that *v* also has a certain self-referential content, namely that the act be performed *by way of carrying out this volition* (1980, p. 53). Since this condition requires that the event represented by *v* be caused by *v*, *v* will not be satisfied if I just happen to perform the action that *v* represents, or perform it due to some other cause than *v*. Searle has intuitive grounds for thinking that intentions have that kind of content (which I will call "Searlean content"): If I order someone to leave the room, and he leaves but for another reason than my command, he can't be said to have obeyed (i.e., satisfied) my order. So my command must have the content that the person leave the room as a result of my command.

[114] What makes *v* a volition as opposed to a standing intention is the represented time of the action is now (see n. 98); the latter's time representation would have the content *at some later time, when appropriate*, etc.

Searlean content prima facie creates a circularity problem for Wegner's account of how the belief in agency arises. If *v* represents not just my action occurring now, but my action occurring now *because of v*, then judging that *v* is satisfied requires believing that I have voluntarily caused the action it represents, e.g., that my willing to turn on the light has caused my turning it on. So judging consistency would be (partly) constituted by an agency belief. It can't then *explain* how that belief arises.

I think there are two ways to solve this problem. First, it is arguable that volitions do not have Searlean content. In the case of an order to another person, that kind of content is plausibly involved. We want the person to act in a certain way *because of our command*, and we have that latter condition in mind precisely since here our agency is a concern to us: The person might do what we want, but for some other reason, and we don't want that. But in the case of our actions, if they occur as we will them to, it's nearly always because our will brought them about. So it's doubtful we have thoughts such as *I will that I grip this mug as a result of this volition*. We probably just represent the desired movement. Second, even if Searlean content *is* part of a volition *v*, we can say an agency belief results (inter alia) from perceived agreement with *v*'s representation of the movement, as distinct from its Searlean content.[115]

Of course, I am not in a volitional state simply by representing an action of mine; I must also will the occurrence of that action. Fodor's (1987) Representational Theory of Mind would analyze such a mental attitude functionally, and in this vein we can say that *v* is a volition only if it tends to bring about the action it represents. More formally: Let $r(x)$ be a function mapping mental representations to realizations of their intentional objects, and let $b(x, y)$ map ordered

[115] Indeed, since to judge agreement with *v*'s Searlean content is to believe that *v* has caused (i.e., I have "voluntarily caused") the movement *v* represents, it follows on Wegner's account that judging agreement with that content *results from* and is explained by judging agreement with *v*'s representation of the movement, along with the exclusivity and priority judgments qua *v*.

pairs of representations and sets of conditions[116] to immediate behavioral results. Then, *v* is a volition only if (i) it represents an action of oneself, and (ii) $r_{(v)} = b(v, y)$ for typical conditions that are values of *y*. The second criterion analyzes the "willing" component of *v*, as distinct from its representational content. Thus, for me to will the occurrence of the act *v* represents – say, my setting the alarm clock – is for the representation to instantiate a certain functional property, namely causing (via my motor system) me to set the alarm clock under typical conditions. (Under atypical conditions, such as motor impairment, the representation of course would not have that effect.) A different attitude toward that content, say a *belief* that I set my alarm now, could not be defined by (ii). For in that case, $b(v, y)$ would *not* yield $r_{(v)}$ given typical values of *y*. Take an average circumstance where *y* = my being asked if I have set my alarm, my being able to speak, etc. Then, $b(v, y) \neq r(v)$, but rather $b(v, y)$ = my reporting that I've set my alarm. We would have to resort to unusual sets of conditions as values of *y* in order for a belief that I've set my alarm to cause me to set my alarm. For example, suppose I am hypnotized to want to try to redo actions I believe I've performed. In that case, $b(v, y) = r(v)$.

It is also possible to functionally define the attitude involved in volition in terms of the representation's *mental* effects, as opposed to its behavioral ones. For instance, if we subscribe to Wegner's theory of felt agency, it is clear that a volition will tend to initiate the judgments that ground that feeling: once I will to set the alarm now, I tend to consider (consciously or not) whether my ensuing act is consistent with *v*. (That judgment process won't arise from the same representation under other propositional attitudes, such as if *I wonder if* I've set the alarm.) In turn, I tend to acquire a belief and feeling regarding my agency. So we can also say a representation of an act is a volition only if, under normal conditions, it works together with act

[116] These would be complete specifications of the person's other psychological and neural states, current sensory inputs from the environment, and physical situation.

perceptions to cause these further mental states. Those conditions would include working visual and proprioceptive systems (necessary to act perception) and, in the case of bodily acts, a working comparator mechanism, as I discuss in the next section.

5. Consistency Assessments and the Comparator Model

One may judge consistency between volitional content and one's bodily movement (accordance with a "resident" goal) or an effect of that movement in the environment (accordance with a "remote" goal), as explained in sect. 2. In the former case, the judgment bears an interesting similarity to the operation of the comparator mechanism proposed as a model of motor control.[117] In this system, the motor system determines which commands will result in a given desired state of the body, generating an "inverse model."[118] A copy of the commands, the "efference copy," is then sent to a comparator or central monitor, which uses the information to build a "forward model" that predicts the sensory feedback the movement will cause. When the actual feedback arrives, the monitor can determine the congruence of the forward model with the estimated actual state. Prior to that, a comparison is also thought to occur between forward model and desired state. If both comparisons result in a "match," the actual state of the body will be the desired state (assuming that the estimation of the actual state is accurate). This serves various functions, including distinguishing those sensory signals that are self-generated from

[117] See, for example: S. J. Blakemore et al. (2001), M. Kawato (1999), and C.D. Frith (1992).
[118] There is more than one possible set of commands to achieve a given desired state, of course. Many different finer-grained movement components could result in a catching-the-child position, for example – and so we have what is known as the *inverse kinematics* problem that the motor system must solve. For a review of this issue, see Haggard (2001).

those that are externally generated, and making online adjustments to subsequent motor commands to improve performance.

I propose that the consistency judgment – in the case of resident goals – supervenes on the process that compares desired state to estimated actual state via inverse and forward models. But this scenario is only plausible if we assume that nonconscious consistency judgments are the norm. Comparator operations are presumably ongoing, and nonconscious consistency judgments may well also routinely occur (unlike *conscious* judgments). Moreover, since nonconscious processing is generally held to be faster than conscious processing, a nonconscious consistency assessment could (theoretically) occur as quickly as a comparator output. The motivation for positing this supervenience is the empirical evidence that felt agency for movement causally depends on comparator outputs. If that is so, then the natural way to preserve Wegner's account of felt agency is to posit that consistency assessment engenders the feeling at the mental level of while comparator outputs cause it at the neural level (that is, they cause the feeling's neural basis). This is an application of Yablo's approach to mental causation, discussed in Chapter 2, sect. 6.

One example of the empirical evidence I allude to is the work of Chloe Farrer et al. (2003). Among the primary neural regions implicated in the matching process are the cerebellum and posterior parietal cortex, and Farrer et al. conducted a study supporting a causal link between comparator operations and agency feelings based on activity in these regions. In that study, (i) subjects continuously moved a joystick with their right hand, which was hidden from view; (ii) visual feedback about their movement was provided by an electronic image of their hand reconstructed at its exact location; and (iii) the image's movements gradually diverged from those of subjects' actual hand. Subjects were instructed to concentrate on their feelings of control

over the electronic hand. Farrer et al. found that activation in the inferior parietal lobule was inversely related to subjects' reported feelings of control: the more the discordance between electronic hand movement and actual movement, the more discordance between visual feedback and forward models; in turn, the greater the parietal activity, and the *less* felt agency. A plausible interpretation is that the inferior parietal lobule codes for specific degrees of agreement/disagreement between forward models and sensory feedback, and that these comparator results are causally linked to different degrees of felt control. As Nicole David et al. write, that sense "strongly depends on the degree of congruence versus incongruence between predicted and actual sensory outcome" (2008, p. 524). Thus, we can expect *degrees* of felt agency to be causally explained by comparator operations. For example, a detected mismatch, note Synofzik et al., can cause a feeling of semi-agency: "In the case of incongruency between these indicators (e.g., a mismatch between proprioception, motor intention and visual feedback), we experience an action as strange, peculiar and not *fully* done by me" (2008, p. 415; emphasis added).

Given this causal link, we might question the need to posit a nonconscious consistency judgment supervening on comparator output: Why not assume a "match" at the neural level fully accounts for felt agency and its degrees, as some proponents of the comparator model do (e.g., Chris Frith)? In response, I would point out the force of Wegner's intuition: to *feel* that one causes an act, one must first *think* that one causes it, by inferring causation from the various principles he cites. Typically, feelings are caused by beliefs: one feels anxious while taking a test because one believes it is important, one thinks it is difficult, etc., though these beliefs often remain nonconscious while one is anxiously at work. Similarly, one feels agentive qua some act *a* because one judges (again, often nonconsciously) that *a* is consistent with one's volition, *a* has

closely followed one's volition, etc. Thus, we causally explain feelings by first appealing to their psychological antecedents like beliefs, and there is no reason to eschew that kind of explanation when it comes to felt agency. The information deployed by the comparator – e.g., visual cortex activity produced by reafference after a movement – is not our rational basis for thinking, and subsequently feeling, that we have caused a movement. So we have justification for positing a mental cause of felt agency along with the neural cause. That's not the case with regard to other (hypothesized) effects of comparator outputs: discriminating self-generated sensory feedback and fine-tuning movement control may be exclusively neural phenomena do not subvene any mental states; thus we have no reason to think they have mental causes.

Now, a volition/act comparison would of course be based on different kinds of representation than the comparator deploys. A volition and a perception of a performed act are mental representations, the kind of representations that *can* become conscious. And whether conscious or not, they are more coarse-grained representations than their neural subvenors: respectively, inverse models and the sensory feedback compared with the forward model.[119] In the relatively rare cases when volitions *are* conscious, as I have argued, they are still more coarse-grained, being phenomenally present to us as a brief feeling of being about to move. Conscious volitions, we may say, *representationally underdetermine* the nature of the movements they precede, more so than nonconscious volitions, and much more so than motor programs.

[119] For experimental evidence of our unawareness of the precise details of action control, see Marc Jeannerod (2003), sect. 1.3.

6. Consistency Assessments and Mental Acts

It's a contingent truth, of course, that conscious volitions tend to represent bodily movements to a lesser degree than nonconscious volitions do. Conceivably, conscious volitions could represent movements in more detail; in fact, they sometimes do. As Marcel writes, actions "will inevitably have details unenvisaged in the intentions. Exceptions to this are actions where great care or self-monitoring is needed" (2003, p. 66). One can even imagine conscious volitions representing movements to the degree of detail that motor commands do. But it so happens that they do not: Even if we bring to mind a very specific idea of the action we're about to perform just before acting, it won't (as a matter of fact) have the specificity of a motor program. This is plausibly due to neural architecture. Consider the evidence that conscious visual states are intermediate-level representations in the visual system: cells in mid-hierarchy areas (V2-V5) tend to encode object size and orientation, unlike cells in the high-level area (IT), which abstract from these features. And unlike cells in the low-level area (V1), mid-hierarchy cells tend to encode illusory contours and fail to encode highly local features like irradiation changes in each retina. Thus, the responsiveness of only the intermediate-level neurons correlates with the contents of conscious visual experience.[120] There is also evidence that conscious states are intermediate-level representations within the hierarchical systems of other perceptual modalities – and within the motor system. The latter's highest level is the prefrontal cortex and its lowest level is the primary motor cortex (MI), which actually executes movements through direct connections to motoneurons in the spinal column. But it is the activity of intermediate-level premotor areas – implicated in the *preparing* of a specific movement – that has been correlated

[120] For an in-depth review of the evidence, see Prinz (2005), Ch. 11.

to conscious volitions to move.[121] The point being, this correlation with mid-level processing is a contingent fact about the brain: we can imagine conscious volitions correlating instead with activity in MI and, accordingly, representing movements at a much finer grain of force, direction, position, velocity, etc., down to individual muscle contractions. That is, we can imagine much less representational underdetermination, or perhaps none at all. Suppose, for example, that we could perform a conscious act of mental simulation so precise that only one movement could satisfy it.[122]

Where the desired act is a mental imaging, such a feat of representation is not only conceivable, but realizable. The problem is, willing the act then collapses into the act itself. Consider the following case: a volition with the conceptual content *picture a bonsai tree* is followed by one's forming a particular mental image of a bonsai tree. The image could have been formed in all sorts of different ways that would satisfy that volition, and even a volition with more specific conceptual content (e.g., *picture a bonsai tree with 12 branches*) can be fulfilled by more than one imaging act. But here a mental simulation *can* be deployed in the volition such that only one act would satisfy it: *picture a bonsai tree in <u>this</u> way* [mental simulation] is capable of fully capturing the detail of the ensuing mental act (i.e., there can be

[121] See, for example, S. Obhi and P. Haggard (2004), p. 361.

[122] Clearly, the kind of representational vehicle I have been assuming is some description – linguistic or mentalese – of the various properties of a movement – speed, direction, force, limb recruitment, muscle recruitment etc. These being quite numerous (to say the least), we can expect representational underdetermination. But a different type of description can be deployed that would be satisfied by only one bodily act, namely, a definite description. So I can will to perform the bodily act that would greet my friend from a distance. This volition (given social context) picks out exactly one desired act: hand-waving. If I wave my hand, my volition *fully* represents that act – there are no degrees of representation. What I have been discussing, however, is not the representation of *acts* – if these are understood as abstractions of movements – but rather the concrete movements themselves. So while there is just one act that would satisfy this volition, there is no one movement that is a hand-waving, but many similar movements. Admittedly, it's possible to formulate a definite description of a movement, and incorporate that into an intention: For example, "I will now duplicate the hand movement I performed 10 seconds ago." Assuming I only performed one hand movement 10 seconds ago, there is just one way I can now move in order to satisfy this intention. But how will I go about satisfying it? I would need to deploy some mental representation of the physical properties my motion of 10 seconds ago, in order to attempt duplication. Thus, the mental state that would actually guide my movement – the state that would be the *volition* and not merely an intention – would need to describe movement properties. And that entails (contingently, I argue) that it underdetermines the nature of the ensuing movement.

perfect consistency), whereas *raise the right arm in <u>this</u> way* [mental simulation] would not fully capture the detail of the ensuing movement. Yet in the first case the intended act is subsumed in the volition. This phenomenon, of course, can only occur in the sphere of mental action. Consider a mental act that has purely intentional content, such as thinking 26. In this case there are no imagistic properties for the guiding volition to represent: unlike picturing a bonsai, there is, arguably, nothing qualitative about thinking 26, or at least there need not be. So the only properties that the volition must represent are intentionality, an intentional object (26), and, arguably, a mode of presenting that object (the mathematical concept *26*). For example, if 26 is the solution to Problem #4, the volitional content might be *think of the answer to Problem #4 under the mathematical concept that expresses it*. Thus we specify the intentionality of the act to be performed (via *think of*), its intentional object (via *the answer to Problem #4*), and its mode of presenting that object (via *the mathematical concept* …). We needn't deploy a volition with the content *think of 26 via the concept* 26 (indeed, we probably *can't* do so since we haven't solved the problem yet).

Unlike an act of mental imaging, then, a strictly intentional mental act can be fully specified by a volition without thereby performing the act. In both cases of mental action, however, it is clear that willing to perform mental act *m* does not (necessarily) consist in performing *m*. It is also clear that when *m* occurs, it can agree with the will's representation of *m*. As Wegner puts it, "It is not that we need to know everything in advance of thinking it, but that we need to know something about where our thought may be going that then is consistent with what we think when we get there" (2002, p. 86). Thus, there can be volition/act consistency in mental action, just as in the sphere of bodily action.[123] Following the Wegnerian account of felt

[123] Frith has even proposed that the comparator model of motor control can be applied to cognition, where a central monitor would compare the actual thought to the parameters for that thought set forth in the volition. We should not,

agency, we can expect the phenomenon to obtain for mental acts as well. Strawson, however, seems to think otherwise.

7. Strawson's Passivist Proposal

Strawson (2003) has argued at length that introspection reveals no sense of one's *causing* one's thoughts, but rather a certain passivity: "We find … that most of our thoughts – our thought-contents – *just happen*," he writes, offering his own introspection as support: "When I consider my mental life I find that things constantly impinge on me. I remember that I have to do X – it strikes me that Y is true" (p. 229).[124] At most, Strawson claims, one prepares oneself to *receive* thoughts, and feels agency in this regard. Such "catalytic" activity, as he describes it, can take various forms: "setting one's mind at the problem" one wants to solve, "focused concentration of will," a "receptive blanking of the mind," "maintaining attention," and so forth. Then, the "content outcomes are delivered into consciousness" by the "natural causality of reason" (or imagination, as the case may be) – if all goes well cognitively. If not, one is left with

---

however, commit ourselves to the claim that thoughts *must* be caused by volitions to think, for since volitions themselves are thoughts, a "never-beginning regress of intentions to form thoughts" would be entailed, as Akins and Dennett (1986) have pointed out. Rather, only conscious thoughts that one feels agentive in thinking must be caused by volitions, as the consistency-based account of felt agency requires. And if the volitions that cause such thoughts are typically nonconscious (as I will argue), then one cannot feel agentive in thinking them. Accordingly, they need not be caused by prior volitions, and so no regress threatens.

[124] Strawson uses the sense of passivity in cognition as a (putative) introspective datum to support his claim that thoughts and imaginings should not count as actions. I am not entirely persuaded by this phenomenological criterion: Does an agent *S* need to *feel* like she caused her thought *T* in order for *T* to be her action? It seems that as long as *S*'s conscious attempt to bring about *T* is a cause of *T* (along with whatever nonconscious "ballistic" cognition is required), *T* is something *S* does intentionally (i.e., an action), especially since *S* is arguably constituted by her nonconscious mind as well as her conscious one. In fact, as noted above, Davidson's theory of action does not require that the subject *consciously* represent an act as intentional: so if *T* "pops up" without my consciously intending it, the nonconscious causes of *T* (if these can be considered intentional) may be enough for *T* to count as my action. Thus, the right kind of causing, not any feeling of causing that may emerge, is what seems to be required for agency in cognition. In any case, the action status of cognition is not my primary concern here, but rather Strawson's premise that cognition is phenomenologically passive.

conscious primings and no results, rather like trying to remember a name and failing to. Strawson also allows that there can be representational content involved in conscious priming, as I have illustrated with the examples of willing to picture a bonsai tree and willing to solve a math problem. The content needn't be as conceptually explicit as the cases I offer, of course. When willing to generate a reply during conversation, Strawson explains, "one often knows … in some ineffably compressed manner, what its content is" (p. 229). But it *can* be: When willing to imagine something, "one must obviously start from some conceptual or linguistic specification of the content (*spangled pink elephant*)" (p. 241). Unlike acts of mental focusing, exertion or "blanking," which presumably have only qualitative character, such conscious volitional activity recruits *representational* contents, with which the mental outcome can agree (or fail to agree), as Strawson also allows: "… given that one's imagining duly fits the specification one may say that it is intentionally produced" (p. 241). But even if the outcome meets such "criteria" imposed at the catalytic stage, and one is conscious of the agreement, Strawson denies that a sense of agency emerges. Introspection reveals a passive reception of the desired thought or image: "What happens then is – a content just comes" (p. 235).

The phenomenology of cognition that Strawson describes is surely accurate in many cases, but not, I think, in all or even the majority. I sometimes feel that a mental image *f* or a thought that *P* "came to me," but more often I feel that I've "brought *f* to mind" or "judged that *P*." Even those thoughts that do seem to "impinge" upon me seem that way only because they carry a lesser degree of felt agency; I do not experience them as *completely* out of my control; indeed, if I did I would likely be suffering from some type of intrusive-thought disorder. Nevertheless, I think we can explain how one might come to believe, as Strawson does, that passivity in cognition is the phenomenological norm. Essentially, it's not that introspection

reveals one's control over acts of catalysis and nothing more. Rather, felt agency commonly accompanies (conscious) cognitions of all kinds, but certain considerations and introspective acts will tend to weaken it. One of those acts is to focus on how conscious catalysis representationally underdetermines the ensuing mental act. Indeed, if conscious catalysis typically consists in the kinds of purely qualitative phenomena Strawson describes – "focused concentration of will," a "receptive blanking of the mind," etc. – these will of course fail to represent the ensuing cognition *at all*. But the consistency-based account of felt agency is still viable in such cases, insofar as the phenomenon can arise from an act's agreement with the *nonconscious* representational content of volitions, as I have argued. So even if one's conscious activity just prior to thinking of the square root of 144 *is just* a certain inclining or blanking of the mind, what follows – *thinking 12* – can still agree with representations that happened to remain nonconscious during the catalytic stage, such as *think the number that times itself will yield 144*. As a result, one will experience control over *thinking 12*. Yet should one consider what one did consciously just prior to that thought – say exert one's attention – one may well be struck by how a mere act of mental effort could have brought about a specific representational content, as that exertion had no representational content itself, and thus could not specify the nature of the desired mental outcome. Basically, by introspecting the fact that the outcome is not at all prefigured at the conscious level of catalysis, one may disrupt the felt agency that naturally arises from volition/act consistency at the nonconscious level. Thus, I suggest that Strawson's introspective focus on catalysis of the purely qualitative kind, which cannot representationally prefigure outcomes, leads him to deny felt agency during cognition.

In turn, this introspection will tend to give rise to a consideration that further weakens the sense of control over mental acts: One realizes how conscious catalysis depends (perhaps in large

part) on nonconscious content, nonconscious inferences, subpersonal mechanisms, etc., in order to bring about a given thought or mental image. For example: "My mere exertion of attention seems insufficient to intentionally control my thought that *P*, so there must have been various nonconscious mechanisms at work. I thus depended on them to think that *P*." This explains Strawson's characterization of content as being "delivered into consciousness" – presumably by nonconscious cognition. On my account, when the entire conscious/nonconscious process is successful, the natural feeling engendered is one of the self *producing* the desired content.[125] So the dependence on nonconscious cognition, the "ballistic machinery," can only attenuate the sense of agency if one deliberately takes it into consideration. Mental contents might start to feel delivered, not produced. But this phenomenology would be artificial.

The overall point, then, is that the sense of control is to some extent under the agent's control: she can weaken it by focusing on representational underdetermination, by taking into account the dependence on nonconscious cognition, and perhaps there are other means. Now, in some cases – again, not the majority – phenomenal passivity in cognition would be natural. Strawson often suggests that one experiences a *delay* between catalysis and outcome: For example, he writes that "action, in thinking, really goes no further than [priming]. The rest is waiting, seeing if anything happens, waiting for content to come to mind, for the 'natural causality of reason' to operate in one" (p. 232). Strawson means this claim to apply to thinking in general, not just challenging mental acts like trying to write a philosophy paper. But the average thought does not seem to be preceded by willing and waiting; these are thoughts like "She shouldn't call me at this hour," "I hope it doesn't rain" or mentally running through a familiar grocery list. Only thoughts that are more difficult to arrive at, such as solutions to problems, creative ideas, hard-to-recall names, etc., typically involve such a lag. The reduced felt agency in

[125] Which is a well-grounded feeling since nonconscious cognition is (arguably) part of the mental self.

the latter case is easily explained by the consistency-based account: A volition not only represents a desired mental act in terms of its content (e.g., recall the name of the actor who starred in *Ben Hur*), but also in terms of its time (recall that name *now*). So if there is a lag in the occurrence of the desired act, the act is to that extent inconsistent with the volitional content, and, ex hypothesi, felt agency is reduced (that is, a certain passivity is experienced).

Admittedly, one may feel quite active in the process of trying to recall a name, for example. One does not necessarily just "wait," but may bring to mind various details and circumstances surrounding "that person I met at the party." If one has a strong sense of agency for these "exploratory thoughts," it's because they have fully satisfied volitions, including the representations of *when* the thoughts are to happen. But felt agency regarding exploratory thoughts does not entail that, once one thinks of the name, one will feel fully agentive about *that* mental act: it may feel like it just "popped into one's head." If it *does* feel that way, it's likely because, as I've argued, the thought fails to satisfy the temporal constraint of the volition. If it does *not* feel that way – i.e., one feels as much control over thinking the name as over the exploratory thoughts – it's likely because one had initiated a *new* volition to recall the name (consciously or not). The act then satisfied that volition, including its temporal constraint.[126]

Note that a delay following a volition to perform a mental act may make one aware of the fact that the volition is causally *insufficient* for the act to occur at the desired time, and the role of nonconscious cognitive factors in "delivering" the content then becomes salient without needing to be considered deliberately. Wegner's exclusivity criterion is also breached here, allowing a

[126] Now, it's also possible that one felt rather passive about the recollection due to the lag and resulting inconsistency with the original volition's temporal constraint; nevertheless, one characterizes the experience in terms that imply robust agency: e.g., "I finally got it" instead of "It finally came to me." This choice of words may be more reflective of one's self-concept as an agent than of the phenomenology of the act. For example, if that concept includes the nonconscious cognition that "delivered" the content (see n. 127), one will still consider it correct to say "I finally got it," passivity feelings notwithstanding.

further explanation of the reduced sense of agency: If one becomes aware of a dependence on nonconscious cognition in order to, say, recall a name, one is becoming aware of *another cause* of recalling that name apart from the conscious will to recall it.[127] Wegner would consider this an "internal" competing cause: "Whenever we become aware of some cause of our action that lies inside ourselves but of which we were previously unconscious, we may lose some of the sense of will" (2002, p. 91). Of course, felt agency is automatically reduced once the act fails to fully agree with the volition due to the delay; the dependence on nonconscious processing need not become salient for phenomenal passivity to set in.

Thus, if "waiting for content to come" – a certain resistance encountered after catalysis – is the norm in cognition (as Strawson implies), then so is phenomenal passivity. But the antecedent is false: typically, a mental act follows promptly and smoothly after being catalyzed. Interestingly, the one case where Strawson *does* allow for felt agency in cognition is imagination, namely, mental image-making, like summoning the image of a giraffe. And that's due to the typical lack of resistance or delay after one wills to imagine something: "We are prone to experience [imagination] as action, as something we do intentionally, when it occurs (as it normally does) without any sort of resistance," he writes (p. 239).[128] Presumably, catalysis also occurs without phenomenal resistance, and that is why he considers it a genuine mental act and not a merely ballistic result one has to "wait for." I'm not sure this claim would be accurate: I

---

[127] Even if, metaphysically, nonconscious cognition is part of the mental self, psychologically, one tends to restrict the mental self (or at least its essence) to conscious cognition. Thus, when one realizes how a given mental act depends on nonconscious cognition, the sense of oneself as the cause of the act may well be weakened, in view of this "other cause." An alternate psychology would subsume both kinds of cognition as part of one causally active mental self, which enables the exclusivity criterion to be met: For example, even if I realize that recollecting a certain name depends little on my conscious primings, and happens only when suitable nonconscious processes transpire, I still judge there to be *one cause* of the recollection whenever it occurs: me.

[128] He further speculates, "It may be that the sense of intentional authorship arises merely from the resistlessness …," and essentially argues that we *shouldn't* feel control simply on that basis, that "glow of ease" (p. 8). Granted. More fundamentally, there must be will/act consistency: Suppose I work with both Sharon and Sally, and I'm thinking about how Sharon might handle a project. I might mistakenly think the name "Sally" quite readily and with no resistance, yet immediately feel a distinct lack of thought control.

can have trouble clearing my mind to solve a problem, or focusing on the problem, etc. But in any case, Strawson effectively implies that *only* mental imaging and catalysis occur resistance-free, and thus only these types of cognition can feel fully agentive. That's a highly questionable proposal.

8. Rosenthal's Account

Rosenthal (2002) advances a different explanation of the sense of agency than the one I am advocating. What he calls our "sense of free agency" I take to be the phenomenon I have been discussing: phenomenologically, "free" simply implies that the control over our actions is *not* experienced as exerted by someone or something else. As a result, we may also feel that we could have acted otherwise than we did, a certain "looseness" or spontaneity during voluntary action.[129] Actions do not seem this way because they seem uncaused, Rosenthal argues, as "we normally experience voluntary actions as caused by conscious volitions" (p. 219). As I've urged in sect. 3, that's not the *normal* experience: We typically feel agentive during action without being aware of a forgoing volition. Indeed, when there is explicit awareness of *what* causes one's voluntary act, the perceived causal agent may very well be *oneself*, not one's volition. Several authors have endorsed a phenomenology of agent causation.[130] But given Rosenthal's account of the sense of agency, we must be aware of causally active volitions prior to acting, and frequently so. On his view, that sense, for some act *a* of ours, is grounded in two more fundamental experiences: (i) the sense of *a* being caused by our volition to do *a* (which must therefore be a

[129] Metaphysically, "free" may imply that we in fact could have acted otherwise, as many free-will theorists hold.
[130] See, for example, Terry Horgan et al. (2003) and Carl Ginet (1990).

conscious state); and (ii) the sense of that volition itself being uncaused. He writes, "Actions seem to be free because the volitions that cause them seem, in turn, to be uncaused" (p. 219). Of course, (ii) surely misrepresents how things are, Rosenthal notes: Even if a volition has no introspectively apparent causal antecedents, it is brought about by the readiness potential, according to Libet's evidence. More proximately, it is likely caused by a lateralized readiness potential (LRP)[131]; more distally, it is caused by the agent's beliefs and desires. Nevertheless, (ii) essentially means that the volition feels spontaneous, and when coupled with (i), means that we do *a* by spontaneously choosing to, which leads us to feel agentive.

While this account seems plausible, I think it faces certain difficulties. First, suppose we *were* conscious of the volition's mental causes. Would our sense of agency in acting necessarily be diminished or negated? If I maneuver my car into a parking slot by a tree while being aware of my will to do so and a mental cause of that will, such as my predilection for shade, it seems I can still feel fully agentive in parking the car.[132] What *does* seem to diminish felt agency, as discussed in the previous section, is realizing that the volition has nonconscious – and particularly subpersonal – causal antecedents. We tend to marginalize such cognitive events within our concept of our mental self as compared to conscious states (whether correctly or not), and so there is the sense that something that is not essentially us is controlling our will. Hence the "folk" reaction to Libet's studies of the readiness potential, which is a (necessarily) nonconscious subpersonal event: "My brain knows what I will want to do before I do," i.e., a

[131] And we may posit a nonconscious volition supervening on that potential, as Rosenthal suggests (p. 218).
[132] Consistent with Rosenthal's approach, we might say that felt control is in this case explained by the predilection's feeling uncaused. But consider this thought experiment: Suppose that for each mental event *x* that we are conscious of as causing our action, we are aware of some other mental event *y* as causing *x*. That is, we are aware of an infinite regress of causal antecedents to our action at the mental level. For example, imagine indefinitely extending the causal chain < predilection for shade→volition to park the car by the tree > to include conscious reasons and desires that are progressively more fundamental. No mental state or event in this chain would feel uncaused, yet it seems that we would still feel that the action that is ultimately produced is mentally controlled by us.

sense of disempowerment. But there is no reason our sense of agency should be diminished when we are introspectively aware of the mental causes of our volitions, as we generally take such conscious states to be essential to our mental selves. We would simply be more deeply aware of our motivations for acting in the course of acting.[133]

Second, when volitions are conscious, we usually feel in control of them as well as the ensuing act. But on Rosenthal's view, we would need a different account of felt agency for volitions than for acts: We feel in control of acts, he claims, when we feel they are caused by volitions that feel uncaused. So we cannot give the *same* explanation of why we feel in control of those volitions: they feel uncaused, and thus cannot feel caused by volitions that feel uncaused. This problem is avoided on a volition/act consistency approach. A person feels in control of both act and volition for the same type of reason: The act feels that way because it agrees with the volition (which is in this case conscious), and the volition feels that way because it agrees with more general volitions and desires (which may be nonconscious[134]). We can thus give a uniform account of felt agency for both bodily actions and thoughts.

Third, consider volitions that we don't feel control over, such as a pathological urge that we feel causes us to act. We may, at least at the time of acting, not be aware of the mental causes of the volition: it seems to just "pop into" our consciousness, and move us to act. So both (i) and (ii) are met, yet we feel agency in neither willing nor acting. The act cannot feel agentive, since it issues from a volition that is unwanted. So the volition's seeming uncaused is irrelevant to felt agency: whether it agrees with our second-order volitions is key. Rosenthal has dismissed this as

---

[133] Such action probably would not feel spontaneous, but it *would* feel agentive, on the consistency-based approach I favor. Indeed, these are distinct phenomena. So we might reserve Rosenthal's account for the feeling of spontaneity, maintaining that an act feels (i) fully agentive when it agrees with a volition that itself agrees with a second-order volition; and (ii) spontaneous when we're unaware of that second-order volition, or any other mental causes of the first-order volition.

[134] Contra Rosenthal's claim that "nonconscious volitions seem irrelevant to the sense of free agency" (p. 219). On his approach, of course, they *are* irrelevant, since he requires felt agency to be grounded in conscious volitions.

a counterexample to his account, claiming that when we experience an intrusive mental state, we *do* tend to have "a sense that it is caused, and a sense of what causes it."[135] But to say we tend to is to allow for exceptions. Indeed, the precise phenomenology of intrusive thoughts likely varies on a case-by-case basis. And if unawareness of the causes of a volition is sufficient for felt agency in acting, there shouldn't be *any* exceptions. That is, there shouldn't be any cases where we are unaware of those causes and feel nonagentive, for that is to falsify the theory.

9. The Function of Felt Agency

While Rosenthal denies that nonconscious volitions can ground the feeling of control over our acts, he does allow that they can provide nonconscious information about agency, which in turn can "play a role in our psychological lives, even if not consciously" (2008, p. 834). I follow him on this hypothesis, although I don't think volitions *in and of themselves* provide information about one's controlling a given act. As argued, a volition enables a consistency assessment, which then causes a belief regarding agency, each of these mental events being typically nonconscious. Reaching this belief, then, depends on volitional content. So a volition certainly provides information that is *relevant* to our belief in agency. But one can't "read off" one's agency from a volition, as the ensuing act may not have satisfied it. It's also the act perception, along with the judgments Wegner identifies, that yield agency information. The information will have been "provided" once one believes in and feels agency on the basis of the judgments.

[135] In correspondence.

How accurate is the information? It depends. As Wegner suggests, “the experience of will” (a term he sometimes uses to designate the sense of agency) “can be an indication that the mind is causing action, especially if the person is a good self-interpreter, but it is not conclusive” (2002, p. 96). That is, if the person judges the conditions to obtain, there is a good chance that the person’s mind *did* cause the action, and so the feeling of agency that results from those judgments becomes a fairly reliable indicator of action-control. But the judgments do not of course *guarantee* that the mind caused the action. First, for Wegner, they would never entail that one’s volition caused the action, insofar as he holds that volitions are epiphenomenal to the neural processes that *do* generate action (e.g., the readiness potential). At best, judgments that priority, consistency, and exclusivity obtain would be correlated with the control of the action by those neural processes, which he considers part of the nonconscious mind. As discussed in sect. 2, I differ with Wegner on this point: I think volitions can produce action. Second, there are cases where a person judges the conditions to hold and her motor mechanisms have *not* caused the action. Primarily such cases can be cited when the action is a remote effect as opposed to a bodily movement, as there is then a greater chance for something unnoticed by the person to be producing the action. Thus Wegner gives the simple example of his playing a videogame he was unfamiliar with (Donkey Kong), and feeling control over the monkey’s leaping over barrels, even though he had actually been “playing” during a pre-game demo (2002, pp. 9-10). Presumably, he had judged that the leaps succeeded his will, were consistent with it, and had no other causes. But his resulting sense of agency, qua those remote events, did *not* correctly indicate that his motor system was a cause of the monkey’s leaps. Still, the agency he felt qua his bodily movements (gripping and moving his hand to control the joystick) *was* accurate: his motor system did control his hand.

Thus, it seems felt agency would more reliably indicate voluntary bodily movement than voluntary remote action, insofar as the judgments grounding it are more reliable in the former case. In particular, one is well positioned to make a correct exclusivity judgment, i.e., whether there are exogenous causes of one's bodily movement. It would also seem that the consistency judgment is on more solid ground: one is more likely to correctly judge that intentions are satisfied by bodily movements than by distal events, especially given the added proprioceptive input from movement. But in fact our conscious awareness of our own movements (one of the two data that Wegner says is used in the consistency judgment), may not always be accurate, given its source. There is evidence that the conscious representation is formed *prior* to actual movement. For example, an experiment by Haggard and Magno (1999) using transcranial magnetic stimulation (TMS) implicated a stage in movement *preparation* – namely, supplementary motor area (SMA) processing – as the source of conscious awareness of movement. Using TMS-induced delays of a key press made in response to a go signal, they found that when TMS was applied to the SMA, subjects showed greater awareness of their slower reaction time than when TMS was applied to the primary motor cortex. Now as discussed, the comparator mechanism receives afferent input from the *actual movement* in making its determination (ex hypothesi), so its output more reliably indicates movement control. For Wegner, this discrepancy between the information used for a consistency judgment regarding bodily movement and the information used by the comparator would be more evidence that "apparent mental causation is generated by an interpretive process that is fundamentally separate from the mechanistic processes of real mental causation" (2002, p. 96). But if, as I argue, felt agency is regularly generated by a *nonconscious* consistency assessment that supervenes on a comparator match, the phenomenon would accrue more reliability as a

movement-control indicator. There would be less separation between the "interpretative" and "mechanistic" processes, to use Wegner's terms – at least when it comes to bodily movement. That such an assessment takes into account input from the actual movement, and not a preparatory stage, does not contravene the Haggard and Magno result, since their experiment concerns *conscious* representation of movement.

Given the general reliability of our agency indicators, it might be thought that the function of felt agency is to yield whatever psychological effects result from one's being informed about action control. But again, these effects could just as well result from a (typically) nonconscious state: the belief in agency. And in fact, it's more plausible that they do result from that state. Consider that the most salient effects of agency information would be on action planning: If I am concurrently performing acts *a* and *b*, and believe that I am in full control of *a* but weak control of *b*, I may well decide to devote more attention to controlling *b*. Such decision-making regarding action is one of the functions commonly ascribed to the *central executive* (CE), a postulated cortical system that draws on rational capacities and memory in order to handle nonroutine behavioral tasks.[136] It is the CE, then, that is "informed" about the state of action control, and I argue that the belief in agency, as opposed to the feeling, constitutes the CE's reception of that information and affects its processing accordingly. That's because the belief likely occurs *in* the cortical regions of the CE itself (primarily dorsolateral prefrontal and anterior cingulate), whereas the feeling likely occurs in the brain's subcortical limbic pathways. Once the belief occurs, it both affects action planning and engenders felt agency. It is less plausible to assume that the causal chain runs the other way; namely, that the CE depends on phenomenal agency to occur – in a separate neural system – in order to reach a belief state on

[136] In D.A. Norman and T. Shallice's (1986) model of action control, the CE is essentially the Supervisory System, which modulates the "lower level" of control by weighting action schemas, thereby altering "automatic" responses to stimuli.

action control. Reaching such a state is a function integral to the CE, as a rational-control system that receives direct input from the premotor system, which presumably subvenes volitions, and perceptual systems that track one's bodily movements and distal acts.

Thus, while both the belief in agency and felt agency represent some degree of action control, only the former on my proposal plays a psychological role qua its representational content, by occurring in and influencing the CE. One may object to this proposal via the following case: Suppose Jones is trying to perfect his golf swing, and in the course of attempting a practice swing, he feels fully in control of the turning of his hips, but poorly in control of the trajectory of the club. Naturally, the latter feeling prompts Jones to focus more on achieving the proper trajectory, on the next swing or perhaps *as* he feels that sense of poor control during the swing. This may entail his bringing to mind a coach's advice, picturing where he wants the club head to be at the end of the swing, etc. Clearly, the weakened *sense* of agency is moving Jones to take these measures in respect of what it represents: poor control over his swing. In response to this putative counterexample, I would note that, ex hypothesi, Jones' feeling a lack of control over his swing is caused by his *belief* that his club is poorly controlled. And it's that belief, combined with certain of Jones' standing desires and beliefs (a desire to have a good swing, a belief that club trajectory is vital, etc.), that moves Jones to devote more attention to the club's trajectory. But since the belief occurs nonconsciously (and immediately upon Jones' motor system detecting a discrepancy between his volition and movement), Jones will tend to think it was the feeling that moved him to take those measures: for he cannot, of course, introspect a nonconscious belief.

If we only need agency beliefs and not feelings in order to make such decisions about the allocation of attention, why then do we feel control over acts (or lack thereof)? While this state,

on my view, is an effect of the process by which we come to an agency belief, I don't think it's epiphenomenal, for it plausibly has its own psychological effects, based on its *qualitative* content, which we may call "control qualia." We are generally drawn to pursue good feelings and avoid bad ones, and the feeling of agency is a positive one of efficacy and harmony during action, while its counterpart is one of impotency and frustration. Thus, successful action, insofar as it entails the feeling of control, is intrinsically pleasurable. As Jonathan Cole and Barbara Montero maintain, movement "is not simply for long-term reward in terms of satiation of hunger, thirst, cold etc. It has some reward of itself" (2007, p. 299). While they suggest that this reward may be "affective proprioception" during movement, mediated by "a system of sensory afferents," they also propose that the intrinsic pleasure may be due to volition/act agreement, the "harmony between intention, action and sensory return," which, as I have argued, grounds the sense of agency. In turn, the pleasure associated with felt agency will drive one to seek to fulfill one's volitions. This psychology fits with evolutionary theory: pleasurable sensations often issue from behaviors that are beneficial to the organism (e.g., eating and rest), while painful ones issue from those that are not (e.g., exposing the body to fire). So because well-controlled voluntary action is generally more conducive to survival than poorly controlled action, we can expect positive and negative sensations to be linked to them respectively. "It is relatively easy to suggest that there was a good evolutionary pressure to reward exercise of the body (in the sense of controlled movement)," write Cole and Montero. Of course, the pleasure involved in feeling movement control will seldom be the *sole* motivation in action choice; many other predicted results of the act are factored in, and usually take precedence. For example, a perfectly executed walk along a high ledge would engender the pleasure of control qualia, but the promise of that sensation alone would motivate few people to attempt the walk.[137]

[137] Infants, who from about two months of age onwards display voluntary movement, may present an example of

Assuming that felt agency is psychologically efficacious, do we have a case of consciousness being efficacious? After all, we would have a causally active conscious mental event. But as discussed in Chapter 2, the key issue is whether said mental event affects cognition *qua being conscious*. Surely, if the sense of agency draws one to seek volition satisfaction (in addition to the desirability of any goals that may be thereby attained), it is in virtue of the pleasing qualitative character of that experience. Qualitative character, however, is not the same phenomenon as consciousness: the notion of conscious qualia is not redundant, for there can be nonconscious qualia. At least this is how things stand if one subscribes to a higher-order theory of consciousness, as I do. On that type of view, control qualia are *conscious* in virtue of being the intentional objects of another mental state, such as a thought or quasi-perception; they are *qualitative* in virtue of being introspectible properties – harmonious, effortless, powerful, etc. – of an affective state.[138] Given that distinction, it is (logically) possible for the c-property of felt agency to be inefficacious while its qualitative character is not. For example, control qualia would have the psychological effects proposed above, while their *being the object of a HOR* is

---

moving simply for the sake of pleasure associated with felt agency. As Cole and Montero note, they seem to enjoy "exploring carefree movements" (2007, p. 310).

[138] Further, we can define token qualitative states via Rosenthal's homomorphism theory (2005, pt. 2). Briefly, the identity of a qualitative state in modality *M* derives from its position in a mental "quality space" that is homomorphic to the quality space of perceptible properties accessible by *M*. So beige appears to resemble brown more than green, and this mental relation would parallel resemblances among the family of visually perceptible properties of beige, brown, and green objects. Following this view, a control quale would be fixed by its position in an affective quality space, e.g., more like a feeling of harmony and less like a sense of effort. And an account might be developed where those mental relations parallel similarities and differences between pairs of agent/environment relations. In addition, a control quale could be defined functionally, in a way that does not involve HORs. So a nonconscious pain would tend to cause aversive behavior, and a nonconscious control quale would tend to cause a desire to repeat the controlled action, or to perform similar acts.

Unlike perceptual states, however, there may not be any actual cases of nonconscious pain or nonconscious control qualia. But they are possible on higher-order theory. For example, a nonconscious control quale might be created by having a person play a familiar piano piece while blocking any HORs that would target the subject's quale. Theoretically, this could be done by applying TMS to increase inhibition in the relevant area of the prefrontal cortex, and checking that the subject reports no feeling of control over his playing. If the activity in his limbic system remains constant, it's plausible that a nonconscious control quale is occurring.

an epiphenomenal property. This would be the case if the HOR plays no causal role qua its h-property.

But assuming that we don't have a case of what higher-order thought theorists call "empty heat" (a HOR without its intentional object), doesn't a HOR play a causal role insofar as it provides information about its target? All representational states (if they don't misrepresent) provide such information: Suppose a person *S* consciously sees a lamp. *S*'s perception informs *S* of a lamp's presence, while *S*'s HOR of that perception informs *S* of *his seeing* a lamp. And to say *S* receives information about *x* is to say one or more of his neural systems have received that information, namely those that have entered into an *x*-representing state. So suppose *S*'s visual cortex receives information about the presence of a lamp in a certain location by entering into state *p*, and *S*'s prefrontal cortex then receives information about *p* by entering into HOR *p´*. Neither state plays a causal role, however, simply in virtue of providing information to their respective brain areas: the information must be *used*, or have effects of its own. That is, *p*'s occurrence *constitutes* the visual cortex's reception of lamp information, it does not *cause* it. Similarly for *p´*: its presence in the prefrontal cortex *is* that brain system's being informed about *p*. But if *p´* does not in turn have mental effects, it's epiphenomenal. Thus, when HOP theorist Lycan writes of an "internal scanner" state "delivering information about … [a first-order] psychological state to one's executive control unit" (1987, p. 72), he is not thereby citing a causal role for a HOP; the delivered information must be put to use.

In the case of felt agency accompanying some act *a*, the HOR that renders its qualia conscious would have the content *I am feeling control over* a. We can assume that this state occurs in some part of the CE. Rosenthal, for example, cites the mid-dorsolateral prefrontal cortex as a plausible locus for HOTs, given the correlation between certain activity there and

consciousness (2008, p. 835). The first-order target state – the control quale – likely occurs in subcortical limbic pathways insofar as it is a feeling, as suggested above. The occurrence of the HOR then constitutes the CE's being informed about a feeling of control in a distinct neural system. But as argued, that HOR would only play a causal role if its content – *I am feeling control over act* a – had some effect on processing in the CE. We would also look for a beneficial and significant causal impact, if we are to claim that this particular type of HOR has *utility* as opposed to mere efficacy.[139] My proposal is that the HOR enables the CE to take the qualia into account in its rational processing, insofar as the HOR relays conceptual information about the qualia's presence. That is what it does qua its h-property, qua representing the first-order state. Thus, HORs about (positive or negative) agency feelings are a necessary condition for thinking and reasoning about the qualia.

Let me address some likely objections to this idea:

(i) *A HOR need not represent the pleasant or frustrating nature of the feeling in order to render it conscious. For example, on HOT theory, the thought that 'I am feeling control over act* a*' is about that feeling despite not expressing its pleasantness in particular; indeed, that theory requires only that HOTs deploy primitive concepts of mentality, which need not pick out every aspect of a first-order state. Thus, an agency feeling's being conscious does not entail that information about its positive or negative affectivity reaches the CE.* The problem with this counterargument is that our ongoing consciousness of control – and occasional consciousness of lack of control – *do* include consciousness of positive and negative affectivity, respectively. So whatever HORs are rendering those feelings conscious must in fact be representing their

[139] See my definition of "utility" in Ch. 3, sect. 2.

pleasantness or frustrating qualities, even if they need not. Furthermore, since these emotions during action can be felt by infants and animals (given the evolutionary account of control qualia I have proposed), no especially sophisticated mental concepts would be required in order to pick them out via higher-order representation.[140]

(ii) *If the impact on the CE is mediated by the HOR, why would the target feeling be needed? Desires and behavior could be driven simply by certain HORs about control feelings. So it becomes difficult to explain the evolution of first-order emotional qualia.* It's unlikely that such HORs would occur without their target qualia, just as first-order perceptions don't (typically) occur without the external objects they represent. From an evolutionary standpoint, it's generally advantageous that mental states don't misrepresent, that they be nomologically tied to the occurrence of their objects.

(iii) *A causal interaction among first-order states would be sufficient for CE processing to be impacted by the agency feeling. For example, if the qualia cause the desire to continue a particular movement, or to seek controlled movement in general, only a direct causal link is needed between the qualia and the desire – both of these being first-order states.* This objection is perhaps stronger than the forgoing ones, and parallels Rosenthal's argument that executive control of first-order states need not involve HOTs about them, but only suitable causal

[140] And surely no *linguistic* concepts: Assuming that infants (under nine months or so) and animals have conscious states, their higher-order states must deploy concepts that are primitive in the sense of not corresponding to a language. These nonlinguistic concepts would presumably be fine-grained enough to pick out the qualitative properties of first-order states, such as degrees of pleasantness or frustration. (Recognitional ability could serve as a criterion for possession of a nonlinguistic concept. In this case, if the infant or animal could recognize the same pleasantness in feeling control across many different successful voluntary movements, it has some concept of that quale.)

connections between first-order states (2008, p. 835).[141] It suggests that a HOR about a feeling of frustrated effort would be *epiphenomenal* to the causal chain running from the feeling (occurring downstream in the limbic system) to the desire to discontinue the behavior that results in that feeling (occurring in the CE). The CE, then, does not need to be "informed" about the state of the limbic system – at least not through meta-representation – in order for its processing to be impacted by that state. The "informing" may involve no more than the qualia causing certain behavioral desires. That is, the mental causation may be entirely first-order, with control qualia *directly* resulting in the desire to do what will make that experience of agency recur (i.e., a certain act); and mutatis mutandis for the feeling of lack of control. In response to (iii), I will argue that if a control quale directly causes a specific behavioral desire, *intervention* by reasoning must still be possible so that one can flexibly respond to the qualitative experience. And that rational control, I claim, would need to deploy a representation of the quale, i.e., a HOR.

Suppose that the control quale that arises from my juggling causes me to desire to continue juggling. If (a) the quale is causally *sufficient* for the desire and (b) the causation is unmediated by mental states following the quale, then the HOR targeting the quale could not play a causal role in producing the desire. But when the cognitive process from quale to desire is rational, neither (a) nor (b) is the case. There would be a state that causes the desire along with the quale, as well as mediating states in the causal process. The co-cause of my desire to continue juggling would be my standing desire (of some strength) to make control qualia occur.[142] The mediating states would be: my belief that my current qualitative state is of the

[141] Armstrong and Lycan, on the other hand, do think that higher-order states (perceptions, in their case) serve to "integrate" first-order ones. I will assess this proposal in Ch. 5.

[142] This desire itself is arguably grounded in what I will call a valuation state. Experiences of control are positively valued (to some degree) while experiences of frustrated effort are negatively valued (to some degree). (There is

control type; my desire to make that state recur[143]; my desire to do what will make it recur; and my belief that juggling causes my qualitative state. Following these intervening states, which constitute a rational process, I would desire to continue juggling. Now, the desire to make control qualia occur is metacognitive, but it would not make any qualitative state conscious, according to higher-order theory: The state is certainly about control qualia, but does not represent that one is in a control-qualitative state.[144] The mediating states, however, do contain such a representation. Thus, each entails that the quale is conscious. To be specific, that metacognitive content is *my feeling of control*, which is used in a different predication in each intervening state. The content can also be expressed as "The control that I am feeling." On HOT theory, that's the kind of assertoric content required for a state to make one aware of being in another state, i.e., to be in a conscious state. We might then distinguish between a "simple" HOR, which is just a representation that one is in a certain mental state, and "embedded" HOR, which is a HOR used in making a predication about its object-state; for example, *I desire to do what will make the control that I am feeling recur*, although a desire, contains the assertoric content *I am feeling control*. An embedded HOR is thus a simple one that is used in ascribing a property to one's

---

nothing intrinsic about their qualitative character that entails they should be so valued.) These valuations may change over time; for example, it seems that infants like the feeling of control that comes with simple movements more than adults do. The valuation system is thus a causal factor determining whether (and to what degree) one will desire to experience control. There is evidence that this system is located in the perigenual anterior cingulate cortex (ACC). According to G. Bush, the perigenual ACC has a role in "processing affective/emotional information, including assigning emotional valence to internal and external stimuli" (2004, p. 208). In psychology, "valence" means emotional force or significance, a measure of how attracting or repelling a stimulus or experience is. Given the perigenual ACC's connections to both lower-level systems that likely subserve the sense of control (motor and limbic) and higher-level systems comprising the CE (dorsal ACC and dorsolateral PFC), its valuation states may well causally mediate between control feelings and the desire to make them occur. Moreover, valuing a token emotive state (such as a particular control quale) would require higher-order representation of that state and its being conscious, as should become clear with my ensuing argument.

[143] Alternatively, the desirability of the quale would not be inferred from its type: Following the experience, one could desire to make *that* state recur. This may be the nature of the cognition when the affective experience is new and difficult to categorize.

[144] The state would be an example of what Kriegel (2005) calls "trait metacognition," which is the representing of a "standing feature" of one's mental life, in this case control qualia. What higher-order theorists think is needed for consciousness is "state metacognition," which is the representing of "an occurrent, local event" in one's mental life.

mental state, beyond the property of one's being in that state. In the forgoing example, I ascribe the property *being a state that I desire to try to make recur* to my feeling of control.

This sort of metacognition utilizes the h-property of the simple HOR – the property of *representing a control quale*. But is *consciousness of control* thereby used in the metacognition? Based on my discussion in Chapter 2, sect. 2, the answer is affirmative if we assume either of two views: (i) The quale's c-property is the complement to the h-property; or (ii) the h-property is sufficient for conscious experience. Given (i), the HOR's representing the quale is a relation between the states, and so the efficacy of the HOR in virtue of *representing the quale* metaphysically entails the efficacy of quale in virtue of *being represented by the HOR*, i.e., its being conscious. Given (ii), the HOR's representing the quale – though it does not constitute a relation with the first-order state – *is* consciousness of oneself as being in control. The efficacy of the h-property is then the efficacy of that phenomenon, whether or not a first-order qualitative state occurs. But it is also possible to hold a view that denies (i) and (ii). First, contra (i), we may hold that the c-property of the quale is merely the property *accompanying the HOR's instantiation of the h-property*, while the h-property is some set of (perhaps teleo-functional) properties that do not involve the quale. It follows that the HOR can play a causal role in a metacognitive process *p* qua its h-property but not qua *being accompanied by the quale*. And so the quale will not cause *p* qua *accompanying the HOR*, i.e., the quale's c-property. Second, contra (ii), we may hold that empty HORs do not yield consciousness, so the quale's occurrence is a necessary condition for consciousness of control, along with the targeting of the quale by the HOR.

Given the tenability of ¬ (i) and ¬ (ii), it seems the causal relevance of h-properties to metacognition concerning control qualia does not entail that *consciousness of control* plays a role

in metacognition of control qualia, whether *consciousness of control* is understood as the c-property of an existing quale or a phenomenal property of a HOR. The entailment fails if ¬ (i), because the h-property is then not a relation to the quale, and it fails if ¬ (ii) because there is no such phenomenal property supervening on (or identical to) the h-property: the HOR does not have the property of *making one feel in control*, simply in virtue of representing the quale. But the causal relevance to metacognition of the c-property of a control quale can be defended without either entailment. The h-property of the HOR is clearly necessary to one's reasoning about how to react to the experience: one needs to represent the quale to think about it. And presumably it is not a random matter whether a HOR occurs in the absence of a target state. Certain psychological mechanisms will usually prevent this from happening, and under those conditions, a control quale is necessary for one to represent oneself as being in the target state and having further thoughts about that state. Under these circumstances, both the quale and the HOR (qua its h-property) cause the metacognition. Equivalently, we can say the quale brings about the metacognition in virtue of *accompanying the instantiating of the h-property by the HOR*. And the property of "accompanying the … by the HOR" is, on the view being considered, the c-property of the quale.

Admittedly, one usually is unaware of an explicit reasoning process leading from an emotional quale to a specific behavioral desire.[145] So the desire to continue juggling may immediately follow the control quale in my stream of consciousness. But the intervening intentional states, those that include representations of the quale, would be nonconscious in that case. Consider that simple reasoning steps (what psychologists have called "automatic

[145] By "specific behavioral desire" I do not mean the desire to do what will make the quale recur; rather, I mean the behavioral desire that does *not* represent the act relative to the quale, e.g., the desire to juggle. The specific behavioral desire, then, does not contain higher-order content.

inferences") often remain implicit,[146] as illustrated by enthymemes. One may consciously think that a certain precious stone will not scratch glass, and consciously conclude that it is not a diamond, while only implicitly thinking that if the stone is a diamond, it will scratch glass. Of course, if the desire to continue juggling immediately follows the control quale in my stream of consciousness, I'm not aware of *any* reasoning steps, including those states with higher-order content. But as Rosenthal (1997) has argued, second-order states are only conscious in the relatively rare cases when we introspect, via third-order states. And furthermore, second-order states need not be conscious themselves in order to confer consciousness on first-order states. Nor does the very brief time interval in which the rational process would need to occur (between the quale and the specific behavioral desire) make its occurrence implausible: "Thought is quick," as Hobbes observed (1994, p. 12), and especially so when it is basic pragmatic reasoning, as in the case I'm describing.

One might concede that such a rational process takes place, but argue that the representation of the quale deployed in the mediating states need not be the type of meta-representation that would make the quale conscious. After all, meta-representations can be about *past* mental states. So I can desire that the quale that I *just experienced* recur, believe that proficient juggling caused the quale that I *just experienced*, etc. But one will be aware of being in a mental state only if one thinks about being in it (or perceives that one is in it, on HOP theory), not if one thinks about *having been* in it. As the control quale that accompanies juggling is presumably ongoing, the states in the rational process would include representations of a *current* state, entailing its consciousness. But in the case of a control quale that is correlated with an act that begins and ends quickly, like throwing a dart, it may be that the reasoning represents the quale as a past state. The nature of the cognition would then be as follows: I am in a control-

[146] See, e.g., B.J. Baars (1997), Table 1.

qualitative state at time t_1, and represent that I am (since it's a conscious state). At t_2, I desire for that quale to recur, but this is a desire that *the feeling of control I had* recur. Subsequently, I reason about how to fulfill that desire. Thus, the consciousness-conferring HOR at t_1, unlike the embedded HOR at t_2, is not used in the cognition. (Both HORs would be "used" in the relevant way – the way that matters to conscious causation – if both are causally relevant to the cognition qua their h-properties.)

While we can't rule out this kind of processing scenario a priori,[147] I think the following argument supports the utility of the HOR at t_1: First, as discussed, reasoning to the specific behavioral desire is typically nonconscious, and nonconscious reasoning, while perhaps limited to routine pragmatic inferences, is generally held to be faster than conscious reasoning.[148] For example, with regard to inferring speaker meaning, Paul Grice has maintained that "We have... a 'hard way' of making inferential moves; [a] laborious, step-by-step procedure [which] consumes time and energy... . A substitute for the hard way, the quick way, ... made possible by habituation and intention, is [also] available to us" (2001, p. 17). That "quick way," he held, is typically nonconscious. Its speed relative to conscious inference can be illustrated by the everyday process of interpreting speaker meaning: Bob comes into the office and utters the sentence 'It's raining' to Jim, who instantly – and consciously – believes him to mean that it's raining. Yet certain

[147] One attempt, based on the nature of memory, might go as follows: The embedded HOR at t_2 is a *memory* of a mental event. And a memory of an event or state of affairs *e* is (normally) caused by *e* itself, via a mental state that represents *e* at the time *e* occurs. For example, my memory at t_2 that there *were* keys on the table is caused by the keys' being there at t_1. But clearly this happens insofar as, at t_1, the keys' being there caused me to perceive that they *are* there, which perception in turn produces the memory. Similarly, my representing at t_2 that I *was* feeling control must be caused by that quale via a representation at t_1 of my feeling control *now*. So a HOR of a past state, and any reasoning involving it, depends upon a prior HOR of that state as occurrent. The problem with this analogical argument is that the causation in the second case is entirely intra-mental: the memory is mental, and so is the event it represents (the quale). So when that event happens, it need cause no immediate mental representation of itself at t_1 to ensure the memory obtains at t_2; the extra-mental state of affairs (the keys' being on the table), in contrast, must so affect the mind at the time it occurs if it is to lead to a memory.

[148] For case studies, see Wegner (2002), pp. 56-59. Now, some psychologists deny that the kind of pragmatic inferences that go on beneath the threshold of awareness count as reasoning, considering them merely "heuristic." This issue does not concern me in the present context, as the utility of HORs to mental processes that are "something like" reasoning still entails they have a mental function.

habitual nonconscious inferences occurred even faster than the conscious belief: that the sentence standardly means that it's raining, and that Bob, like most people, intends the standard meaning by his utterances. Jim might also instantly believe that it's raining, prior to which he would (nonconsciously) think that Bob is capable of knowing whether it's raining, has no reason to deceive him in this context, etc.[149]

Second, note that the HOR at t_1 enables the *most immediate* reasoning about how one will react to the quale; such reasoning would take place while the quale still obtains. So, it's plausible that the cognition leading to the specific behavioral desire, insofar as it is a case of fast, nonconscious inference, deploys the HOR at t_1. This means that *before* the brief control quale resulting from the dart-throw elapses, I would think: I want *this* state I am in to recur; I want to do what will make it recur; I believe that my dart-throw caused *this* state. Now, even though one can run through such thoughts quite quickly, it may seem improbable that one could do so as quickly as the example implies. But of course, when we think about how fast *it seems* we can run through the thoughts, we're considering how we *consciously* run through them. So that consideration cannot legitimate judgments about the speed of implicit reasoning. Moreover, instantaneous pragmatic reasoning about one's response to an affective state can be advantageous: consider a pain one must try to relieve as soon as one feels it. And for that ability, I argue, the mind needs to deploy a representation of the state as occurrent. A parallel suggests itself with the representation of external events: The hunter who sees a flash of his quarry in a bush and immediately shoots at it presumably engages in a few implicit reasoning steps just prior to shooting; such states will be fastest if they represent the prey's brief appearance as occurrent.

[149] One explanation for the relative slowness of conscious processing is its (proposed) correlation with "global broadcasting," or the relay of information to various systems that handle memorization, association, emotional response, verbal reporting, etc. (See Baars [1988].) It is natural to suppose that the broadcasting would require additional milliseconds (automatic, nonconscious reactions have been timed at 200-300 ms; conscious ones at about 500. See Wegner [2002], p. 57).

In addition, the rational processing of a control quale allows for the rational *control* of behavioral desires that result from it: For example, if I did *not* desire that the quale recur, or did *not* believe that the quale was caused by my juggling, the desire to continue juggling wouldn't arise. Regarding the first case, I may develop a Frankfurtian second-order desire[150] that I not want the quale to recur, because I find myself getting addicted to juggling. This second-order desire may eventually weaken, if not remove, the first-order one, thus resulting in my not wanting to juggle. Regarding the second case, suppose I knew that neuroscientists had stimulated my limbic system to produce that control quale peculiar to successful juggling: I would not then believe juggling caused it, and so not want to persist in the act. Rational control of one's response to a qualitative experience is thus advantageous.

Now that I have described how HORs enable the rational processing of control qualia, I can better address objection (iii). Someone who puts forth (iii) might concede that a rational process mediating between a control quale and specific behavioral desire requires a representation of that quale. She might also concede that when the process is fast and nonconscious, it likely deploys a representation of the quale as occurrent. Nevertheless, she would deny that such reasoning is needed to mediate the causation at all. Why can't the causation be direct? First, note that in order for direct causation to be plausible, we would have to posit a qualitatively distinct sense of control accompanying each successful voluntary action. So there must be a control quale peculiar to juggling, separable from the visual, tactile, and proprioceptive experience of juggling. A *general* feeling of control could not be causally sufficient for a desire to continue juggling, as that same feeling, when it accompanies successful cycling, would cause one to want to continue juggling – which is absurd. So let us grant that a

[150] This is a desire about a first-order desire, namely that the first-order desire guide (or not guide) one's behavior. Harry Frankfurt (1997) discusses the concept.

distinct juggling-control quale is sufficient to directly cause the specific behavioral desire. In that case, even if I *lack* the desire to make that state recur, the desire to do what will make it recur, or the belief that juggling causes my qualitative state, I will want to juggle. Perhaps it is plausible that the causation could occur in mere absence of these intentional states. But surely we must allow that holding desires and beliefs whose content is *contradictory* to those states would (tend to) stop the specific behavioral desire from arising. That is, suppose I do not simply lack the belief that juggling causes the quale, but believe that something else does (e.g., the neuroscientists who are stimulating my limbic system to produce the quale). Similarly, suppose I do not merely fail to desire my quale's recurrence, but desire that it *not* recur (for some reason). As argued, such states represent the quale, and entail its being conscious. And if we assume that (i) they have the causal power described (indeed, to deny they do conflicts with their functional identity); and (ii) a mental state's causal powers are determined by its content; then the state's representations of the quale (the HORs) are required for that power. For example, if the desire that the quale not recur failed to be about that quale, it would be a desire that something *else* not recur. As such, it would not work inferentially with the belief that juggling is causing the quale to block the desire to juggle. In short, it is plausible that rational control can *intervene* in any case where the causation from quale to specific behavioral desire is direct, as this allows cognitive flexibility: One's beliefs and/or desires about that quale may call for responding to it differently.

I should add that while an emotive state's being reasoned about requires higher-order representation, the converse does not hold: simple HORs can occur, and render their target states conscious. So the fact that control qualia are regularly conscious does not imply that each token state is reasoned about. For some agents, only more salient agency feelings (e.g., those that issue from successful juggling) may engender a desire for recurrence and the cognition following that

desire, but not less salient ones (e.g., those that issue from successful basic movements, like lifting a pen). Still, both feelings would, ex hypothesi, be the object of HORs. Note that the simple HORs do make their content *available* for use in immediate reasoning about that state. Similarly, not all first-order representations (FORs) are used in reasoning about their objects: one continually perceives a multitude of features of one's surroundings, though not all (perhaps relatively few) of these representations inform one's action-oriented reasoning. But any token FOR is the consequence of the ongoing operation of a first-order representational system, which usefully generates environmental information *in case* it may be selected for further processing. Ongoing representation of affective states, such as control qualia, is generally useful in same way, even though it sometimes generates unused higher-order information. Now one may argue, based on dispositionalist higher-order theory, that while consciousness does enable reasoning about first-order states, HORs needn't regularly occur for that purpose. Only the FORs' access to an HOR *system* is needed. I address this position, advanced by Carruthers, in Chapter 5.

10. Conclusion

In this chapter I have argued several points regarding (i) the role of consciousness in producing the feeling of agency, and (ii) how consciousness contributes to the function of felt agency. Let me briefly review the main ones. As to (i), the role of consciousness in producing the feeling is constitutive, not causal. Felt agency is a conscious phenomenon, but it originates from (typically) nonconscious mental states, including volitions and the judgments of priority,

consistency, and exclusivity that Wegner identifies. Since these states *need not* be conscious to generate the feeling, when they are conscious, they aren't generating it qua being conscious.

As to (ii), I've argued, based on higher-order theory, that felt agency is constituted by a control quale targeted by a HOR, which renders the quale conscious. The function of control qualia, as a pleasurable psychological state, is to move one to seek well-controlled action, which is generally advantageous from an evolutionary perspective. For this "moving" to occur is for a control-quale to affect CE's action-oriented decision-making, in particular to cause a desire (of some strength) for that quale's recurrence. I've argued that this causation is mediated by fast, usually nonconscious pragmatic inference, and that the quale's c-property is essential to that reasoning. A representation of one's being in a control-qualitative state is utilized in immediate reasoning about how one will respond to that experience. So the function of consciousness, with regard to feelings of control, is to enable rational processing of such qualia at the moment they occur. In the next chapter I expand on this approach to the function of consciousness, arguing that HORs also subserve useful reasoning about one's perceptions, volitions, and thoughts.

V. THE UTILITY OF HIGHER-ORDER STATES

1. Introduction

While higher-order theories of consciousness come in several varieties, all assume the Transitivity Principle: a mental state is conscious in virtue of one's awareness of that state.[151] This account has the advantage of being very intuitive: we find it hard to say a person *consciously* desires a slice of pie, for instance, if that person is not at all aware of so desiring. Thus, state consciousness is explained in terms of transitive consciousness (consciousness *of*). In turn, transitive consciousness is explained in terms of mental representation: Awareness of the state is effected by a mental representation of that state, which is therefore a HOR. Theories then diverge on the representational nature of that HOR: is it a thought or more like a perception? Following our example, is the person aware of desiring a slice of pie in virtue of thinking or believing that she has that desire, or in virtue of experiencing that desire in a way that to some extent parallels how she would see or touch the slice itself? They also diverge on the metaphysical nature of higher-order representation: is the h-property instantiated by the conscious state, entailing that the state self-represents, or is the h-property instantiated by a distinct state, a HOR? Would the desire for pie and the thought or experience of that desire comprise one state? Furthermore, if the meta-representational vehicle is a distinct state, is its occurrence necessary for the first-order state's being conscious, or is the mere disposition for its occurrence enough? In addition, higher-order theorists are divided as to the *functional* nature of this meta-representation. Rosenthal (2008), for instance, has argued that HOTs contribute little to cognition, relative to first-order content. This would mean that, relative to the cognitive impact

[151] Or awareness of oneself as being in that state, as the higher-order thought theorist would have.

of the person's seeing a slice of pie, believing that it's on the table and/or desiring it, concurrent thoughts about such mental states would have little use or effect, if any. But Armstrong and Lycan have maintained that HOPs serve to "integrate" or "coordinate" first-order contents, which is necessary to sophisticated behavior.[152] Similarly, I will argue in this chapter that HORs – whether they are construed as thoughts or perceptions – have an information-processing role, although that role would be better described as intramental "communication" than "coordination."

As I have maintained in previous chapters, the efficacy of a state's c-property on the higher-order view depends upon the efficacy of HORs qua their h-properties. But it's not obvious how h-properties could be cognitively useful, given the *contents* of HORs. For example, a conscious perception of a house is about a house, a conscious volition to run is about one's running, but the HORs that target those states are about *seeing* a house and *willing* to run, respectively. These contents seem irrelevant to processing information about the environment or to producing action: One needs a state about a house in the environment to get that information, not a state about one's perceiving that object; similarly, one needs a state about running to drive one's running, not a state about one's willing to do so. But while HORs do not plausibly generate environmental information or bodily movement, they may be recruited in reasoning that ultimately benefits thinking or behavior.

In fact, that's my general proposal in this chapter. More specifically, I argue that HORs are used in *inferentially reacting* to one's current first-order representations (FORs). An inferential reaction, as I define it, is a type of cognitive reaction that can guide behavior, the other type being a "schematic" reaction. Let us first examine this dichotomy with respect to a reaction to events in one's immediate environment (or body), as opposed to events in one's

[152] See Armstrong (1997) and Lycan (1997).

mind. That reaction can be accomplished strictly via FORs: perceptions provide "what" and "where" information on extra-mental events, while volitions hooked up causally to those perceptions execute the reaction. Such reacting would be *cognitive* insofar as it is guided by mental states; we can also noncognitively react to the environment, via reflexes. Now, if the causation from perception to volition is "direct" in the sense of involving no inferential steps, it is *schematic*, meaning that it is driven by an action schema. Schemas are goal-activated stimulus-response associations that constitute the "lower-level" of action control.[153] So if I have the goal of driving to the airport on my usual route, the sight of the Dixie-Forest Drive intersection may *directly* trigger a volition to make a right turn (or to make a turn signal, check for oncoming traffic, etc.). That is, it may do so without my thinking (consciously or not) "I'm at Dixie-Forest Drive. Turning right here should get me to the airport" or engaging in similar practical reasoning. In other cases, inferential steps may be involved. For example, if that intersection were unfamiliar to me, I may launch into a deliberation – involving, say, a desire to get to the airport quickly, beliefs about MapQuest, etc. – that yields a volition to turn right. This would be an *inferential* reaction to my arriving at Dixie-Forest Drive.

It is clear that either type of cognitive reaction to an event *x* in one's environment requires representing *x* (via some sensory modality). But where *x* is a mental event, we *need not* represent *x* – i.e., deploy a HOR about *x* – in order to cognitively react to it. Since *x* is already in the mind, it can prompt cognitive reaction without first being represented. Importantly, however, that's only in the case of a *schematic* reaction to that mental event. For example, consider a person suffering from intrusive-thought disorder who has become accustomed to snapping a

[153] For more on the distinction between S-R associations and executive control, see Norman and Shallice (1986). I would also consider more complex associative processes to be schematic, provided they are "pre-established" and not driven by reasoning. I have in mind M. Posner and C. Snyder's (1975) notion of "preconscious spreading activation," where a presented word (for example) automatically activates both memory traces and representations of semantically related words.

rubber band on his wrist whenever the thoughts arise; it's plausible that the thoughts and the volition to snap the band have been schematized, so that the thoughts directly trigger the volition. On the other hand, someone who has recently learned the technique may well run through a quick, likely nonconscious inference about what to do: "There go those thoughts again. Snapping the band should distract me from them ..." Essentially, then, inferential reaction requires asserting the occurrence of *x*, the particular mental event one reacts to.[154] More specifically, the HOR would be used in predicating something about one's current mental state. For example, the ratiocination might begin with thoughts like *my perception p is occurring at time t*, *my perception q might be inaccurate*, etc. Such metacognition entails that the object of such thoughts is conscious, according to the higher-order view. Inferential reaction to *x* is thus sufficient – though not necessary – for *x*'s being conscious. What is necessary is just the occurrence of the HOR, which need not be deployed in inference.

In Chapter 4, sect. 9 I argued that the function of HORs with regard to control qualia is to enable (immediate) inferential reaction to that qualia. In this chapter, I wish to show that the general function of HORs is to allow reasoning about one's current mental states. This is a function they have in virtue of their h-properties specifically. HORs have other properties not relevant to such reasoning, for example, *being introspectible* (via third-order states) or perhaps the properties of constitution they have corresponding to the color and temperature of their neural-object constituents. The causal relevance of h-properties, as I argued in that section, is not *sufficient* for the causal relevance of state consciousness; that will depend on the nature of representation (specifically, whether a HOR's representing a state is a relation to that state) and

[154] Thus, HORs provide the "premise" for reasoning about their target state, and if that premise is false, any reasoning based on it will be unsound. Now, some higher-order theorists do allow for "empty" HORs (lacking actual target states) and those that misrepresent their targets; so there is at least a theoretical possibility (if not an empirical one) that we can be conscious of being in states we are not actually in. But the role I posit for HORs requires only that they accurately represent their targets most of the time, which no doubt they do.

on the necessary conditions for state consciousness to arise (specifically, whether HORs are sufficient for conscious experience). But the causal relevance of h-properties is at least *necessary* for the causal relevance of state consciousness. The causal contribution of target states themselves to metacognitive reasoning may also be necessary, and I have tried to show that is arguable: In the "causal field"[155] of typical psychological conditions, metacognitive reasoning about a given target state would not arise without the HOR's representing an *actual* target state. So the first-order state can thus be said to bring about the reasoning. And it would do so qua *accompanying the HOR's representing it*, which is one plausible interpretation of the target's c-property.

I will refer to my hypothesis as the Inferential Metacognition Theory (IMT) of the function of state consciousness (and refer to inferential metacognition as "IM"). For clarity, that function should be characterized as "inferential metacognition" instead of "thinking about one's mental states" because the latter could occur just in virtue of the presence of a HOR. There is inferential cognition, on the other hand, only if that representation is put to use in reasoning.[156] This would be a kind of "executive" processing insofar as reasoning is the domain of the executive system – primarily the dorsolateral prefrontal cortex. But the information that's processed is about an *internal* condition, not an external one like the presence of a stop sign.

Thus, we would seem to have theoretical backing for the empirically established link between executive processing and consciousness: HORs are constitutively linked to IM, and IM is a kind of executive processing. But clearly, not all (or even most) cases of executive processing are inferential reactions to one's current mental states. Inferential reaction to objects

[155] See Mackie (1965), sect. 2.
[156] I should note that HORs may be useful in this way even if they do not constitute state consciousness. Block, for example, considers such representations at best a byproduct of the kind of neural activation that *does* constitute consciousness.

in one's environment also counts as executive processing, yet it can proceed via FORs. Suppose I consciously see a puddle in my way, desire to get past it, determine that circumventing it is easier than stepping over it, and will to do the former. The reasoning here does not concern my seeing the puddle, nor any mental state of mine. Yet the information that there is a puddle in a certain location reached my executive system (ex hypothesi) and generated a HOR, which went unused in the subsequent reasoning. Still, executive processing of first-order perceptual content seems to be *associated* with HORs of those perceptions: When we reason about perceived objects, those perceptions tend to be conscious. In sect. 2 I give an account, based on IMT, of why this is so. And there are other empirical correlations and disassociations between consciousness and various kinds of mental processing that IMT needs to account for. For example, if perceptual consciousness is ongoing, while volitions are seldom conscious (as I argued in Chapter 4), how would IMT explain these facts? It must be shown that enabling inferential reaction to one's volitions is on the whole less crucial than enabling inferential reaction to one's perceptual states; thus HORs about perceptual states will arise more frequently. Yet volitions do seem to become conscious when one is attempting difficult and/or critical actions, and I devote sect. 3 to an IMT-based account of this correlation. Similarly, reasoning that challenges the subject is often conscious, and in sect. 4 I explain the utility of IM in that case.

Before proceeding to specific applications of the theory, clarification is in order about what type(s) of higher-order views IMT is compatible with. I intend it to be compatible with both HOT and "inner sense" or HOP versions of higher-order theory. The former, whose advocates include Rosenthal and Carruthers, posits that a mental state is conscious in virtue of one's

having[157] an assertoric thought that one is in that state, while the latter sort of theory, held by Armstrong and Lycan, says that the HOR is instead an experience of being in that state, generated by a (hypothesized) internal monitor. As discussed, reasoning about one's mental state requires deploying representations of it in propositional-attitude states. For example, one believes that one's doubt (say, that taking a driving test will be easy) has arisen at an inopportune time. Here, a representation of one's doubt is deployed in a belief state. Since such a state is generally held to represent conceptually, the embedded representation of the doubt, it seems, must also be purely conceptual, and not in any sense perceptual. So the theory I am proposing may appear committed to the HOT view.[158] But it is not clear that the kind of "quasi" perceptions that inner-sense theorists posit are *not* the kind of representations that can be deployed in belief states and reasoning. They might be perceptual in the sense of mobilizing a nonlinguistic "introspective concept" of the first-order state[159] or having an analog structure that picks out more detail in the target state than is linguistically expressible. And such a representation may be suited to be deployed in a propositional attitude.[160] For example, let 'Ψ' be an analog representation of one's doubt that the driving test will be easy. One may then think: Ψ *has arisen at an inopportune time*. Perhaps it is not plausible that there can be such representations of first-order states when these are propositional attitudes, as opposed to first-order perceptions, emotions, or other qualitative states, since propositional attitudes may not *have* more detail to represent than what is linguistically expressible. But that would be a problem for HOP theory as applied to conscious propositional attitudes. Assuming that it can be resolved, I argue that HOPs

[157] Or being disposed to have, for Carruthers.
[158] The theoretical possibility, admitted by Rosenthal (2005), that HOTs may be "empty" (lacking actual target states) or misrepresent their targets isn't a problem for the functional role I am advocating, as it is surely rare that such HOTs would occur and lead us to reason about states we are not actually in.
[159] See Lycan (2004).
[160] Fodor (1975, Ch. 4) in fact allows that analog representations can be syntactic and play a computational role in a "language of thought."

can plausibly be embedded in IM about all kinds of first-order states. Note that HOPs may also subserve IM in a different way: It is possible that a thought about one's being in a mental state *x* is normally brought about by perceiving *x*, and perceiving *x* is what makes *x* conscious. In that case, the utility of the HOT in metacognitive reasoning confers utility on the HOP it arises from.[161]

Above I noted that inner sense theorists suggest an intramental "coordination" role for higher-order representation, and that my proposal would be better characterized as intramental communication. A coordination of FORs may indeed result from reasoning about one's mental states, as I will illustrate in sect. 4. But HORs would not *themselves* do the coordinating; rather, they would enable it, by containing the information about their target states used in the reasoning. Higher-order representation constitutes intramental "communication" in the following way: A HOR, and any reasoning that involves it, is presumably located in executive areas of the cortex, while its target state – at least when that state is a perception, volition, or emotion – is presumably located in a distinct neural system (e.g., visual cortex, motor cortex, limbic system). So the HOR is essentially informing the executive about a state of a separate functional area of the brain; its occurrence is an instance of *cognitive access* to that first-order mental state, as opposed to what the first-order state represents. For example, the information *I see that it's raining*, as distinct from the information *it's raining*, becomes "poised" to impact my belief-desire complex and rational control of behavior, as writers like Block and Tye have characterized access.[162]

---

[161] Carruthers argues that if the role of higher-order experiences serve "in underpinning and providing content for higher-order thoughts" then the faculty of inner sense would be "redundant" (2000, sect. 5.2). But that wouldn't be the case if HOPs and HOTs serve different roles: Suppose only HOPs can implement our awareness of our own mental states (i.e., the arguments of the inner-sense theorists are successful), while only HOTs can be deployed in IM.

[162] See Block (1997a), p. 382; and Tye (2000), p. 62.

Given this scenario, IMT may appear incompatible with *dispositional* higher-order views. If the HOR's mere tendency to occur were sufficient for the target state's consciousness, then its information would not be available for IM. So it seems we can't hold that the role of state consciousness is to subserve IM on the dispositionalist view. However, Carruthers has argued that there is no need for a representation of a given first-order state *m* to be available prior to IM, or what he calls "reflexive thinking," about *m*. All that is needed is the availability of *m* itself to a "mind-reading" system that deploys HORs and can engage in IM.[163] Since consciousness *is* that availability, consciousness subserves IM. Note that unlike the dispositionalist, the actualist must explain why HORs are regularly present (constituting the "stream of consciousness"), while only sometimes being deployed in IM, as seems to be the case. The dispositionalist is only committed to HORs occurring just when IM does, as she does not posit actual HORs to explain consciousness. Now, the actualist might concede that HORs need occur only when IM does, but argue that IM may well *routinely* occur, thus explaining why HORs are ongoing. Insofar as I favor actualist higher-order theory, I will defend its compatibility with IMT along these lines. The intramental communication effected by HORs would then happen only (or primarily) when the information is put to use in reasoning, but that reasoning is fairly common.

Lastly, I should add that, to some extent, IMT is incompatible with *self-representational* views of state consciousness. If the HOR relays information about its target state to the executive system by occurring *in* that system, we have theoretical motivation to count the HOR as distinct from that target state, at least when the latter occurs in a separate functional area, say downstream in visual cortex. Of course, this kind of criterion for distinct HORs would not entail

[163] See Carruthers (1996), pp. 200-202. I prefer the term "inferential metacognition" to Carruthers' "reflexive thinking" as a description of the use of state consciousness. The latter may describe simply having a HOR – say, a belief that one is in a certain mental state – without deploying that representation in reasoning. And such reflexive thoughts are not, in and of themselves, useful to have. They are what consciousness *is* (at least on actualist higher-order theory), not what it *does* or *facilitates*.

the distinctness of an HOR and its target that occur in the *same* neural system. For example, in the case of a conscious desire, both the desire and its accompanying HOR may occur in the prefrontal cortex.[164]

2. Why Conscious Perception?

On the whole, the most salient kind of consciousness is perceptual: our states of hearing, smelling, touching, tasting, and seeing are often the qualitatively strongest types of conscious states, as compared to thoughts, mental images, emotions, etc.[165] Of course, one can become absorbed in thoughts and feelings to such an extent that perceptions are pushed to the fringe of awareness, but perception retains the capacity to "intrude" and take the focus even in these moments. It's natural to give an evolutionary explanation here: ongoing information about the environment is critical to a creature's fitness. And since perceptions are often conscious (indeed, from the first-person perspective, they are *always* conscious), it's also natural to think that their being conscious is necessary to processing that information, or at least facilitates the processing. Given the evidence for *nonconscious* perception, however, the latter claim would be the more plausible one for the anti-epiphenomenalist to make. (In the present context, anti-epiphenomenalism is the claim that a perception's c-property is efficacious in perception.) Some of these well-known bits of evidence are anecdotal (e.g., the long-distance truck driver) while

[164] Thus, Rosenthal's argument for higher-order states' being extrinsic to their targets is perhaps stronger. Essentially, we individuate states on the basis of the attitude (if any) internal to them. I may both wonder whether *P* and hope that *P*, but these can't be the same state, as the attitudes are different. Now, a HOT must have an assertoric attitude, e.g.: I see that it's raining. In contrast, my visual state representing the rain, which is conscious in virtue of that HOT, has no internal attitude. So it must be a state distinct from that HOT. Regarding the conscious desire, the first-order attitude differs from that of the HOT that makes it conscious, so again they count as distinct states, contrary to the self-representational view.

[165] It has even been argued that *all* phenomenal consciousness is perceptual in some sense. See Prinz (2007).

others are based in experiments involving normal subjects presented with masked primes or special kinds of patient with impaired visual systems (e.g., blindsighters, prosopagnosiacs). Results show that participants can perform tasks where certain perceptual information would be required, while reporting no awareness of that information. Such cases of nonconscious control are unsurprising on higher-order theory: first-order perceptions provide information on one's immediate environment that, together with one's first-order belief/desire complex, sufficient to guide action. So, for example, blindsighters' supposed inability to form suitable HORs about perceptions in their blind field should not inhibit their ability to react – inferentially or schematically – to objects and events in that area. All they would need is some representation (which presumably counts as a perception) of those things.

It's likely, however, that blindsighters' above-chance ability to reach for objects in their blind field is *schematically* mediated, and does not involve reasoning. As Lawrence Weiskrantz has observed, "manipulating items in thought and imagery" seems possible only for *conscious* perceptual content (2007, p. 171). That is, only when an object is consciously perceived can we use representations of the object in reasoning about how to react to it. Thus, "the patient with unilateral neglect may covertly process visual information in his neglected field, but he himself ignores food on the left side of his plate!" (p. 170). Most patients, he also notes, will not duck an object flying at them in their blind field. The simple, practical reasoning that guides such reactions to objects and events in the blind or neglected field is apparently unable to engage.[166] Experiments also support the correlation between consciously perceived stimuli and "adaptive"

[166] Ex hypothesi, the minority of patients who *do* duck would have a schema governing the reaction. Now, I do not deny that actions guided by nonconscious perceptions and schemas can be called "rational," if they satisfy rational goals. Ducking might be an example. But that doesn't entail that such actions guided by reasoning.

as opposed to "automatic" response to those stimuli – the difference lying, I argue, in the use of reasoning in the former case. Consider the following two kinds of experiment[167]:

(i) *Exclusion tasks*. Subjects are presented subjects with single words on individual trials, either very briefly (50 ms) so that the word is seen nonconsciously, or for a longer duration (150 ms) allowing conscious perception. Immediately after the presentation of each word, subjects are shown the first three letters of the word and asked to complete the stem with any word that came to mind *except* the word that had just been presented. It is found that they have difficulty "excluding" subliminally presented words (e.g., they tend to complete "dou-" as "dough" if "dough" had been presented) but little difficulty excluding those words they consciously see.

(ii) *Prediction based on stimulus redundancy*. Experimenters flash either the word "red" or "green" and then show either a red or green patch whose color participants are to name as quickly as possible. Due to the Stroop effect,[168] subjects take longer to name the color of the patch when the forgoing word is incongruent, whether or not the word is flashed subliminally. However, when the incongruent word/color patch pairings are presented much more often than the congruent ones, participants capitalize on the pattern and their response time become faster for the mismatches – but *only* when the words are flashed to allow conscious perception. Otherwise, the Stroop effect continues.

[167] For a more detailed review, see P.M. Merikle and M. Daneman (1998).
[168] In 1935, American psychologist J.R. Stroop published a now famous study showing that naming the color of a word printed in color *C* takes longer when that word denotes a color different than *C*, as the automatic semantic processing interferes with the naming task. Similarly, in experiment (ii), semantically processing the flashed word creates the disposition to name the color it denotes, which participants must override when the patch is of a different color.

In both (i) and (ii), there is a correlation between the stimulus perception's being conscious and the subject's ability to use the perceptual information in a rational way: In (i), they reason (explicitly or not) about the consciously flashed word in a manner such as: "I'm supposed to exclude 'dough,' so I'll complete 'dou-' some other way." In (ii), they reason inductively about the pattern of word/patch presentation based (in part) on the conscious information about the flashed word. When the stimuli in the experiments are nonconsciously perceived, however, subjects react schematically: Mere stimulus-response associations dispose them, in (i), to complete a word stem with a just-presented word and, in (ii), to name a color patch according to a previously denoted color (hence the Stroop interference for incongruent pairs).

However, the mere correlation between inferential reaction to a stimulus and the conscious perception of that stimulus does not entail that the c-property of the perception enables the reaction. The nonconscious perceptions of the long-distance truck driver, the blindsight patient, the subject presented with a masked prime, etc., all presumably lack not only accompanying HORs, but *also* access to the executive system: The first-order representations may reach the lower level – the units that encode schemas by producing certain volitions in response to those representations – but not the executive, which is where inferential reaction takes place. In short, the perceptual states are not "access-conscious," in Block's sense. And, as Block argues, it is wrong to assume that enabling rational control is a function of phenomenal consciousness[169] instead of access consciousness, when both are missing. Now suppose phenomenal consciousness is reduced to higher-order representation, what Block calls "monitoring" or "reflexive" consciousness. While Block resists this move, his reasoning still

[169] Block defines this term along Nagel's lines: a state is phenomenally conscious iff there is "something it is like" to be in that state, which (for Block) is metaphysically independent of any functional relations the state has.

applies: we can't just assume that enabling rational control is a function of monitoring consciousness instead of access consciousness, when both are missing. It may be that access consciousness is what enables rational control, while monitoring consciousness is *epiphenomenal* to access consciousness. And that would explain why, when a certain perception *p* lacks access, there is no HOR targeting it. On this scenario, *p*'s c-property would enable neither *p*'s occurrence, nor inferential reaction to *p*'s content, since *p* acquires that property *after* access has occurred.[170] And in fact, we have little theoretical reason to think that a representation *about one's perceiving* would be useful to either perception or access. Perception only requires information about the *object of perception*, and that information's reaching the executive requires only suitable causal connections. Why should those connections be mediated by a representation of the perceptual state? Prima facie then, the correlation between the two is best explained through epiphenomenalism, with the representation of *p* being a byproduct of *p*'s content becoming available to "flexible thinking." As Rosenthal notes: "When flexible thinking does enlist perceptual contents, it very likely brings along HOTs about those contents, resulting in the perceptions' being conscious. So if their being conscious is blocked, flexible thinking about them is likely to be as well. Benefit in the normal case, however, may then be due just to the flexible thinking, and not to the perceptions' being conscious" (2008, p. 835).[171]

---

[170] Velmans (1991) supports a similar claim with a variety of evidence for the "lateness" of consciousness with respect to focal attention. Access to the executive system, focal attention, and "global dissemination" all describe relatively late information-processing stages associated with state consciousness. Some first-order theorists *identify* a conscious state's c-property (at least in part) with such functional properties: e.g., being focally attended. Velmans does not, since he thinks the evidence shows consciousness arises after content has been focally attended. For example, subjects in W.D. Marslen-Wilson and L.K. Tyler (1980) apparently applied context-based analysis to complete word fragments within 200 ms of seeing the fragment, which is likely *before* consciousness of a stimulus can set in.

[171] I should note that Rosenthal's position is that consciousness "has no significant function," not that it is epiphenomenal. HOTs surely have some cognitive impact, but one that is "too small, varied, or neutral in respect of benefit to the organism," he writes (p. 831). IMT, in contrast, ascribes a stable and significant use for higher-order content, as I hope will become evident.

The epiphenomenalist scenario, however, begs an important question: Why should representations of perceptions "very likely" arise from access? Rosenthal has given a causal explanation of why such representations tend to arise in general,[172] but another reasonable hypothesis is teleological: they are subsequently useful to flexible thinking involving perceptual contents. This would mean that once a representation of a perception *p* results from *p*'s access to the executive, the HOR can then be deployed within the executive processing of *p*'s content.[173] And that utility would explain why the HOR tends to co-occur with access to *p*'s content, i.e., why *p* tends to be conscious when it impacts the executive. What is that utility? First, note that on IMT, a representation of *p* is useful because it enables inferential reaction to *p*. It follows that inferential reaction to *p* must be relevant (at least in some cases) to inferential reaction to the external object that *p* represents (i.e., executive processing of that content, flexibly thinking about it, etc.). To see why this is so, consider the following type of cognition by a person *S*, where *p* is *S*'s perception of an object *x* over a duration *d*:

(i) *S* wants to alter *p* for some reason; (ii) *S* believes that *x* is causing *p*; (iii) *S* wills to alter *x* so as to alter *p* as desired.

---

[172] In brief, HOTs about one's present perceptual states have their origin in error detection, though their function is not error detection. Detecting that one has perceived erroneously requires thinking about that perception – though as a *past* state. But since such detection occurs fairly often, he argues, one will get habituated to having such HOTs. Eventually, perceptions will dispose one to think about them when they occur, rendering them conscious. See Rosenthal (2005), pp. 303-305.

[173] Velmans (1991) also considers whether consciousness may impact information-processing *after* content has been focally attended. But "it is not clear what additional activation *mediated by consciousness* would achieve," he writes. "Only information selected for focal-attentive processing enters consciousness. According to 'activation' theories of attention, this will only occur if it is *already* more activated than competing, less relevant information – in which case no additional, 'consciousness-mediated activation' would be required to ensure its prominence in subsequent activity." It is indeed difficult to see what efficacy consciousness might have on the nonreductive view of the property that Velmans takes (see his sect. 9.2). In fact, even if consciousness regularly *preceded* focal-attention, *how* it facilitated that stage would remain mysterious. But on the higher-order view, the property is reduced to a special type of metacognitive information. As such, it can enable processing that deploys that information, after the onset of focal attention.

Assuming that (i)-(iii) occur over *d*, *S* would be thinking about *p* as a current state, entailing that *p* is conscious. So (i)-(iii) is a case of IM. But *S* is also reasoning about *x*; in fact, (i)-(iii) guides *S*'s reaction to *x* – seeking to alter it. So in this type of case, inferential reaction to *x includes* inferential reaction to *p*. This means both the FOR (*p*) and its accompanying HOR are relevant to the cognition.

Why might one engage in the sort of IM described in (i)? For various reasons, one may wish to alter, upend, or preserve one's perceptual state. One may be hearing a very loud sound one realizes is unsafe to hear, looking at the front of a computer tower when the sought-after serial number is on the back, getting the ideal view of a portrait subject, feeling the correct part of the car engine while under the chassis to make a repair, etc. Accordingly, one will seek to adjust one's perceptual state in the first two cases and maintain it in the latter two. Thoughts about one's current perceptual state, then, can engender desires for certain perceptual acts: e.g., to get a clearer view of *x*, to shift focus from *x*. In turn, one will seek to alter or preserve the object(s) causing the current perception. Note that the action target is oftentimes the object's *egocentric location*. Perceptions, particularly visual ones, don't represent an object in the environment in isolation; rather, the object is represented in a certain spatial relation to one's body (as well as in allocentric relations with other objects). So one doesn't just see Mt. Rushmore, but also how one is situated with respect to it. Since the egocentric location is also a cause of the perceptual state (it results in a certain view of the object, affects the audition of any sound the object is making, etc.), one may seek to alter or preserve that location in order to fulfill one's desires with respect to the perceptual state. For instance, I may realize I'm not seeing the part of Mt. Rushmore I want to see; clearly I won't want to alter anything intrinsic to Mt.

Rushmore in order to achieve the desired perceptual state, but rather I'll seek to change its egocentric location simply by standing somewhere different.

It may be objected that reasoning about perceptual *objects* would suffice in putative cases where IM is involved in rational control of behavior. So once I consciously hear a screeching sound, I simply believe it's occurring in my immediate environment, think it may cause auditory damage, and decide I should cover my ears. This reasoning deploys no representation of my auditory state. But of course if I wasn't *hearing* the sound, it's not likely putting my auditory system at risk. Thus, my belief that I'm hearing it, not just the belief that it's occurring in my vicinity, surely figures in my reasoning, at least implicitly.

As noted, reasoning that leads to a desire to alter, upend, or preserve one's current perceptual state must deploy a (higher-order) representation of that state, entailing that the state is conscious. So if such a desire factors into one's inferential reaction to the object of perception, the HOR has been efficacious within the executive processing of that first-order content. This potential utility explains the correlation between a perceptual content's reaching the executive and the occurrence of a HOR: When the brain makes the content of an act of perceiving available to the executive, it also tends to make available information about that act of perceiving, as reasoning about the perceiving may impact reaction to its content. Put differently, my claim is that when a person is *enabled* to react inferentially to object *x* being in egocentric location *l* (via the perceptual content *x*-in-*l* accessing the executive), she is also *enabled* to react inferentially to her perceiving *x*-in-*l* (via a suitable HOR of her perceiving *x*-in-*l*), since reasoning about the perception of *x* may be relevant to reasoning about *x*. For the dispositionalist, the person is enabled to react inferentially to her perception of *x* via that perception's accessing the HOR system, not via an actual representation of the perception occurring. But since access to that

system constitutes the perception's c-property, the link between conscious perception and executive access to first-order perceptual content holds for the reason given, except that the link is between executive access to the perceptual content *x*-in-*l* and HOR-system access to perceiving *x*-in-*l*.

Here is where the compatibility of actualist theory with IMT might be questioned: Clearly HORs are deployed in IM, but need they be available beforehand for IM to occur? If that were required, it would explain the continual HORs posited by the actualist. But it's doubtful that prior availability is needed. Given a capacity for IM about perceptions,[174] HORs can arise *just when* one engages in IM. As Carruthers has argued (regarding HOTs), "[T]here is no reason why my own experiences and thoughts should actually give rise, routinely, to HOTs concerning themselves. It would be sufficient that they should be available to HOT [the "mind-reading" system], so that I can entertain thoughts about the relevant aspects of my experiences or thoughts when required" (2000, sect. 6.1). This plausible scenario seems to saddle the actualist with a great many unused HORs producing the stream of perceptual consciousness – unless IM occurs fairly continually as well. I'll argue that a relatively simplistic IM about one's perceptions routinely takes place.

First let us consider when and how IM about a perception accessed by the mind-reading system would begin. As explained in sect. 1, a representation of a mental state's occurrence – as opposed to its own content – serves as a rational basis for IM about that state. So a HOR of a perception *p* would combine with standing metacognitive beliefs and desires to prompt reasoning about *p*. An example: Suppose I want to avoid seeing the potentially high charges on a bill until I clear up some other distressing matter. I therefore want to avoid looking at the document at all

[174] Based in the mind-reading system's rational abilities and theory-of-mind concepts, i.e., those concepts necessary to attribute mental states to oneself and to others.

from any close distance, since I might then notice the charges. How would this metacognitive desire become involved in reasoning about a current state of seeing the bill (call that state *s*) that I find myself in – say because a family member had absentmindedly left the bill on the kitchen table? Surely, the desire must rationally interact with a HOR about *s*, say a thought that I am now seeing the bill. For example: *I desire to avoid seeing this bill. I am now seeing it. So, I should upend my current state. My looking in direction x is causing that perception. So, I'll turn away from x.* Thus, the metacognitive reasoning is (partly) grounded in the second state, the HOR with the content *I am now in s*. It is not rationally grounded in *s* itself, which carries the information that the bill is on the table in front of me. That information would not work with my standing desire to avoid seeing the bill to get the reasoning started (the bill could be on the table in front of me without my seeing it); so I clearly need the information that I am seeing the bill.

What engenders a HOR on the dispositionalist theory, then, is both a context of beliefs and desires relevant to IM, such as desires to avoid certain types of perceptions, beliefs that certain types are useful, etc., along with the availability of the target state to the mind-reading system. So if I desire to avoid seeing the bill (say, to avoid visual states of type *B*) and if *s* is available to the system, I will tend to realize that *s* is of type *B*, that I must upend *s*, and so forth. This scenario indeed limits the occurrence of HORs to cases when beliefs and desires relevant to IM are present; one then "checks" whether current perceptions are consistent with or fulfill those beliefs and desires, and the perceptions are represented in the process. But here the actualist can give a plausible account of why HORs are ongoing, namely, because there is general metacognitive content that current perceptions are routinely checked against. This would be a desire to avoid unpleasant and/or disadvantageous perceptions, or alternatively, a belief that such perceptions should be avoided. ("Disadvantageous" would of course be relative to one's

perceptual goals, which are usually to be receiving information on the environmental features that matter to one.) Fast, constant assessments of whether one's current perceptual state falls into these categories would be a basic level of IM, one that may or may not lead to further metacognitive reasoning, depending on whether the system flags a current perception as unpleasant and/or disadvantageous, in which case further reasoning would concern how to alter or upend the states, as described in (i)-(iii) above. Or perhaps, the state is assessed as especially pleasant and/or advantageous, in which case further IM would concern how to preserve the state.[175] Now, even if HORs are regularly generated for this purpose, it is still theoretically possible that a conscious perception's c-property is constituted by its availability to the mind-reading faculty, not its being targeted by an actual HOR. My present objective, however, is not to argue against the dispositionalist theory of consciousness, but only to make a case that the routine occurrence of HORs posited by the actualist is not a detriment to his account if we assume IMT.

It's also important to note that not all reasoning about one's perceptual states deploys the kind of HOR that makes its target state conscious. The state must be represented as occurrent, according to actualist higher-order views. So if ratiocination leads me to think I should ensure I hear my alarm clock, my reasoning deploys a representation of a state of audition. But since that state is represented as desired and not as present, I won't have a conscious experience of hearing my alarm.[176] Thus, reasoning about a perception that one should have (or will have, or has had) is inferential metacognition, but not the kind that entails state consciousness.

---

[175] Admittedly, ongoing IM would require more cognitive resources than ongoing simple ("unembedded") HORs, so the proposal seems even more vulnerable to the "objection from cognitive overload" than higher-order theory is in general. The objection is that the HORs required to create all of our conscious experiences would strain cortical resources. But whether this is true is an empirical matter. See Rosenthal (2004), p. 25.

[176] If I did deploy a HOR of the suitable kind, I *would* have such an experience, argue theorists who hold that a suitable HOR is sufficient for state consciousness (e.g., Rosenthal). Others hold that the HOR must target an actual first-order state.

I would also point out that voluntarily shifting perceptual focus does not require representing one's current perceptual state in that modality. Prima facie, it might seem so: Suppose I want to see the dog in the middle of the street, and I'm now looking at the left side of the street. I have to turn my head (or eyes) to the right, but I won't know that unless I represent that I'm now seeing the left side. Yet to guide my head turning, it is sufficient that I represent the *allocentric location* of the part of the street that's currently in view (with respect to the dog), which is not to represent that *I am seeing* that part. So a representation of my present visual state isn't needed. Indeed, blindsight patients have been shown to be better than chance at shifting their gaze to a flash of light in an area of the visual field wherein they had reported not being able to see flashes of light.[177] In the course of this perceptual act, the patient would need to shift his gaze from some part *A* of his blind field to the location of the flash in the field. If that required a representation of *his seeing A*, instead of just *A*'s being currently in front of him, then he would be conscious of *A*. But the evidence weighs against that possibility.

3. Why Conscious Volition?

A volition, I said in Chapter 4, is the most immediate mental cause of a voluntary action. More specifically, it is an attitude of willing directed to a proposition expressing the immediate occurrence of a certain act of oneself: e.g., *I lift the cup now*. Such a state, as I've argued, is not often conscious. And when conscious, it is typically phenomenologically thin, meaning that it is a rather brief and mild feeling (sometimes described as an "urge") of being about to move in the way expressed by *P*. Assuming the higher-order view, it follows that volitions are seldom

[177] See E. Poppel et al. (1973).

accompanied by HORs. And when they are, the HOR (presumably) represents the volition's content in a fairly minimal way, explaining the phenomenal thinness. Following IMT, the explanation for the relative latency of volitions among conscious mental phenomena is that it is seldom useful to react inferentially to one's volitions, which is just what HORs enable. As I will argue, volitions tend to become conscious primarily when we are performing acts that we perceive to be difficult and/or critical. By "critical" I mean that the act or its potential consequences are either very desirable or undesirable to the agent. In those (relatively infrequent) cases, it *is* useful to react inferentially to one's volitions, and that is why representations of them arise.

In fact, I will argue that inferential metacognitions about conscious volition and conscious perception serve a similar purpose. A HOR of a perception can be deployed in reasoning about whether and how to alter, upend, or preserve that perception; that is, the objective is the control of that perception. It is natural that this ability should be regularly facilitated, since one's perceptual states can change unpredictably. Environmental conditions may result in unwanted perceptions: one may desire to turn a corner but *not* desire to see the road kill that lies there. In contrast, volitions normally occur in a controlled fashion. That is not to say they satisfy desires, for one need not (and rarely would) desire to have a volition. Rather, volitions are controlled in that they are caused by desires (along with beliefs), and *support* desire satisfaction. So if I want peanut butter (and believe that now is a good time for a snack), a volition to open the jar will tend to arise to help satisfy that desire.[178] The question, then, is that since volitions are generally well-controlled psychological states that don't contravene one's desires, why would they need altering or upending? (Preserving them isn't an issue since their

[178] The beliefs and desires that guide volitions are not their *exclusive* cause, of course; perceptions also factor in. The volition to open the jar won't arise until I believe it is within reach, which belief is the result of *seeing* it is within reach.

causal function is served as brief, pre-movement states, while it is sometimes beneficial that a perception be ongoing.) It would seem that control of behavior, not volition, is the issue, since behavior *can* upset desires: The jar might be stuck, I have difficulty opening it, and decide to abort the attempt or try opening it a different way.

But while volitions unproblematically support desires,[179] those desires might change given new considerations just before action, at the moment of volition. In that case, the ability to alter or stop the volitional process would be useful. Libet has proposed that consciousness has this function with regard to volition, so that a volition is conscious (about one second before movement) just when the brain can upend it with a conscious "veto." "The existence of a potentiality to veto is not in doubt," Libet writes. "Everyone has experienced having a wish or urge to perform an act, but vetoed the actual performance of the act" (2003, p. 25). Libet's further claim that the veto preserves free will (or at least "free won't") is problematic, since there are surely nonconscious neural antecedents to the conscious veto itself. So if the conscious volition isn't "free" because it is predetermined by a readiness potential, then the conscious veto isn't free for the same reason. Free-will issues aside, however, the veto capability appears useful, and there is evidence that the phenomenon has its own neural substrate. Haggard, who characterizes the veto as resulting from a "late whether decision"[180] (whether to go ahead with the action), and Marcel Brass conducted a study (2007) that implicates an area of the anterior frontomedian cortex, rostral to the pre-SMA, in the veto process. Activity in the pre-SMA itself is thought to subvene conscious volitions; its direct stimulation, for instance, produces a feeling

[179] There are cases of compulsive behavior where volitions tend to support desire *A*, which is opposed to another desire, *B*, and the agent would want to be guided by *B*. That is, the person has the second-order desire that *B* is his will, i.e., that it be effected through his volitions. Even in this case, volitions still support *a* desire, and is to that extent "controlled" by the agent. (On second-order desires, see Frankfurt [1997].)

[180] Specifically, the veto is a cancellation of the volition that results from a late "no" decision about the action. Note that this cognition would likely *precede* the first comparator operation, where a forward model is constructed and compared to the volitional content.

of being about to move.[181] A third element here is the c-property of the volition: the HOR targeting the volition need not be located in the pre-SMA, and in fact it is more plausibly located in higher cortical areas outside the motor system.

Now, on my view, such a representation would not arise very often: Volitions primarily tend to be conscious in those situations when we think we are about to perform a difficult and/or critical act. Before discussing this claim, let me point out that it is rather different from Haggard's phenomenal picture. "Conscious intention," for him, admits of qualitative degrees, but he seems to think it regularly occurs to some degree: "One can be barely conscious that one is going to take the next step when walking, but intensely aware of pulling a trigger" (2008, p. 942). As I've argued, the great majority of mundane acts – such as walking – aren't preceded by *conscious* volitions at all. Nonetheless, Haggard illustrates theoretically significant situations where volitions *do* tend to be conscious. One is pulling a trigger, an act that is often "critical" in the sense defined above: The potential consequences of the act, killing a person or animal, may be very undesirable (emotionally), and yet very desirable (practically) if one is being mortally threatened. Prior to such critical acts, we often seem to become distinctly aware of being about to perform them. The acts themselves need not be difficult at all. Pulling a trigger is quite easy, in contrast to, say, playing an arpeggio on a piano when one is a beginner. Yet volitions also seem to become conscious prior to difficult acts, whether or not they are critical. The beginner may be practicing the piece at home as opposed to in recital, for instance, where misplaying the arpeggio is undesirable, but not greatly so. Still, she may well be conscious of being about to attempt that tricky passage.

Conscious volition is especially salient when a veto occurs, as the experience of willing doesn't "flow through" to the experience of acting: the process appears cut short at the willing

[181] See I. Fried et al. (1991).

stage. Haggard gives a good example: "Most people recognize the situation of being about to say angry words or send an angry email and refraining at the last moment (usually wisely!)" (2008, p. 939). Similarly, the neophyte pianist may catch herself about to play the arpeggio incorrectly. It is plausible that volitions are always conscious prior to being vetoed. But this would not establish that a volition's being conscious in any way facilitates the cancellation. In fact, the volition to do *a* might become conscious *as an epiphenomenon of* the veto process qua *a* initiating.[182] Against this scenario, Libet suggests that one becomes aware of the "decision to veto" *after* conscious volition (2003, p. 25). But, first, it's doubtful that this decision is part of the fleeting phenomenology during the second (or so) in which a veto occurs: In deciding not to pull the trigger at the last moment, I will likely just become aware of "desisting," which could just be an agentive experience of *not* acting after my volition. And second, even if there is a conscious decision to veto after the conscious volition, it remains questionable whether the volition enables the veto decision *qua being conscious*.[183] It might even be that the decision process began nonconsciously before the volition itself became conscious, and the volition's becoming conscious is epiphenomenal to the decision process. At least from the first-person perspective, we can't rule out these possibilities.

Note that if a conscious volition's c-property *were* instrumental to the veto, we would have a good explanation of why volitions to perform difficult and/or critical acts tend to be conscious; namely, these are cases when the ability to veto is most useful. For critical acts, it is natural that the mind should enable desisting at the last moment, in case new considerations

[182] In saying the volition *becomes* conscious, I am assuming a theoretical model on which volitions first occur nonconsciously, and correspond to preconscious neural initiators of action, such as the lateralized readiness potential (LRP). See Rosenthal (2002), pp. 217-18.

[183] Libet seems to think the efficacy of consciousness here lies in the efficacy of the conscious decision to veto, not so much the conscious volition. But again, the c-property of that decision may be causally inert. As Rosenthal suggests, "A neural event may operate to veto before becoming conscious" (2008, p. 834).

arise. So the angry email may be a critical act in that the person thinks it has potentially very undesirable consequences (while still really wanting to send it). A last-moment realization might arise that motivates withholding that click (e.g., realizing that the undesirable consequences outweigh the desirable ones). Similarly, the pool player about to make an easy but critical shot, say one that effectively wins the tournament, would be advantaged to be able to veto the attempt, just in case a late consideration arises as to how to better strike the ball.[184] Perhaps the act itself is highly undesirable, and a person may want to be open to last-moment reasons (or "excuses") not to do it (someone about to step into an ice-cold shower might be a case in point). With regard to acts that are difficult for the agent (whether or not they are critical), a late reconsideration of the ideas that guide the act may occur, since these ideas may be wrong given the difficulty. Perhaps the pianist about to play an arpeggio realizes that she is about to position her finger incorrectly, or play it too soon in the context of the piece. It would thus be advantageous for the person to be able to veto.

The argument under consideration, then, is that volitions tend to be conscious prior to acts that the agent perceives as difficult and/or critical because (i) the veto ability is most useful in those situations and (ii) the c-property facilitates that ability. But this remains a "just-so story," no more plausible than an epiphenomenalist account, as long as (ii) is unjustified. How is consciousness relevant to the veto ability? I think IMT provides a good answer. Essentially, the veto of an act *a* is guided in part by a reasoning process (a fast, likely nonconscious one) that deploys an HOR targeting the volition to do *a*. While the above examples make it plausible that a veto is based in practical reasoning just prior to action, we may nevertheless question whether a

[184] Note that experts such as professional pool players may not have conscious volitions even with critical performances, due to being "immersed" in the activity. (This may be advantageous, if conscious volition interferes with the fluidity of the performance.) That's why I say volition only *tends* to become conscious prior to critical acts.

representation of one's *willing* the act, not just a representation of the act, need be recruited in that reasoning.

To see why metacognition is plausibly involved, let us first examine the kind of information that enters into the "late whether decision." Suppose a volition to do *a* becomes conscious at time *t*. At *t*, or perhaps while the volition is still nonconscious, the agent reconsiders doing *a*.[185] As a result, he might will to go ahead with *a* or desist from *a*. Thus, the reconsideration arises while the volition to do *a* is in progress – that's what makes it a *late* whether decision with respect to *a*, and not just a whether decision. This happens, presumably, because something prompts reconsideration just prior to action, as may be the case with difficult and/or critical acts. Or a reason to desist from *a* may arise at the last moment even if *a* is neither difficult nor critical. For instance, most of us have had the experience of being about to reach for something (say, a dropped pen) to give it to a person, only to desist at the last moment when we realize she's about to get it for herself. But apart from reasons to reconsider *a*, I argue that the late whether decision must also be informed by the *belief that a is imminent*, which is (plausibly) caused by the volition. So we have a line of causation from a representation of movement *a* in the preSMA (where it would have the functional property that defines volition: tending to cause *a*) to the activity in the anterior frontomedian cortex that underlies the late whether decision (per the Haggard-Brass results). Now it may be that certain proprioceptions associated with preparing to do *a* cause the belief that *a* is imminent. But, unlike volitions, it is unclear whether such sensations always happen prior to voluntary acts. I may be aware I'm about to reach for a pen without feeling anything different in my arm or hand muscles. And would these sensations by

[185] I am simplifying matters by assuming the final check (as Haggard calls it) would focus on the volitional goal (*a*) as opposed to the desire that the volition subserves. My volition, for example, would be to pull the trigger now, but that subserves my desire to kill the deer. What would most naturally be reconsidered at *t* (or before) is killing the deer.

themselves prompt me to think I'm about to reach for the pen? They might be associated with a different upcoming motion.

Essentially then, the belief that *a* is imminent is needed as a "premise" to the reconsideration so that the reconsidering yields the right results: Suppose one decides (after reconsidering) that one in fact should do *a* now. A volition to do *a* would naturally arise. But then one would have *two* volitions to do *a*, which is implausible. Thus, no volition should result from a "late re-approval" of *a*, since one is already in progress. But if no volition arises from reasoning that would normally lead to a volition to do *a*, it must be because one believes that *a* is imminent.[186] Alternatively, suppose one decides that one *shouldn't* do *a* now. Of course, no new volition to do *a* would arise. But one would then still do *a*, since a volition is in progress. Thus, the result of a reconsideration that ends in a disapproval of *a* must be a decision to *withhold a*. The effect is a veto, essentially the cancellation or inhibition of the volition to do *a*. And what best explains why that veto process occurs – apart from the disapproval of *a* – is one's belief that *a* is imminent, what I will refer to as the "signaling belief."

Still, it doesn't appear that a *representation of the volition*, and hence consciousness, is required for these cognitions. The reconsideration is about whether to do *a* now; it's not about the volition. Neither is the signaling belief metarepresentational: it's also about *a*. Furthermore, there is no reason to think that a representation of the volition is required for the volition to cause the signaling belief. Essentially, the content of a volition to do *a* is just *a*, and the state has the functional property of tending to cause *a*.[187] So it's natural that such a state could directly cause the belief that *a* is imminent. A representation of my willing *a* is thus not needed for me to believe that *a* is imminent. Similarly, there is no reason to think that a representation of a

[186] It may also be because the neural substrate for volition is impaired, but I am assuming normal conditions.
[187] See Ch. 4, sect. 3.

perception *p* is needed for the rational centers to access, or acquired beliefs about, *p*'s content, namely some object in the environment.

What I argue here is that the signaling belief, along with the late disapproval of *a*, are *insufficient* to yield the decision to veto and the cancellation of the volition. One must also think that *a* is about to happen *due to one's willing it* (i.e., due to the presence of a representation whose content is *a*). That belief explains why one cancels the volition in order to prevent the movement. The mere belief that *a* is about to happen, where *a* is conceptualized just as a bodily movement, might be held if *a* is an *involuntary* movement, brought about entirely by nonvolitional factors. One can be about to kick up one's leg because of a spasm, or because the physician is about to strike one's patellar tendon with a reflex hammer. One can trip and be about to bump into a vase. Suppose one makes a late decision that such a movement should not happen. Clearly, one will not cancel any volition to do it one has; rather, one may stiffen one's muscles, or otherwise attempt to stop the movement. So a belief about the potential cause of the imminent movement informs the preventative measure. As Armstrong has suggested, "Knowledge of the presence within us of potential causes of behavior is obviously valuable … If I know that I am set towards hitting you before I hit you, I may be able to control my impulse, in a way that I could not do if I knew nothing about the impulse until it manifested itself" (1993, p. 99). When that potential cause is a present volition, the knowledge will be metarepresentational: *a* is imminent due to my present volition.

That entails *conscious* volition, on the higher-order view. The volition's c-property, then, facilitates the veto by enabling the representation of the imminent movement's mental cause. And since vetoes are often useful for difficult and/or critical acts, we have an explanation of why

volitions tend to be accompanied by HORs prior to such acts.[188] We also have an explanation of why HORs accompany volitions prior to "mundane" acts (i.e., easy and not deeply consequential) when one desists from those acts at the last moment. Note that one property of consciousness-conferring HORs – that they represent their targets as occurrent – fits this functional role: The decision to veto must be immediate if there is a late disapproval of the movement (or of goals the movement would further); it must occur while the volition still obtains. The decision must therefore be informed by the belief that the volition *is* happening, which is just the content of the HOR.[189]

It might be insisted that the veto need not be guided by such a belief. Perhaps certain first-order mental states can combine with the disapproval of an act *a* to yield a veto. For example, consider the belief that *a* is a voluntary movement – that it is about to happen due to my *will*, conceptualized as a general mental faculty and not as a state I am in. That state seems able to rationally direct the veto, assuming one has changed one's mind about doing *a*. But such a belief lacks content specific enough to guide an *immediate* veto, since it might be held if one thinks one is about to generate a volition to do *a*, but at present still has not. One must therefore think of *a* as a movement one is now willing.

Alternatively, perhaps the veto occurs simply when there is both a disapproval of *a* and a volition to do *a*. Thus the volition causally participates in its own cancellation. I think this scenario ultimately entails representation of the volition. Note that the cause of the cancellation is upstream of the preSMA: The prefrontal area where the late "no" decision occurs presumably sends a cancellation signal to the preSMA, where the volition is in progress. How would this

[188] Even if they are not used in the veto process, as the "final check" might result in a go-ahead.
[189] This is an example of a HOR serving an intramental communication role, as described in sect. 1. In this case, the HOR informs the prefrontal area where the late reconsideration takes place of a current state of the preSMA, namely the volition.

signal be sent *just when* the preSMA is in a certain volitional state, if the prefrontal area did not receive information as to the occurrence of that state? So that's how the volition would help to bring about its own cancellation – by sending information about its occurrence that works causally with the disapproval of the act to generate a veto. But that information would arguably qualify as a HOR about the volition.

There is a more plausible way of dispensing with rational guidance from a belief that the volition is occurring. The volition brings about the signaling belief (which has no metarepresentational content), and that belief works with the disapproval to generate a veto under the following condition: The agent does *not* think *a is imminent due to some nonvolitional cause* (e.g., he doesn't think he suffers from spasms, or that somebody will force him to do *a*, etc.). For if the agent *did* have such a belief, no cancellation signal should be sent to preSMA. Note that in this scenario, the veto is to a degree guided schematically. I've defined schemas as direct causal links between perceptions and volitions (inputs/outputs), that is, as devoid of intervening reasoning. Here, I'm considering other sequences of mental states as schematic when an intervening rational step is supplanted with mere causation. So, for instance, the disapproval of *a* along with the belief that *a* is imminent bring about the cancellation of the volition *without* the agent thinking that *a* is imminent due to the volition. The idea is that this schema would be "blocked" if the agent thinks that he is about to be forced to perform *a*. It is perhaps natural that such a schema should be in place, as one's movements are seldom forced, and still less often would one be aware of a movement about to be forced. That is, if one does have a belief about an imminent movement, it's typically volitional.

The problem is that late "no" decisions are themselves relatively rare, and schematic connections are honed over many similar cognitions. Suppose I oftentimes decide not to perform

an act just prior to movement, when I'm already willing that movement. A cancellation of that volition may well be generated less rationally and more automatically: I may only need to believe that the movement is about to happen for a cancellation signal to be sent to preSMA, or one may be sent in absence of a signaling belief.[190] After all, in this case, late "no" decisions often happen, and thus volition cancellation is often needed. So, for example, once one decides that one shouldn't reach for the coffee mug now (for whatever reason), the veto isn't guided by further ratiocination: *My present volition to reach for it will cause it to happen; I should stop this volition*. This would be a case of IM (see sect. 1), and thus entail a representation of the volition.[191]

But late "no" decisions *don't* happen for the vast majority of mundane acts like reaching for a mug. As I've argued, they tend to be restricted to difficult and/or critical acts, and to situations where a reason to desist happens to arise just prior to movement. So it's quite plausible that, when such a decision does happen, there is no schema in place associating it with a veto. We would then need to react inferentially to our volitional state, which we want to upend, just as we do in those cases of undesired perceptions. And insofar as that practical reasoning is metacognitive in a certain way – i.e., representing the volition as occurrent – it necessarily involves consciousness.[192] Indeed, Libet's own characterization of the reasoning that leads to the veto implies metacognition: It "presumably occurs when a given *W* [the wish to act a certain

[190] Similarly, there would be a schematic connection between a re-approval of *a* and no volition to do *a* arising, as one is already in progress.

[191] It bears reemphasizing that these basic inferential steps need not be conscious, and seldom would be. But their higher-order content need not be conscious to make the volition conscious.

[192] I should add that this hypothesis about the function of HORs in the veto process is empirically falsifiable, in the following way. Let *C* stand for the substrate for the HOR targeting the volition, presumably a certain pattern of prefrontal activation. Let *D* stand for the activation in the anterior frontomedian cortex that the Brass-Haggard results correlated with the late whether decision and veto process. Finally, let *V* stand for the preSMA activation subvening the volition. While *C* and *D* both occur "upstream" of *V*, in the prefrontal cortex, it may be that *C* is neither partly constitutive of *D* nor causally interactive with *D*. In that case, we would have reason to doubt that the HOR is integral to the veto process.

way] is recognized as being incompatible with social acceptability and with one's personality" (2003, p. 25). One is thus thinking about (representing) *W*, a current mental state one is in. However, I would stress that the reconsideration is not about the *wish* to act, it's about the act: one would realize the act is incompatible with one's goals. Rather, the metacognitive reasoning is about preventing the act. That's where one represents one's volitional state and thinks it must be cancelled.

4. Why Conscious Thinking?

While volitions are seldom conscious, what we might call purely intentional states – namely those that (arguably) lack any qualitative character – are often conscious. Propositional attitudes like beliefs, desires, doubts, etc. are conscious throughout cognitions that range from wondering about an issue to step-by-step problem-solving to everyday practical reasoning: I may believe that Sue will come to the party, doubt that she would like to be in the company of my friend Jones, and then desire that she be advised that Jones will be there. But to say we engage in "conscious thought" is not to commit to the implausible claim that *all* the states in a conscious thought process are conscious. Rather, to say that a thought process is conscious need only imply that some of the states are. So I may consciously believe that *P* and then consciously conclude that *Q* while my belief that $P \supset Q$ remains nonconscious, but the latter still factors into what can be called a "conscious" ratiocination. Although folk psychology has accepted the idea of nonconscious thinking, it still tends to hold that consciousness improves thinking. For example, we accord special status to "conscious decisions" as being more reflective and less automatic or

impulsive. Here it seems consciousness has the role of bringing more considerations to bear on a problem, or perhaps increasing the flexibility of our thinking. We also tend to hold that intricate reasoning, such as solving a challenging math problem, *can't* be done nonconsciously. Now, folk psychology does allow that good ideas sometimes "pop into" one's mind, or result from "sleeping on a problem." But in these cases of nonconscious ratiocination, we have done some conscious work on the problem beforehand. At least, there has been a state of conscious *mental effort*[193] toward solving the problem prior to the nonconscious processing – and it may be that the c-property is efficacious at that point. An *entirely* nonconscious solving process, where we're first exposed to the complex problem and the next thing that enters consciousness is the solution, would surely be quite rare. Normally, we're conscious of several reasoning steps or exploratory thoughts, along with exertions of mental effort, in the course of arriving at a solution or decision.

Given the efficacy of the mental, these conscious states would drive the thinking process along with the nonconscious states. But that utility does not entail the utility of the states' c-property, which may be merely associated (perhaps epiphenomenally) with the ratiocination. On the higher-order view, the question would be put this way: Assume that I consciously arrive at some belief that *P* in the course of trying to resolve an issue, and that idea brings me closer to a resolution. Would the state with the content *I believe that P* – the state that renders the idea conscious – be of any use in that deliberation? Prima facie, it does not seem so: For example, given my desire to go to the market, the content of my conscious belief – *the market opens at 10 a.m. on Sundays* – is what leads me to decide on what time I should wake up on Sunday, not the representation of my holding that belief. Or suppose that, in deciding between two products *A* and *B*, I consciously realize that the virtues of *A* outweigh those of *B*. Surely that content, carried

[193] This being a purely qualitative state that Strawson has described in various ways: "setting one's mind at the problem," "focused concentration of will," a "receptive blanking of the mind," etc. See Ch. 4, sect. 7.

by the first-order state, moves me to choose *A*. The content of the higher-order state – that *I have come to believe* that the virtues of *A* outweigh those of *B* – is not needed to reach my preference. In short, it is only the propositional attitude and intentional content of the *first-order* states that seems to matter to decision-making and problem solving. Thus Rosenthal writes that "the rationality of thoughts and desires is a matter of their intentional content. … And since thoughts and desires have intentional content independently of being conscious, rational connections among them will tend to occur independently of whether they are conscious" (2008, p. 832).

When an intentional state enters into rational connections, we can say it is *engaged* in thought; that is, it causes other states in ways that reflect good (though sometimes poor) inferences, instead of being merely disposed to have those effects. So, for example, I presumably have a latent belief in the Pythagorean Theorem, but that belief would become engaged if I'm trying to figure out the length of the hypotenuse for a given right triangle. An intentional state need not be latent before it is engaged, of course: A doubt may arise that I've got the theorem right, and that doubt may cause a rechecking of the formulation, or recollections about a math lesson from grade school, but I did not have the doubt before it had those rational effects. An engaged intentional state is thus defined as one I'm reasoning *with*, as opposed to reasoning *about*. And it seems plausible that reasoning with states can go on without one's reasoning about, or merely representing, any such state.

Now, cases of nonconscious reasoning are widely accepted, and the states in such processes would, ex hypothesi, not be targeted by HORs. But the theoretical considerations I am raising imply a stronger claim: that when reasoning *is* conscious, consciousness isn't facilitating it, as the content of HORs is not rationally relevant to the first-order inferences. We are thus left with a scenario on which conscious propositional attitudes constitute thinking processes, but

their *being conscious* is merely associated – perhaps epiphenomenally – with the ratiocination. The property would not be epiphenomenal, however, if there were a use for HORs in *meta*cognitive reasoning. In line with my discussion of IM during perception and volition, a representation of a purely intentional state would enable one to reason about whether to upend the state, which may mean simply canceling it or reducing it from engaged to latent. Other kinds of metacognition relevant to the control of first-order thinking *processes*, as opposed to individual states, would also be enabled by HORs, as I will discuss.

First, let me address a fundamental objection: Why aren't the state's rational connections to other first-order propositional attitudes enough to guide its cancellation? For example, it seems one would stop believing that *P* (or at least rethink whether *P* is the case) if it occurred to one that *Q*, and one realized that $Q \rightarrow \neg P$. Based on these realizations, one need not conclude that *My holding P makes no sense, I should stop believing that P* or otherwise deploy a representation of one's belief before abandoning it. Neither is a HOR needed to begin deliberating over a belief content: Simply believing that *P*, along with considerations that weigh against *P*, may spur one to question *P* or its likelihood, and in turn preserve or abandon that belief. Similarly, the adjustment of other attitudes toward *P* – doubting that *P*, hoping that *P*, etc. – can occur without representation of those states. Where a HOR *would* be necessarily involved, I argue, is in deliberating whether to continue engaging that state. The issue in that case is not whether the attitude toward *P* is rational, but whether it is rational to let that attitude toward *P enter one's thinking at the present time*. Of course, the former issue is relevant to the latter: If holding a certain attitude toward *P* is irrational, one may very well not want the state engaged. But here we must be wary of "denying the antecedent": If holding a certain attitude toward *P* is rational, it does not follow that one will *want* the state presently engaged. Just because it is rational for one

to believe that 2+2=4, or to wonder whether there is life on other planets, does not mean it is rational for one to have those thoughts now, instead of, say, expecting an oncoming fork in the road. It may well be more rational to let the expectation become engaged, to become causally active in one's psychology and prompt one to decide between the roads as soon as one sees the fork.[194]

Thus, a given attitude toward a proposition *P* (call it mental state *a*) may become disengaged due to two different kinds of ratiocination: (i) Considerations about the truth (or likelihood) of *P*; this requires deploying a representation of *P* but no representation of *a*. (ii) Considerations about whether to let *a* continue to engage one's thinking; this *does* require deploying a representation of *a* (a HOR). Note that in the latter case, *a* may not be cancelled, but simply become latent. So for instance, the ratiocination may be: *Right now, I shouldn't be wondering what to have for dinner; it's better that I focus on what the lecturer is saying*. As a result, I may remain in that state of indecision about dinner, but its rational connections will cease to be active; it will no longer prompt me to consider dinner options, for instance. And deciding that I shouldn't be wondering about *P* requires a representation about my current state of wondering, which entails its being conscious. That representation is needed to ascribe a property such as *being presently undesirable* to my state. As Rosenthal writes, "with doubting, hoping, or expecting that my mental state has some property … I must at least have the assertoric thought that I am in that state" (1997, p. 742). Now, this view seems to entail that a current intentional state can't go from engaged to latent if that state is nonconscious. Considerations that weigh against holding the attitude toward *P* will tend to cause a *cancellation* of the state, not merely a disengagement from current thinking. So it appears that only a deliberation about

[194] This of course assumes a limited capacity of states that can be engaged at once, and indeed working memory – the ability to "hold in mind" information for the purpose of reasoning and learning – is thought to have a limited capacity.

whether to continue engaging the state (entailing higher-order representation) can motivate its relegation to latency, as in the forgoing example. But in fact this is not so: Suppose I am getting certain ideas from the lecture, but suddenly am distracted by someone entering the room in flamboyant dress. Those ideas may become psychologically inactive as I begin to wonder about the person's choice of attire. I certainly need no representations of the thoughts I'm gathering from the lecture in order for that to happen. The key point, then, is that HORs are only needed to disengage a state *on the basis of reasoning about whether to be in it*. Suppose I am working on a timed math test of five problems and I've reached what I think is the solution to the first problem, but I start doubting it's correct. I then realize I only have a short time left to work on the other four problems, and reason: *Instead of doubting whether my work is correct, at this point, I should start thinking about the next problem.* Here it might be objected that one need only reason about what state one should be in, which does not require representing one's current state. But I think it does. Surely such reasoning is premised on the belief that one is not in the right state, and one would come to that belief only by representing what state one *is* in. So my belief that I'm not currently thinking about the next problem comes from knowing that I'm doubting my solution to the first problem.

Whether or not this ratiocination itself is conscious, it deploys a representation of my doubt that would make that state conscious. While that HOR has no bearing on the rational connections the doubt has – e.g., to rechecking my work – it is needed to reason about whether to continue engaging the doubt, or change the direction of my thinking. Thus we can see how higher-order information enables a coordination of first-order states, as Armstrong and Lycan have proposed, although the coordination results from the reasoning that deploys the HOR, not

directly from it. In addition, as I hopefully have made clear, first-order states (specifically propositional attitudes) do not *require* HORs in order to become rationally coordinated.

Just as HORs enable the disengagement of an intentional state via IM, they also enable the cancellation of a first-order reasoning process via IM. Recall that most "conscious ratiocinations" involve both conscious and nonconscious states, so HORs are only directed at some of the beliefs, doubts, and desires that comprise the process. At some points, they may target merely a sense of cognitive effort or confusion instead of a propositional attitude. But even if, taken together, the HORs don't target *all* the states in the process, they certainly provide "snapshots" of what is mentally transpiring. As such, they support the metacognitive belief that one is reasoning or wondering about some issue *x*. On the basis of that belief, along with other considerations, one might conclude that instead of reasoning or wondering about *x*, one should be thinking about *y*. So while only first-order propositional attitudes and contents are relevant to the rationality of the cognition, higher-order contents can be relevant to its conative aspect: information about what one is pursuing mentally is a basis for deciding whether to continue pursuing it. I say "can be" because as long as there is no decision to upend the process, the volition to resolve the issue, solve the problem, etc. (a first-order state) continues to drive the thinking, along with the person's rational capacities. However, Armstrong and Rolls have argued, respectively, that HORs are needed for volition to drive thinking and for one to be able to locate and correct errors in thought. I doubt they are needed for either function, and will address these views in turn.

While a volition may be cancelled because one has decided one should be engaging in different mental activity than what the volition drives, it may alternatively be cancelled once it has "served its purpose." Indeed, Armstrong likens mental to bodily action in that both are

purposive, meaning that they are caused by intentions. And central to the operation of will is a "feedback" system that can determine when the goal has been reached and then cancel the intention (1993, pp. 138-144). Armstrong argues that in the case of mental action, the reaching of the goal must be introspected, which on his HOP view entails that state's being conscious. He gives the following example:

> An intention to work out a certain long-division sum purely in one's head is a mental cause that initiates and sustains mental activity. We are informed of the results reached at each stage of the activity by introspection (just as in the case of physical action we are informed by perception). The introspectively acquired information reacts back upon the mental cause, so that further steps in the calculation are made in accordance with the currently reached mental situation. Finally, recognition that the answer to the sum has been obtained "switches off" the sustaining cause of the activity. (1993, p. 162)

Introspection thus drives the calculation by enabling feedback: the intention to solve, say, 330 / 15 continues to cause stages in the calculation (along with rational connections to previous stages) as long as introspection of the current stage does *not* reveal information about a number that the one thinks equals 330 / 15. It might be held that Armstrong's picture should be refined to include a "main intention," e.g., solve 330 / 15, and subintentions: solve 33 / 15, solve 33 - 30, etc. The mental states that follow from these subintentions would then not need to feed back to the main intention, but only to the subintention; once a subintention is satisfied and dismissed, the current mental state would again feed back to the main intention, again via introspection. In any case, it is clear that Armstrong's picture requires regular introspection of the mental states

constituting a ratiocination in order to keep it going, and following HOP theory, those states become conscious. But given that thinking can proceed *nonconsciously* for significant intervals of a "conscious ratiocination," this picture is problematic. In what follows I argue that what must be fed back to the intention is not information about the current state (which entails higher-order representation), but rather information about the state's *content*, both conceptual and referential.

Armstrong's notion of feedback is akin to what I referred to as the will/act consistency judgment in Chapter 4. As discussed, I think these judgments regularly occur, although volitions are seldom conscious. Thus, the judgments do not require representing the volition, but rather the volitional content as a means to determining the act's agreement with it. So I don't need to represent that *I will that I grip the mug now* in judging the agreement of some bodily movement *x*, but only deploy a representation of my gripping the mug. That same representation is a volition to move just when it tends to bring about its intentional object via my motor system (presumably the representation has that tendency when it occurs in preSMA). So for any representation *a* of bodily movement, *a*'s having that causal disposition constitutes my being in the propositional-attitude state *I will that I do a now*. Judging consistency thus requires the following representations: *a*, which represents the desired bodily movement, and a thought about what I've actually done, which is based on a perception of what I've actually done. None of these are HORs.

Volition and consistency judgment in the case of mental action are given a parallel analysis, except that the feedback need only be based in thoughts about the *content* of the mental act, not the act itself. This distinction clearly doesn't arise for bodily acts, which have no representational content; so representation of the act in that case is required. In the mental scenario, the volition represents a mental act as willed insofar as it tends to bring about that

intentional object via one's higher cognitive systems (e.g., dorsolateral PFC). Whether the mental act occurs will of course also depend on one's cognitive abilities, just as whether the bodily movement occurs doesn't depend solely on the volition to move, but also on one's physiological abilities. As I have argued,[195] the volition picks out the desired intentional state in terms of the state's intentional object and its mode of presenting that object, the concept(s) under which the object is thought. The content of the volition to perform a mental act will thus have the form *think of o under concept c*. For example, suppose that Problem #4 requires me to solve for the area of a 5-by-5 inch square. My volition can't merely have the content *think of the area of a 5-by-5 inch square*, since "the area of a 5-by-5 inch square" picks out 25, so I would be performing the act just by having the volition. Also, I could satisfy that volition by simply thinking *the solution to #4*. Clearly this won't do; what I want is to think of 25 under a concept that not only picks it out, but expresses it: the concept *25*. So my volition must have the content *think of the area of a 5-by-5 inch square under its numerical concept.* What would it be for me to judge that a current intentional state *x* of mine agrees with that volition? Armstrong would say that introspection of *x* is needed, a higher-order state with the content *I am thinking of 25 under the concept 25*. That content will allow me to judge consistency: *My thinking of 25 under the concept 25 is a numerical representation of the area of a 5-by-5 inch square*.

But in fact feedback only requires a *subjective measure* that I'm in the intentional state I want to be in, which needn't be a *judgment* to that effect. Since the identity of my state is fixed by its referential and conceptual contents, if I judge that these contents are as I want them to be (as my volition stipulates), it will be subjectively established that my state is consistent with my volition, without my thinking about my state. The two content constraints given by my volition in the present example are that my ensuing thought should (i) represent the area of a 5-by-5 inch

[195] See Ch. 4, sect. 6.

square (ii) under its numerical concept. So suppose I then enter intentional state *x*, which is a thinking of *o* under concept *c*. The feedback process for *x* consists of the following two judgments about *x*'s content, neither of which involves a representation of *x*: First, I judge that *o* is the area of a 5-by-5 inch square. This subjectively establishes that *x* satisfies (i), as I believe that *o* is what I want to be thinking of. Second, I judge that *c* is the numerical concept for *o*. This subjectively establishes that *x* satisfies (ii), as I believe that *c* is the kind of concept I want to deploy. What results from both these judgments is a subjective measure that the state I'm in, *x*, is consistent with my volition. So the volition can be cancelled on the basis of those judgments, which is the function of feedback for Armstrong. Similarly, suppose my volition has the content *think of the color of Jane's dress at last week's party under its color concept*. I then think of red, under the concept *red*. The feedback process consists of my thinking that *red was the color of Jane's dress at last week's party* and that red *is the color concept for red*. In general, the feedback for any thought *t* (which is a thought of *o* under concept *c*) to volition *v* (which has the content *think of o* under concept c**) can be defined as follows: I judge that *o* is *o**, and that *c* is *c**. These first-order thoughts subjectively establish that *t* agrees with *v*. So I need not think that *t* agrees with *v*, which entails *t* is conscious (if it is represented as a current state). In sum, the feedback process is indeed a coordination of first-order states, enabling the cancellation of the volition to have a certain thought once that thought occurs. Feedback also drives the ratiocination, guiding its continuation when a solution hasn't been found. But feedback itself can be carried out via first-order states, contra Armstrong.

Likewise, only first-order judgments are needed to locate errors in a reasoning process. These would be judgments about the *content* of the states in the reasoning process, not about the states themselves. Rolls argues that "by thinking about lower order thoughts, the higher order

thoughts can discover what may be weak links in the chain of reasoning at the lower order level, and having detected the weak link, might alter the plan, to see if this gives better success" (2004, p. 151). The "weak link" would presumably be some belief that $F(a)$, that $P \supset Q$, etc., that is part of the ratiocination and is discovered to be poorly supported or false. In detecting the error, one realizes that *a* is not *F* (or probably not *F*), that *Q* doesn't follow from *P*, etc. That is, one corrects one's attitude toward the *content* of the false or poorly supported belief, which need not involve metacognitive thought. Instead of thinking *My belief that F(a) is false* one simply thinks *'F(a)' is false* or *a is not F*, and accordingly, adjusts or abandons the ratiocination that includes the weak link. One comes to such a realization as considerations arise that weigh against believing that $F(a)$ or as Rosenthal puts it, via the belief's "first-order dissonance with other antecedent beliefs" (2008, p. 836). Or the corrective thought process may be metacognitive in representing belief states generally, but not one's being in any such state: *Believing that P conflicts with believing that Q, and the belief that Q is better grounded*.

Note that if Rolls' theory were true, it would teleologically explain the correlation that seems to hold between state consciousness and thinking that challenges the individual, such as solving a difficult problem: Errors are most likely to occur in that case, and if HORs are needed to detect errors, it's no wonder they tend to occur as well. But there is an alternate teleological account of the correlation. Perceived progress on a complex problem can naturally be unsatisfactory, and if so, that realization ("progress assessment") may lead one to abandon the effort, perhaps in favor of a different activity (mental or otherwise) that may be more productive at that time. The feedback mechanism, which as I've argued is entirely first-order, does not have the effect of a progress assessment: It only cancels a volition to think of *x* under concept *c* when it determines one has done that; it does not cancel the volition when one is having difficulty

satisfying the volition. Nor does the feedback mechanism provide information on whether a subintention has been satisfied, which would be the kind of data relevant to a progress assessment. As discussed, the feedback for any thought *t* (which is a thought of *o* under concept *c*) to volition *v* (which has the content *think of o* under concept c**) would be constituted by the judgments that *o* is *o**, and that *c* is *c**. Neither is a higher-order judgment that *t* is the thought that *v* represents (where *v* is a subintention). But such a judgment is integral to an assessment of how well my thinking is going, how far along I'm getting toward a solution.

A progress assessment would involve more than a judgment of whether a single subintention has been satisfied, of course. One would consider how many have been satisfied, over what time period, and with what degree of cognitive effort. For instance, if one's intention to mentally solve 330 / 15 is followed by the subintention to solve 33 / 15, and one realizes that it has taken a bit too much time and effort to think of 2.2, the main intention might just be cancelled as one opts for a calculator. Now, many volitions are not followed by subintentions. For example, in trying to recall the name of an actor from a movie, or deciding whether a painting counts as "surreal," or determining whether a prospective campsite is advantageous, one may not identify more specific volitions to fulfill, but rather open one's mind to promising ideas on the issue: Respectively, a thought that the actor's name starts with an "T", that the faces in the painting are distorted, that there is a good amount of shade at that site, might arise and seem as if they could lead toward the objective. A progress assessment would then not identify these thoughts as satisfying subintentions, but just as being promising; but here again, representations of one's having such ideas must be deployed. And rather than identifying ideas one has had that failed to satisfy subintentions, the assessment may simply register states of mere cognitive effort or even confusion over some time period following the volition.

In brief, I argue that HORs not only support the metacognitive belief that one is reasoning or wondering about some issue *x*, enabling one to think, e.g., *Instead of thinking about x, I should be thinking about y*. HORs also support beliefs relevant to an assessment of how well one's thinking is progressing, essentially beliefs that certain thoughts have or have not satisfied volitions, seemed promising, etc. That progress assessment in turn may be the basis for a decision to alter or continue one's present mental activity, e.g., *My current efforts to solve issue x aren't going well, I should do something else for now*. Carruthers has described this function similarly in his RT (reflexive thinking) theory: "A faculty of reflexive thinking would get us the ability to think about, and hence modify and improve upon, our own thoughts and patterns of thinking on a regular basis. … By thinking about what we have just thought, we are able to assess our thoughts for truth, plausibility, and appropriateness. … And by thinking about the manner in which we have been trying to think about a problem, we can sometimes see the possibility, or the need, of approaching it differently" (1996, p. 200). But Carruthers' characterization of the use of metacognition is too broad. First, as I have argued contra Rolls, we don't need HORs to assess our thoughts for truth or plausibility, as we can just consider the propositional contents apart from the mental states that serve as vehicles for those contents. Nor do we need HORs to register the appropriateness of a thought to an intention; the feedback system handles that, as discussed. And we can also change the course of our thinking – "approach a problem differently" – on the basis of first-order judgments that certain thought contents are false or implausible. In particular, one may dismiss a subintention as misguided without representing it as a mental state one is in, e.g.: *Trying to determine the product of pi and the radius isn't the right way to get the area of the circle; first the radius must be squared*. But in order to decide whether to continue one's *mental act* of trying to work out the problem, one

judges whether subintentions seem correct, whether thoughts are satisfactory, etc. Here one *does* represent the mental states, as one is not considering the problem per say, but one's solving efforts.

It may be objected that beliefs that certain thoughts have or have not satisfied subintentions, seemed promising, etc., deploy HORs of *past* states, so any utility their h-properties have is not relevant to that of c-properties.[196] Only a HOR's representing its target as *occurrent* will make that state conscious, according to higher-order theory. But surely, a progress assessment that takes place at some time *t* after the volition occurs will be based not only on representations of one's mental states during the interval up to *t*, but also on a representation of one's mental state at *t*, which is the furthest point of progress. An idea that satisfies a subintention, or one that seems promising, may well be one's latest. So the HOR targeting that thought would be relevant to assessing progress in virtue of a judgment about the thought's rational contribution entering into the assessment. That judgment (again, a case of IM) would utilize the h-property of the HOR that renders the first-order thought conscious.

Given that such metacognitive judgments are more useful during complex thinking, where progress may be unsatisfactory, we have an explanation for the prevalence of conscious states during such thinking. Such states provide us access to both "the contents and occurrences of our acts of thinking," as Carruthers puts it (1996, p. 202). Of course, as discussed in sect. 2, for Carruthers we don't access the occurrences of our thoughts in virtue of actual HORs, but rather in virtue of those states' access to the HOR system, which makes those states conscious. So he is only committed to HORs occurring when IM does, not when consciousness does. But as I've argued in the case of conscious perception, it is plausible that there is a basic, ongoing IM

[196] It might be argued that these representations are memories, which require the objects of memory to have been represented at the time they occurred. But see Ch. 4, n. 147.

that deploys HORs, conforming to the prevalent HORs posited by the actualist. In the case of conscious reasoning, a HOR about one's present state is useful for a progress assessment, as explained. Now, such an assessment is a fairly simple kind of IM, classifying current reasoning as successful or not, promising or not, in accord with one's general metacognitive desire to be productively engaged mentally. As such, it may be fairly common during reasoning or wondering about a problem, thus routinely utilizing HORs about present states.[197]

It follows that consciousness is not *necessary* to (first-order) thinking, complex or otherwise; rather, it is necessary to thinking about the current stage in one's thinking, which is useful in making rational decisions about whether to continue that mental act. Essentially, it adds another layer of cognitive flexibility. This result is inconsistent with what appears to be the folk psychological view: Intricate reasoning, such as solving a challenging math problem, *can't* be done nonconsciously, presumably because consciousness is needed to somehow drive the inferences.[198] On my view, since consciousness is not integral to (first-order) thinking, it's theoretically possible for complex thinking to proceed nonconsciously. It may not be *empirically* possible in some cases due to the strong tendency for HORs to attend such thinking and be deployed in IM, which, as I've argued, is especially useful during complex thinking.

My view is thus consistent with the results of an important study by Ap Dijksterhuis et al. (2006) that presumably shows that complex deliberation can not only proceed nonconsciously, but tends to be more successful when performed that way. The study focused on consumer choices between products with many attributes to be compared, such as cars or furniture, and success was defined in terms of participants' satisfaction with their choices. The deliberation in

[197] I am not suggesting that *all* HORs that transpire during thinking or perception are put to use in IM; some may occur due to mere habituation. My claim is that the actualist is not saddled with a great many HORs that go unused.
[198] Now, it may be that a metacognitive "progress assessment" leads one to postpone a complex ratiocination to a later time when (for whatever reason) it proceeds with more success. But that is not a case of consciousness improving one's rational ability by (directly) causing better inferences to be made.

question is thus "complex" in virtue of the *amount* of information to be processed, not in virtue of the *inferential* difficulty involved. In the first of the four experiments in the study, for example, "conscious thinkers" would choose between cars characterized by four attributes (simple) or by twelve attributes (complex) after deliberating on the respective choices for four minutes. "Unconscious thinkers" would deliberate over the same choices for four minutes while being distracted (they solved anagrams). The key finding was that conscious thinkers reported less post-choice satisfaction than unconscious thinkers when it came to the complex choice. Dijksterhuis et al. explain this result through their Unconscious Thought Theory (UTT), which claims that "conscious thought suffers from the low capacity of consciousness, making it less suitable for very complex issues." In contrast, "during unconscious thought, large amounts of information can be integrated into an evaluative summary judgment" (p. 1006). A second hypothesis of UTT is that when the complexity of thought is founded in rule-following (e.g., arithmetic or logical rules) as opposed to the integration of large amounts of data, "conscious attention is necessary. For example, one cannot do arithmetic without conscious attention" (p. 2006). Choosing between the cars characterized in terms of a dozen attributes, then, is not a matter of following a train of inferences over the four minutes, but rather weighting and comparing the various attributes and arriving at an estimation of which car has the most desirable attributes overall.[199] That's why, ex hypothesi, the unconscious thinkers were better at making these estimations.

On higher-order theory, the claim that conscious thinking has a lower capacity than nonconscious thinking suggests that HORs reduce the capacity of first-order thinking. If we assume that greater capacity is achieved through the production of more numerous FORs in order

[199] Some relatively simple inferences may be involved here: "If the car has an airbag, I'll be safer. Car A has an airbag…" etc.

to integrate many bits of data, we can speculate that the added causal activity required to produce HORs targeting (some) of those FORs interferes to a degree with the production of FORs. This would be a matter to be decided empirically, based on information about the neural substrates for these cognitive systems and their causal interaction. IMT only entails a "trade off" for the unconscious thinker: Her first-order thinking would (presumably) have a higher capacity, but she would lose a certain metacognitive ability that is enabled by HORs, namely to think about that first-order process, its progress, and whether to continue it. Thus, IMT is consistent with the results of Dijksterhuis et al.: It does not deny that complex nonconscious thinking ("deliberation without attention") is possible, nor does it preclude the processing advantage that UTT attributes to nonconscious thinking (the "deliberation-without-attention effect").

However, as to UTT's claim that conscious attention is needed for complex rule-based thinking, IMT only implies that they would tend to be correlated, since HORs enable a progress assessment that is useful in such a case. In fact, I argue this correlation tends to hold even for thinking that is complex in terms of data integration, like the 12-aspect car comparison in Dijksterhuis et al.'s first study. The fact that participants were distracted by solving anagrams doesn't guarantee that *no* states relevant to their decision process were conscious over the four minutes, only that fewer states were. But the improved choices under those conditions offer evidence that conscious attention (i.e., higher-order representation) isn't needed *at all* for that type of complex thinking. Now suppose participants did arithmetic under the same distraction condition. Their performance (in terms of correctness of response, speed of calculation, etc.) at the main task would surely worsen. Would this support the view that conscious attention *is* needed for complex inferential thinking, as UTT also claims? I don't think so. Under the distraction condition there is less conscious attention to the main task, and on higher-order theory

that means *two* things are missing: (i) HORs targeting many of states in the main task; and (ii) the exclusive focus of first-order states on the arithmetic (some are involved in solving anagrams). So the worsened performance on the main task might be due to the divided attention at the first-order level, not to the reduction in HORs targeting states in the arithmetic solving process. On the other hand, the improved performance on the car comparison task when both (i) and (ii) are missing clearly suggests that no HORs targeting the states in the main task are required.[200]

In sum, Dijksterhuis et al. provide good evidence that the integration of large amounts of data does not require conscious attention, a claim that is compatible with the role that IMT posits for consciousness in thinking. But better support is needed for their claim that conscious attention is needed for inferentially complex thinking: The general correlation between the two may be due to the mere *usefulness* of evaluating the progress of complex thought – and indeed the ongoing occurrence of (fairly simplistic) IM during such thought. HORs on my view, then, are not involved in the *progress* of first-order thought, but merely in the assessment of that progress. In contrast, Armstrong and Rolls have argued that HORs *are* required for that progress, in different ways: (i) the abortion of volitions once satisfied in the course of one's thinking, and (ii) the detection of erroneous thoughts or inferences. I have argued that (i) and (ii) can occur nonconsciously, thus attributing (correctly, I think) more sophistication to nonconscious thought.

[200] It also suggests that exclusive focus at the first-order level isn't needed for data-integration tasks.

5. Conclusion

On first-order theories of state consciousness, it is usually clear that a state's c-property serves a significant function. For example, if a perception's c-property reduces to that state's having a certain type of first-order content, say nonconceptual, it is arguable that the perception thereby has a distinct causal potentiality. On functional theories, the identification of the property with a certain access relation, say access to the executive system, entails that a conscious perception has a stronger cognitive impact. But on higher-order views, state consciousness isn't necessary to any form of first-order processing. Being about a mental state, a HOR does not serve to gather information about the environment, will a bodily action, or think about any state of affairs outside of one's own psychology. Neither is there reason to think that HORs somehow facilitate first-order perceiving, willing, or thinking. For example, why should a representation of a perception render its first-order content more cognitively accessible?

Rather, I have argued that state consciousness is necessary to *meta*cognitive processing, and more specifically to reasoning about one's current mental state. The basic idea of IMT has been mentioned in passing by Robert Lurz. Higher-order states, writes Lurz, "enable us to think and reason about our own intentional states" (2004, p. 246). But there are straightforward objections to the claim that this is the function of state consciousness, which I have attempted to address. First, it may be held, correctly I think, that introspectively it doesn't seem we often engage in IM – often enough to make it a significant function for HORs.[201] Here, I noted that IM may often transpire nonconsciously, and the HOR(s) deployed, despite being nonconscious themselves, would still render their target states conscious. Second, it may be objected that IM

[201] As Rosenthal points out, "We seldom if ever operate psychologically on our own thoughts and desires as we think about the objects of our perceptions" (2008, p. 832).

need not deploy HORs about current mental states one is in – the kind of HORs that entail consciousness – but rather representations of past states or general mental traits. I have responded to this point with examples of how reasoning about one's current perceptions, volitions, and purely intentional states is useful. In addition, it may be argued that IMT is more compatible with a dispositionalist higher-order theory, in that (i) IM occurs irregularly and (ii) state consciousness subserves IM as the *ability* to engage in thought about FORs. In contrast, actualist theory posits that state consciousness is constituted by the ongoing occurrence of HORs, which largely fails to subserve IM. In defending actualist theory on this point, I have argued against (i): it is plausible that HORs are regularly deployed in a basic level of IM that assesses current thoughts and perceptions for compatibility with one's general metacognitive beliefs and desires.

Based on IMT, I have also endeavored to explain three correlations that generally hold between state consciousness and types of first-order cognition. The first is between conscious perception and executive access to first-order perceptual content. I've argued that the latter allows planning one's reaction to external objects, and information about one's perceptual state can be relevant to that planning. The second is between conscious volition and acts the agent perceives to be difficult and/or critical. In these cases, a veto or late "no" decision may be especially useful, and representations of one's current volitional state plays a role in the veto process. The third is between conscious thoughts and challenging first-order ratiocination. Here, I've argued that a progress assessment is particularly useful, and that takes into account (inter alia) information about one's current state in the reasoning process, provided by a consciousness-conferring HOR.

VI. CONCLUSION

My main objective in this project has been to give a theory of the cognitive function of consciousness, but I have first addressed the issue of causal role, which is logically prior to that of function. If consciousness has utility – if it is beneficial for a creature to have conscious mental states – such states must have a causal impact qua being conscious. But merely establishing that the c-property is efficacious does not entail that it has *beneficial* effects. Its effects may be disadvantageous or merely insignificant, in which case we might expect consciousness to be eventually selected out in the course of evolution. The more fundamental claim that the property is efficacious may be challenged in two ways, which I elucidated in Chapters 2 and 3, respectively. First is the claim that a mental property such as consciousness is not the sort of thing that *can* affect cognition, given plausible views such as externalism about mental content, anomalism of the mental, and the causal exclusion of the mental by the neural. The second kind of challenge allows that the metaphysical nature of consciousness is consistent with its efficacy, but denies that it actually causes; that is to say, it happens to be epiphenomenal. We can call these the issues of *causal viability* and *de facto efficacy*. Just as de facto efficacy is logically prior to utility, viability is logically prior to efficacy: If consciousness is de facto efficacious, it must be metaphysically possible for it to affect cognition. But merely establishing that the c-property is causally viable does not entail that it affects the mind/brain in the actual world. Given my position that phenomenal consciousness is higher-order representation and has utility as such, I will review my points on these two issues from that theoretical perspective.

As to causal viability, the view that consciousness can be reduced to representation, and in particular higher-order representation, has several important implications. Assuming that

representational content is wide, it becomes necessary to argue that wide content can have "local" effects, or that there is such a thing as narrow representational content, in order to preserve the causal viability of consciousness. The view that consciousness is *internally directed* intentionality does not avoid this issue, as I argued in Chapter 2. Neither does the higher-order approach allow the nonreductive physicalist to avoid the "qua issue" once the causal exclusion problem has been solved. The causal viability of all supervenient mental events does indeed entail the causal viability of HORs, but not necessarily that of their higher-order representational properties (h-properties). And the causal viability of the h-property is critical to that of the target state's c-property, as discussed in Chapter 2, sect. 2.[202] The putative anomalism of the mental, however, is largely a mere prima facie problem for the causal viability of consciousness on the higher-order view. I've argued that even nonstrict laws subsuming consciousness aren't discernible due to the variety of mental processes it attends, and the comparative uniformity of the property's introspective appearance. But the anomalistic mental appearance of the property turns out to be misleading if consciousness is higher-order representation, as consciousness then *does* arguably participate in nonstrict regularities that can ground its efficacy. For each conscious state, consciousness is the same type of phenomenon – higher-order representation – but its *tokens* differ in each case: a conscious state *x* will be targeted by a representation of *x*, a conscious state *y* by a representation of *y*, and so forth. As follows from my arguments in

[202] On an alternative theory where HORs are sufficient for consciousness, I've argued the h-property's efficacy is critical to that of a certain phenomenal property instantiated by a HOR. If the HOR is about an intentional state representing *x*, the phenomenal property will be: *making the subject consciously represent x*, or *creating something it is like for the subject to represent x*. If it is about a purely qualitative state such as an emotion, feeling *y*, the phenomenal property of the HOR will be: *making the subject consciously feel y*. The HOR would have these properties solely in virtue of its h-property, ex hypothesi.

Chapter 5, these token HORs would fall into nonstrict correlations with reasoning about *x*, reasoning about *y*, etc. and therefore with the *results* of those metacognitive ratiocinations.[203]

As to the issue of de facto efficacy, I first approached it in Chapter 3 by examining what it is for consciousness to be epiphenomenal. With regard to higher-order theory, a HOR is epiphenomenal to a mental process *T* iff it is causally irrelevant to *T*'s outcome, *T* causes the HOR, and the HOR is "secondary" to the outcome, relative to our teleological intuitions.[204] Furthermore, the HOR is epiphenomenal *relative to the mind/brain* iff it is also causally irrelevant to the outcome of every other mental/neural process. If *every* HOR meets these criteria with regard to some base process, then consciousness is epiphenomenal. Note that this scenario is consistent with HORs being causally viable; the states just happen not to have effects due to the causal structure of the mind/brain. The epiphenomenalist scenario is also consistent with HORs being *nomologically necessary* to their base processes; they can be shown to be causally irrelevant to the outcomes via Lewis' counterfactual approach. Per my argument in Chapter 3, the exception would be a HOR that serves as a controlling effect of its base process; such a state would count as causally relevant to the outcome on Lewis' analysis. For example, if the base process for HORs targeting perceptual states is executive access to those states, then that access would be blocked – when needed – via the inhibition of the relevant HORs.

But for consciousness to have a mechanical role of this sort in first-order information processing is implausible if it is constituted by higher-order representation, for then the phenomenon carries *its own kind of information* that is advantageously processed, namely that

---

[203] Note that if consciousness is involved in such rationally mediated regularities, then its efficacy would potentially conflict with the physical laws that govern neural processes, per the argument by Davidson I discussed in Ch. 2, sect. 5. I tend to favor the approach to this problem offered by E.C. Tiffany (2001) as well as other authors, namely that, for all we know, physical laws may happen to preserve the rational connections between propositional attitudes. So the conflict is merely potential.

[204] It is natural to regard the first-order processes that yield a HOR – perhaps access to working memory in the case of perception – as being more central to cognition, so this last criterion may well be met.

some particular first-order process is transpiring. This basic idea underlies the theory of function I developed in Chapters 4 and 5. To briefly review one example from Chapter 5, the information that a volition is occurring may be processed via what I have described as an inferential reaction to being in a volitional state. Specifically, the HOR-involving belief that one's act is about to happen *due to one's volition* can combine with other reasons to yield a "late" decision either to desist from the act (by canceling the volition) or go ahead with the act (by not canceling the volition). I've argued that such rational processing is advantageous particularly in the case of difficult and/or critical acts, where reasons to desist from an act may arise at the last moment. This explains why HORs targeting volitions – and thus conscious volitions – tend to arise prior to such acts. But for the vast majority of acts, "late whether decisions" are unnecessary, and so volitions tend to remain nonconscious. This entails that *conscious* volition doesn't play a role in causing our ongoing feeling of agency over our movements. And as argued in Chapter 4, the typical nonconsciousness of volition also entails that (i) the judgments Wegner identifies as grounding felt agency must not require conscious volition, but only a volition's representation of the act[205]; and (ii) the judgments themselves must usually be nonconscious. If consistency assessments (for example) *were* routinely conscious, they would be targeted by HORs: one would routinely represent one's judging acts to be consistent (or inconsistent) with volitions. Thus one would represent the volitions (inter alia), and they would routinely be conscious.

But there is normally no reason to think about *one's judging* consistency, priority, or exclusivity, hence no representations of those mental acts usually arise. On the other hand, the control quale that often results from those judgments and attends many of one's movements is routinely conscious: it is represented, yielding the feeling of agency. And here there *is* a reason

[205] Note that while the priority judgment does require representing the volition, surely this judgment occurs at the time of the act; at that point, the volition is represented as *having occurred* just before the act. And if it is not represented as a present state, it won't become conscious.

one thinks about the first-order mental event: it is pleasurable, and a desire for its recurrence will naturally arise. That desire represents the quale, as would a thought that a certain act is its cause, which informs how to make it recur. Such states utilize the HOR's h-property, its targeting the quale. And that utilization is necessary to (if not also sufficient for) the efficacy of the quale's *being represented*. This account of the utility of HORs with regard to an affective mental state can, I think, be straightforwardly extended to other such states, such as nervousness, elation or frustration. Assuming they are conscious, the idea would be that they are targeted not merely by HORs, but HORs deployed in valuations of the states, desires about whether to be in them, beliefs about their cause, and so on. As argued in Chapter 5, metacognition proceeds similarly in the case of perceptions: one regularly categorizes current perceptual states as advantageous or not, pleasurable or not, and then (perhaps) reasons about how to alter or preserve the states. Comparably, one can reason about whether to continue engaging a purely intentional first-order state like a doubt via a HOR about that state, or assess the progress of first-order thought via HORs targeting states in that process.

IMT combines these plausible examples of metacognition concerning occurrent mental states with the assumption that HORs underlie phenomenal consciousness, thus ascribing phenomenal consciousness a significant use. I haven't made a case, then, for the higher-order theory of phenomenal consciousness; that is a separate project that has been pursued at length by others. But I have in effect responded to one possible objection, namely that the theory leaves the utility of consciousness in question, unlike first-order and functional theories. That utility does not follow, I should add, from the fact that higher-order theory can block the possibility of "zombies," namely creatures that physically duplicate us but lack phenomenal consciousness. Presumably such creatures would lack HORs, and therefore whatever neural states HORs reduce

to or supervene on. So they *wouldn't* be physically just like us. Yet the metaphysical impossibility of zombies doesn't guarantee the efficacy of consciousness: even if physical, or "fixed" by the physical, HORs could be *epiphenomenal* physical states. Or, if HORs must have *some* effects simply in virtue of being physical, these might be slight and result in no discernable behavioral differences. If we were then to perform the "subtraction" of consciousness per the zombie thought experiment, we would still be left with a *behavioral* duplicate of ourselves. We would also be left with a creature that is very close to a cognitive duplicate, if thinking, perceiving, and willing remain intact at the macro-level with the subtraction of HORs and their effects.

This would mean Conscious Inessentialism, as defined by Flanagan, is true: "the view that for any intelligent activity *i*, performed in any cognitive domain *d*, even if we do *i* with conscious accompaniments, *i* can in principle be done without these conscious accompaniments."[206] These "activities," Flanagan notes, are behaviors, so the claim is for input-output equivalence sans consciousness. But this possibility is ruled out if we assume IMT, as there are behaviors that do require consciousness on that theory, namely those that result from inferential metacognition, what we may call *IM-behaviors*. Consider two creatures, *A* and *B*, who are physical duplicates save for the fact that only *B* has conscious states, i.e., *B* has HORs targeting some of its first-order states. *A* would thus not qualify as a zombie version of *B*. The question is whether *A* would be not just physically, but also behaviorally distinct from *B*, if we assume they receive the same sensory input at time *t*. Let us assume that at *t*, *A* and *B* are each in perceptual state *p*, volitional state *v*, and propositional attitude state *a*, but only *B* has HORs targeting *p*, *v*, and *a*. Based on my arguments in Chapter 5, only *B* would be capable of assessing the desirability of *p*, and thus (perhaps) acting on the environment so as to alter *p*; realizing that

[206] Flanagan and Polger (1995), p. 313.

she is about to move voluntarily in a certain way – the way represented by *v* – and thus (perhaps) desist from that movement at the last moment; and considering whether she should be engaging *a* at *t*, and (perhaps) alter her course of mental activity as a result. That is, only *B* would be capable of IM-behaviors, and so *A* would not necessarily be input-output equivalent to *B* at *t*.

It might be counter-argued that Conscious Inessentialism assumes only the metaphysical or logical possibility of subtracting consciousness and preserving input-output equivalence, and so we can support the theory by simply imagining a possible world where *A* matches any IM-behaviors of *B* without *A* representing any of her mental states, thereby showing that consciousness is inessential to those behaviors. But showing inessentiality is not that simple. Presumably, physical laws governing neural states underlie the psychological process of IM that takes us from certain first-order states to certain behaviors, and in the possible world being imagined, those laws would be "adjusted" so that IM is no longer causally involved. But that would not show that HORs are inessential to those behaviors *in this world*. Since physical laws are part of a world's identity conditions, we would be imagining HORs as inessential in a different sort of world.[207] And if we were merely concerned to show *possible* inessentiality, we could show it just as well for first-order states, such as desires and volitions: there is a possible world where all of our behaviors issue directly from our perceptual and belief states without desires or volitions occurring.

So to review, the zombist claims that the actual physical world could remain as it is without consciousness, since that scenario is conceivable, and so consciousness is nonphysical. But the physicalist must argue that the scenario is inconceivable, or that its conceivability doesn't entail its metaphysical possibility. As a higher-order theorist in particular, she would be

[207] Andrew Bailey stresses that the zombie scenario, if it is to threaten *actual* epiphenomenalism, must reflect "a final completed physics," "everything that the physicalist takes to be required for physicalism to be true" (2009, p. 132). And that includes actual-world causal laws.

committed to the zombie scenario being impossible since the neural states that are, or subvene, HORs wouldn't occur in that possible world, so it wouldn't physically duplicate ours. The Conscious Inessentialist might then claim that creatures in that world – which physically duplicate actual-world creatures save the lack of those neural states – would nevertheless be our input-output equivalents, on the assumption that HORs make no behavioral difference. But I argue that the nonconscious creatures would be incapable of IM-behaviors, assuming, of course, duplication of the relevant actual-world laws.

I take IM-behaviors to add to a creature's fitness – its ability to deal successfully with its environment. Only creatures with conscious states can, through metacognitive reasoning, optimize their current perceptual states, desist from acts they are currently willing, and alter or abort current thought processes in favor of more promising activities. With this theory in mind, we can address the issue of why phenomenal consciousness evolved. As Flanagan puts it, "Why did evolution result in creatures who were more than informationally sensitive?" First, on higher-order theory, evolution did *not* result in such creatures, as phenomenal consciousness is just a type of informational sensitivity: sensitivity to information about one's own mental states. But we might then ask, why did evolution result in creatures who were sensitive to more than first-order information (i.e., information about "external" conditions)? I expect that this project has given a plausible answer to that question.

BIBLIOGRAPHY

Akins, K. and Dennett, D. (1986). Who may I say is calling? *Behavioral and Brain Sciences* 9: 517-518.

Armstrong, D.M., ed. (1965). *Berkeley's Philosophical Writings*. New York: Collier.

Armstrong, D.M. (1993). *A Materialist Theory of Mind*. London: Routledge.

Armstrong, D.M. (1997). What is consciousness? In Block et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

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

Baars, B.J. (1997). In the theatre of consciousness: global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4: 292-309.

Baars, B.J. (1997). Some essential differences between consciousness, attention, perception, and working memory. *Consciousness and Cognition* 6: 363-371.

Bailey, A. (2009). Zombies and epiphenomenalism. *Dialogue* 48: 129-144.

Bauer, R.M. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the guilty knowledge test. *Neuropsychologia* 22: 457-469.

Bayne, T. and Levy, N. (2006). Conscious intention and the sense of agency. In Sebanz, N. and Prinz, W. (Eds.), *Disorders of Volition*. Cambridge: MIT Press.

Blakemore, S.J., Frith, C.D. and Wolpert, D.M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *NeuroReport* 12: 1879-1884.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9: 261-325.

Block, N. (1997a). On a confusion about a function of consciousness. In Block et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Block, N. (1997b). Inverted earth. In Block et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Block, N. (2005). Two neural correlates of consciousness. *TRENDS in Cognitive Sciences* 9: 46-52.

Block, N. (2009). Comparing the major theories of consciousness. In Gazzaniga, M. (Ed.), *The Cognitive Neurosciences IV*. Cambridge: MIT Press.

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis* 3: 419-431.

Brown, L., ed. (1993). *The New Shorter Oxford English Dictionary*. Oxford University Press: New York.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy* 4: 73-121.

Bush, G. (2004). Multimodal studies of the cingulate cortex. In Posner, M. (Ed.) *Cognitive Neuroscience of Attention*. New York: Guilford.

Carruthers, P. (1996). *Language, thought and consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. (2000). The evolution of consciousness. In P. Carruthers and A. Chamberlain (Eds.), *Evolution and the human mind*. Cambridge: Cambridge University Press.

Cole, J. and Montero, B. (2007). Affective proprioception. *Janus Head* 9: 299-317.

David, N. et al. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition* 17: 523-534.

Davidson, D. (1967). Causal relations. *Journal of Philosophy* 64: 691-703.

Davidson, D. (1980). *Essays on Actions and Events*. New York: Oxford University Press.

Davidson, D. (1991). Mental events. In Rosenthal, D. (Ed.), *The Nature of Mind*. New York: Oxford University Press.

Dennett, D. (1991). *Consciousness Explained*. New York and Boston: Little Brown.

Descartes, R. (1991). To Princess Elizabeth, 28 June 1643. In Cottingham, R. et al. (Eds.), *The Philosophical Writings of Descartes*, vol. 3. Cambridge: Cambridge University Press.

Dijksterhuis, A., Bos, M.W., Nordgren, L.F., and van Baaren, R.B. (2006). On making the right choice: the deliberation-without-attention effect. *Science* 311: 1005-1007.

Dretske, F. (1988). *Explaining Behaviour*. Cambridge: MIT Press.

Dretske, F. (1997). What good is consciousness? *Canadian Journal of Philosophy* 27: 1-15.

Dretske, F. (2003). Experience as representation. *Philosophical Issues* 13: 67-82.

Ducasse, C.J. (1993). On the nature and the observability of the causal relation. In Sosa, E. and Tooley, M. (Eds.), *Causation*. Oxford: Oxford University Press.

Farrer, C. (2003). Modulating the experience of agency: a PET study. *Neuroimage* 18: 324-333.

Flanagan, O.J. and Polger, T.W. (1995). Zombies and the function of consciousness. *Journal of Consciousness Studies* 2: 313-321.

Flanagan, O.J. (1997). Conscious inessentialism and the epiphenomenalist suspicion. In Block, N. et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.

Fodor, J. (1989). Making mind matter more. *Philosophical Topics* 67: 59-79.

Frankfurt, H. (1997). Freedom of the will and the concept of a person. In Pereboom, D. (Ed.), *Free Will*. Indianapolis: Hackett.

Fried, I. et al. (1991). Functional organization of human supplementary motor cortex studied by electrical stimulation. *Journal of Neuroscience* 11: 3656-3666.

Frith, C.D. (1992). *The Cognitive Neuroscience of Schizophrenia*. Hove, UK: Lawrence Erlbaum.

Gallagher, S. (2004). Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology* 37: 8-19.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Grice, P. (2001). *Aspects of Reason*. Oxford: Clarendon.

Haggard, P. and Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research* 126: 128-133.

Haggard, P. and Magno, E. (1999). Localising awareness of action with transcranial magnetic stimulation. *Experimental Brain Research* 127: 102-107.

Haggard, P. (2001). The psychology of action. *British Journal of Psychology* 92: 113-128.

Haggard, P. (2006). Conscious intention and the sense of agency. In Sebanz, N. and Prinz, W. (Eds.), *Disorders of Volition*. Cambridge: MIT Press.

Haggard, P. and Brass, M. (2007). To do or not to do: the neural signature of self-control. *Journal of Neuroscience* 27: 9141-9145.

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience* 9: 934-946.

Harman, G. (1997). The intrinsic quality of experience. In Block, N. et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Heil, J. and Mele, A. (1991). Mental causes. *American Philosophical Quarterly* 28: 61-71.

Henderson, M. (2006). The man who can open his e-mails by the power of thought. *The Times*, July 13.

Hobbes, T. (1994). *Leviathan*, Curley, E. (Ed.). Indianapolis: Hackett.

Horgan, T. (1989). Mental quausation. *Philosophical Perspectives* 3: 47-76.

Horgan, T. et al. (2003). The phenomenology of first-person agency. In Walter, S. and Heckmann, H-D. (Eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter: Imprint Academic.

Hume, D. (1988). *An Enquiry Concerning Human Understanding*. Amherst, NY: Prometheus Books.

Huxley, T. (1874). On the hypothesis that animals are automata, and its history. *Nature* 10: 362-366.

Jackson, F. (1996). Mental causation. *Mind* 105: 377-413.

Jeannerod, M. (2003). Consciousness of action and self-consciousness: a cognitive neuroscience approach. In Roessler, J. and Eilan, N. (Eds.), *Agency and Self-Awareness*. Oxford: Oxford University Press.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9: 718-727.

Kim, J. (1966). On the psycho-physical identity theory. *American Philosophical Quarterly* 3: 277-285.

Kim, J. (1973). Causation, nomic subsumption, and the concept of event. *Journal of Philosophy* 70: 217-236.

Kim, J. (1991). Epiphenomenal causation. In Rosenthal, D. (Ed.), *The Nature of Mind*. New York: Oxford University Press.

Kim, J. (1998a). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge: MIT Press.

Kim, J. (1998b). *Philosophy of Mind*. Boulder, CO: Westview Press.

Kriegel, U. (2004). The functional role of consciousness: a phenomenological approach. *Phenomenology and the Cognitive Sciences* 3: 171-193.

Kriegel, U. and Williford, K., eds. (2006). *Self-Representational Approaches to Consciousness*. Cambridge: MIT Press.

Kriegel, U. (2007). Intentional inexistence and phenomenal intentionality. *Philosophical Perspectives* 21: 307-340.

Kriegel, U. (2009). Self-representation and phenomenology. *Philosophical Studies* 143: 357-381.

Lewis, D. (1993). Causation. In Sosa, E. and Tooley, M. (Eds.), *Causation*. Oxford: Oxford University Press.

Libet, B. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* 106: 623-642.

Libet, B. (2003). Can conscious experience affect brain activity? *Journal of Consciousness Studies* 10: 24-28.

Lurz, R. (2004). Either FOR or HOR: a false dichotomy. In Gennaro, R.J. (Ed.), *Higher-Order Theories of Consciousness*. Amsterdam and Philadelphia: John Benjamins Publishers.

Lycan, W. (1987). *Consciousness*. Cambridge: Bradford Books/MIT Press.

Lycan, W. (1997). Consciousness as internal monitoring. In Block, N. et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Lycan, W. (2001). The case for phenomenal externalism. In Tomberlin, J.E. (Ed.), *Philosophical Perspectives*, vol. 15: Metaphysics. Atascadero, CA: Ridgeview.

Lycan, W. (2004). The superiority of HOP to HOT. In Gennaro, R.J. (Ed.), *Higher-Order Theories of Consciousness*. Amsterdam and Philadelphia: John Benjamins Publishers.

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*: 245-264.

Mangan, B. (1999). The fringe: a case study in explanatory phenomenology. *Journal of Consciousness Studies* 6: 249-252.

Marcel, A.J. (1986). Consciousness and processing: choosing and testing a null hypothesis. In *Behavioral and Brain Sciences* 9: 40-41.

Marcel, A.J. (2003). The sense of agency. In Roessler, J. and Eilan, N. (Eds.), *Agency and Self-Awareness*. Oxford: Oxford University Press.

Marslen-Wilson, W.D. and Tyler, L.K. (1980). The temporal language of spoken language understanding. *Cognition* 8: 1-71.

Merikle, P.M. and Daneman, M. (1998). Psychological investigations of unconscious perception. *Journal of Consciousness Studies* 5: 5-18.

Metzinger, T. (2006). Conscious volition and mental representation: toward a more fine-grained analysis. In Sebanz, N. and Prinz, W. (Eds.), *Disorders of Volition*. Cambridge: MIT Press.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. Cambridge: MIT Press.

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

Nelkin, N. (1995). The dissociation of phenomenal states from apperception. In Metzinger, T. (Ed.), *Conscious Experience*. Paderborn, Germany: Schöningh.

Norman, D.A. and Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In Davidson, R.J. et al. (Eds.), *Consciousness and Self Regulation: Advances in Research*, vol. 4. New York: Erlbaum.

Obhi, S. and Haggard, P. (2004). Free will and free won't. *American Scientist* 92: 358-365.

Poppel, E., Held, R., and Frost, D. (1973). Residual visual functions after brain wounds involving the central visual pathways in man. *Nature* 243: 295-296.

Posner, M. and Snyder, C. (1975). Facilitation and inhibition in the processing of signals. In Rabbitt, P.M.A. and Dornick, S. (Eds.), *Attention and Performance*. Waltham, MA: Academic Press.

Putnam, H. (1975). The meaning of 'meaning.' In *Mind, Language, and Reality: Philosophical Papers*, vol. 2. Cambridge: Cambridge University Press.

Prinz, W. (2003). Experimental approaches to action. In Roessler, J. and Eilan, N. (Eds.), *Agency and Self-Awareness*. Oxford: Oxford University Press.

Prinz, J. (2005). A neurofunctional theory of consciousness. In *Cognition and the Brain*. New York: Cambridge University Press.

Prinz, J. (2007). All consciousness is perceptual. In McLaughlin, B. and Cohen, J. (Eds.), *Contemporary Debates in Philosophy of Mind*. Malden, MA: Blackwell.

Rey, G. (1998). A narrow representationalist account of qualitative experience. *Philosophical Perspectives* 12: 435-458.

Robinson, W. (2012). Epiphenomenalism. In Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy*. URL: http://plato.stanford.edu/archives/sum2012/entries/epiphenomenalism.

Rolls, E. (2004). A higher order syntactic thought (HOST) theory of consciousness. In Gennaro, R.J. (Ed.), *Higher-Order Theories of Consciousness*. Amsterdam and Philadelphia: John Benjamins Publishers.

Rosenthal, D.M. (1997). A theory of consciousness. In Block, N. et al. (Eds.), *The Nature of Consciousness*. Cambridge: MIT Press.

Rosenthal, D.M. (2000). Metacognition and higher-order thoughts. *Consciousness and Cognition* 9: 231-242.

Rosenthal, D.M. (2002). The timing of conscious states. *Consciousness and Cognition* 11: 215-220.

Rosenthal, D.M. (2004). Varieties of higher-order theory. In Gennaro, R.J. (Ed.), *Higher-Order Theories of Consciousness*. Amsterdam and Philadelphia: John Benjamins.

Rosenthal, D.M. (2005). *Consciousness and Mind*. Oxford: Clarendon Press.

Rosenthal, D.M. (2008). Consciousness and its function. *Neuropsychologia* 46: 829-840.

Rosenthal, D.M. (2011). Exaggerated reports: reply to Block. *Analysis* 71: 431-437.

Searle, J. (1980). The intentionality of intention and action. *Cognitive Science* 4: 47-70.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Segal, G. and Sober, E. (1991). The causal efficacy of content. *Philosophical Studies* 63: 1-30.

Shoemaker, S. (1999). Causality and properties. In Kim, J. and Sosa, E. (Eds.), *Metaphysics: An Anthology*. Malden, MA: Blackwell.

Sider, T. (2003). What's so bad about overdetermination? *Philosophy and Phenomenological Research* 67: 719-726.

Strawson, G. (2003). Mental ballistics *or* the involuntariness of spontaneity. *Proceedings of the Aristotelian Society* 103: 227-256.

Synofzik, M. et al. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and Cognition* 17: 219-239.

Tiffany, E.C. (2001). The rational character of belief and the argument for mental anomalism. *Philosophical Studies* 103: 285-314.

Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences* 14: 651-726.

Von Wright, G.H. (1993). On the logic and epistemology of the causal relation. In Sosa, E. and Tooley, M. (Eds.), *Causation*. Oxford: Oxford University Press.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford: Clarendon.

Weiskrantz, L. (2007). *Consciousness Lost and Found*. Oxford: Oxford University Press.

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge: MIT Press.

Wegner, D. (2003). The mind's best trick: how we experience conscious will. *TRENDS in Cognitive Science* 2: 65-69.

Wilberg, J. (2010). Consciousness and false HOTs. *Philosophical Psychology* 23: 617-638.

Yablo, S. (1992). Mental Causation. *Philosophical Review* 101: 245-280.

Zigmond, M. et al., eds. (1999). *Fundamental Neuroscience*. San Diego: Academic Press.