For a festschrift for J. Kim, co-edited by T. Horgan, M. Sabates, D. Sosa, and me, MIT Press, in press

CAUSAL COMPATIBILISM ABOUT AGENTIVE PHENOMENOLOGY

Terry Horgan
University of Arizona.

I will here describe a philosophical problem that surely should fall under the rubric "philosophical problems about mental causation," but has not been much noticed or much discussed in recent philosophy of mind. Briefly, there is a problem about whether or not human beings are really agents of the kind they experience themselves to be—whether their *agentive experience* is veridical or instead is illusory. Two considerations together generate the problem. First, there is prima facie reason to think that agentive experience could not be veridical unless human beings exert a form of control over their behavior, quite different in kind from mental *state*-causation, that is posited by some philosophers who espouse metaphysical libertarianism about freedom and determinism—what is sometimes called "agent causation." Second, there is also good reason to think that there is really no such phenomenon—that all human behavior is actually a part of the state-causal nexus.

I will set forth and motivate this problem in a way that highlights its distinctness from more familiar issues about mental causation. I will argue (i) that the satisfaction conditions for human agentive experience do *not* require humans to exert libertarian "agent causation" of behavior, and (ii) that these satisfaction conditions are very likely actually *satisfied* in normal human behavior. I call this view "causal compatibilism," because it affirms that the veridicality of agentive experience would be compatible with various plausible metaphysical theses that might initially seem to threaten such compatibility—for instance, the thesis that all human behaviors are physical events that have physical causes.

Familiar discussions of mental causation in recent philosophy of mind focus on mental *state*-causation—on whether mental events, qua mental, have causal efficacy vis-à-vis other mental states and vis-à-vis behavior.[1] It might be thought that these familiar discussions already address the problem I am referring to. After all, such discussions are often initiated by asking whether or not humans are really agents of their own actions, as they take themselves to be—where it is presupposed that being a genuine agent is a matter of the state-causal efficacy of one's mental events, qua mental, in generating one's behavior. But there are two reasons why these discussions do not directly speak to the philosophical problem I am posing here. First, they typically do not focus on agentive *phenomenology*—the what-it's-like of doing (or of *seeming* to do, at any rate). On the contrary, they typically proceed as if there were no such thing as distinctively agentive phenomenology at all. Second, as I will argue below, the metaphysical hypothesis that human behaviors are state-caused by mental events, qua mental, is itself among the prima facie *threats* to the veridicality of agentive experience! This fact is palpably ironic, given how often it is claimed that the way to validiate our belief in our own agency is to vindicate the state-causal efficacy of mental events vis-à-vis behavior.

Much of the recent literature on mental causation has been predicated on acceptance of *strong externalism* in philosophy of mind—roughly and generically, the view that all mental intentionality depends constitutively on certain external connections (e.g., causal, and/or covariational, and/or historical, and/or evolutionary connections) between a mental subject and that subject's wider environment. I myself maintain that strong externalism is deeply mistaken, even though philosophers like Putnam, Kripke, and Burge have made important and correct observations about how *some* aspects of mental intentionality constitutively

involve external linkages to the mental subject's wider environment. The specific problem of mental causation I will pose and address is especially apt to be overlooked by advocates of strong externalism. For, the problem emerges most clearly and vividly, I think, once one sees that—and why—strong externalism is so deeply misguided.

In the first section of this paper I summarize two related lines of thought in my prior (collaborative) philosophical work that form the background out of which has arisen, in my own work, the problem I will be addressing. First I describe work with John Tienson and George Graham on what we call "phenomenal intentionality" (Horgan and Tienson 2002, Horgan, Tienson, and Graham 2004). Our approach firmly repudiates strong externalism about mental intentionality, while also incorporating the important lessons learned from Putnam, Kripke, and Burge. Then I describe work with Tienson and Graham directed more specificially at *agentive* phenomenology—the "what it's like" of acting (Horgan, Tienson, and Graham 2003, Horgan and Tienson 2005). We maintain that agentive phenomenology is richly intentional, and thus that philosophical questions arise about its satisfaction conditions and about whether or not it is veridical.

In the second section I describe three metaphysical/scientific theses each of which is fairly credible, and I explain why all three of them might initially seem to threaten the veridicality of agentive phenomenology. Each thesis generates a specific version of the generic problem of compatibility I am highlighting in this paper—compatibility between the content of agentive experience on one hand, and one or another plausible metaphysical/scientific thesis on the other hand.

In the third section I set forth and defend a version of compatibilism about agentive phenomenology that purports to solve these problems. The defense draws in part upon earlier work of mine (some collaborative with George Graham and with David Henderson) articulating and defending a contextualist form of compatibilism concerning freedom and determinism (Horgan 1979, Horgan and Graham 1994, Henderson and Horgan 2000).

Finally, in the fourth section I connect the present discussion to the more familiar issue of mental *state*-causation. One plausible version of my generic compatibilist position about agentive phenomenology incorporates, as a component element, the claim that the satisfaction conditions for agentive phenomenology require one's behavior to be state-caused by certain mental events, qua mental. That requirement is typically satisfied, I maintain. Here I invoke a position I have set forth and defended elsewhere—a form of contextualist causal compatibilism about mental state-causation (Horgan 1989, 1991, 1993b, 1998, 2001a, 2001b).

1. Groundwork

My discussion in the later parts of the paper will presuppose some views of mine that I have developed and defended elsewhere, in collaboration with John Tienson and with George Graham. In this first section I will briefly summarize these views. Although it might turn out that the main problem I want to address can be motivated and addressed without presupposing everything I will say in this section, my own thinking about the problem is so heavily affected by the views I will now describe that I should begin by setting them out.

1.1. Phenomenal Intentionality and Externalistic Intentionality

Lately Graham and Tienson and I have been articulating a general approach to mental intentionality that is very much at odds with the various versions of strong externalism that have recently been fashionable in philosophy of mind. Although our approach repudiates strong externalism, we claim that it nonetheless does justice to the genuine insights about

matters of intentionality and reference in the work of philosophers like Putnam, Kripke, and Burge. Central in our approach is the role of *phenomenology* or *phenomenal consciousness*, by which we mean those aspects of one's mental life such that there is "something it is like" to undergo them. Briefly, the position goes as follows.

Phenomenology is *narrow*: it is not constitutively dependent upon anything "outside the head" (or outside the brain) of the experiencing subject. Indeed, it is not constitutively dependent upon anything outside of phenomenal consciousness itself; in this sense, it is *intrinsic*. Your phenomenology, being narrow and intrinsic, supervenes at least nomically upon physical events and processes within your brain.

Phenomenology is also richly and pervasively *intentional*: there is a kind of intentionality that is entirely constituted phenomenologically (we call it *phenomenal* intentionality), and it pervades our mental lives. Among the different aspects of phenomenal intentionality are the following. First, there is the phenomenology of perceptual experience: the enormously rich and complex what-it's-like of being perceptually presented with a world of apparent objects, apparently instantiating a rich range of properties and relations—including one's own apparent body, apparently interacting with other apparent objects which apparently occupy various apparent spatial relations as apparently perceived from one's own apparent-body centered perceptual point of view. Second, there is the *phenomenology of agency*: the what-it's-like of apparently *voluntarily controlling* one's apparent body as it apparently moves around in, and apparently interacts with, apparent objects in its apparent environment. (This is a central focus of the present paper, of course.) Third, there is *conative and cognitive* phenomenology: the what-it's-like of consciously (as opposed to unconsciously) undergoing various occurrent propositional attitudes, including conative attitudes like occurrent wishes and cognitive attitudes like occurrent thoughts. There are phenomenologically discernible aspects of conative and cognitive phenomenology, notably (i) the phenomenology of *attitude type* and (ii) the phenomenology of *content*. The former is illustrated by the phenomenological difference between, for instance, *occurrently hoping* that Hillary Clinton will be elected U.S. President and *occurrently wondering* whether she will be elected —where the attitude-content remains the same while the attitude-type varies. The phenomenology of content is illustrated by the phenomenological difference between occurrently thinking that Hillary Clinton will be elected U.S. President and occurrently thinking that she will *not* be elected—where the attitude-type remains the same while the content varies.)

Since phenomenal intentionality is entirely constituted phenomenologically, and since phenomenology is narrow, phenomenal intentionality is narrow too. Hence, there is *exact match* of phenomenal intentionality between yourself and your brain-in vat (BIV) physical duplicate. This exactly matching, narrow, intentional content involves exactly matching, phenomenally constituted, *narrow truth conditions*. But whereas the narrow truth conditions of your own beliefs are largely satisfied, those of your BIV physical duplicate's matching beliefs largely fail to be satisfied; thus, the BIV's belief system is systematically nonveridical.

On the other hand, exact match in narrow content between your own intentional mental states and the corresponding states in your BIV physical duplicate does not require or involve exact match in *referents* (if any) of all the various matching, putatively referring, thought-constituents. For instance, certain of your own occurrent thoughts that you would express linguistically using certain proper names—say, the thought that Bush is not a genius—involve singular thought-constituents whose referents (if any) are determined partly in virtue of certain external relations that obtain between you and those referents. Thus, your

occurrent thought that *Bush is not a genius* involves a singular thought-constituent that purports to refer to a particular specific person (viz., Bush); its *actually* referring, and its referring to the specific individual to whom it does refer, depends upon there being certain suitable external relations linking you to a unique eligible referent (viz., Bush). A Twin-Earthly physical duplicate of yourself, in a Twin-Earthly duplicate local environment, would refer to a *different* individual (viz., Twin-Bush) via the corresponding singular thought-constituent of the corresponding occurrent thought. And in the case of your BIV physical duplicate, the matching singular thought-constituent *fails to refer at all*, because the BIV does not bear suitable externalistic relations to any suitably reference-eligible individual in its own actual environment. (Parallel remarks apply to thought-constituents that purport to refer to natural kinds, such as the thought-constituent that you yourself would express linguistically with the word 'water'.)

For mental states involving thought-constituents for which reference depends upon externalistic factors, there are two kinds of intentionality, each involving its own truth conditions. First is the kind of intentionality already mentioned above: *phenomenal* intentionality, with truth conditions that are phenomenally constituted and narrow. Second is *externalistic* intentionality, with wide truth conditions that incorporate the actual referents (if any) of the relevant thought-constituents. Your own thought that Bush is not a genius, and the corresponding thoughts of your BIV physical duplicate and your Twin Earth physical duplicate, have matching phenomenal intentionality, with matching truth conditions. (These truth conditions are satisfied in your case and in the case of your Twin Earth duplicate, but not in the case of your BIV duplicate.) On the other hand, your own thought that Bush is not a genius and your Twin Earth duplicate's corresponding thought do not have matching *externalistic* intentionality, because the externalistic truth conditions of these respective thoughts do not match: the truth value of your own thought depends upon the intelligence level of *Bush*, whereas the truth value your Twin Earth duplicate's corresponding thought depends upon the intelligence level of an entirely different individual, viz., Twin-Bush. (Each thought's wide truth conditions are indeed satisfied.) As for your BIV duplicate's thought, it lacks externalistic intentionality and wide truth conditions, because its singular thought-constituent purporting to refer to a person called 'Bush' does not actually refer at all.

Our account rests heavily and essentially upon two key contentions. First, mental reference to many properties and relations—including various spatiotemporal-location properties, shape-properties, size-properties, artifact-properties, and personhood-involving properties—is wholly constituted by phenomenology alone. Even systematically *nonveridical* phenomenology, as in the case of the BIV, provides *reference-constituting experiential acquaintance* with such properties and relations. It makes no difference to such experiential acquaintance with such properties—and hence it makes no difference to mental *reference* to such properties—whether or not the properties with which one becomes experientially acquainted are ever actually instantiated in one's ambient environment.

Second, in the case of thought constituents whose reference (if any) depends constitutively upon certain externalistic elements, the mechanisms of reference-fixation crucially involve phenomenally constituted *grounding presuppositions* (as we call them). Thus, phenomenal intentionality is more basic than externalistic intentionality, since the latter depends in part upon the former (as well as depending in part upon externalistic factors). Suppose, for example, that you have an occurrent thought that you could express linguistically by saying "That picture is hanging crooked," where the singular thought-constituent expressible linguistically by 'that picture' purports to refer to a picture on the wall directly in front of you. This thought-content involves phenomenally constituted grounding presuppositions that

must be satisfied in order for the singular thought-constituent to refer: roughly, there must be an object at a certain location relative to yourself (a location that you could designate linguistically by a specific use of the place-indexical 'there'), this object must be a picture, there must not be any other picture at that location that is an equally eligible potential referent of 'that picture', and this object must be causing your current picture-experience. If these grounding presuppositions are satisfied by some specific concrete particular in your ambient environment—some particular object that is a picture and is uniquely suitably located—then your singular thought-constituent thereby refers to that very object. *Which* object your thought-constituent refers to, if any, thus depends jointly upon two factors, one phenomenally constituted and one externalistic: on one hand, the phenomenally constituted grounding presuppositions, and on the other hand, the unique actual object in your ambient environment that *satisfies* those presuppositions.

1.2 The Agentive Phenomenology of First-Person Agency

Graham and Tienson and I have also have been urging specific attention to the phenomenology of first-person agency—the "something it is like" to experience oneself as behaving in a way that constitutes action.[2] We maintain that agentive phenomenology is richly intentional, presenting in experience a self that is an apparently embodied, apparently voluntarily behaving, agent. Because agentive experience is intentional, it has *satisfaction conditions*—which raises two philosophically important questions. First, what are those satisfaction conditions? I.e., what is required of the world, including oneself and one's own body, in order for one to be an agent of the kind one experiences oneself as being? Second, are those conditions actually satisfied? I.e., are humans *in fact* agents of the kind they experience themselves as being? Such questions have received very little attention in recent philosophy of mind, largely because the phenomenology of agency itself has received very little attention. But Graham and Tienson and I have been arguing that this needs to change. In this section I will briefly summarize some of what we have had to say descriptively about the phenomenology of doing—about what this kind of "what it's like" is like. Issues about satisfaction conditions will be central in the remainder of the paper.

We employ the term 'behavior' in a broad sense, one that is neutral about whether or not any particular instance of behavior counts a genuine *action*. Paradigmatic behaviors are certain kinds of bodily motions. (Although there can be other forms of behavior, such as remaining still or remaining silent, we largely set them aside for simplicity.) The point of using 'behavior' in this broad sense is to remain neutral about the question whether the bodily motions called behavior really meet the satisfaction conditions imposed upon them by the phenomenology of doing.

What is behaving like phenomenologically, in cases where you experience your own behavior as action? Suppose that you deliberately perform an action—say, holding up your right hand and closing your fingers into a fist. As you focus on the phenomenology of this item of behavior, what is your experience like? To begin with, there is of course the purely behavioral aspect of the phenomenology—the what-it's-like of being visually and kinesthetically presented with one's own right hand rising and its fingers moving into clenched position. But there is more to it than that, of course, because you are experiencing this bodily motion *as your own action*.

In order to help bring into focus this specifically actional phenomenological dimension of the experience, it will be helpful to approach it a negative/contrastive way, via some observations about what the experience is *not* like. For example, it is certainly not like this: first experiencing an occurrent wish for your right hand to rise and your fingers to move into

clenched position, and then passively experiencing your hand and fingers moving in just that way. Such phenomenal character might be called *the phenomenology of fortuitously appropriate bodily motion*. It would be very strange indeed, and very alien.

Nor is the actional phenomenological character of the experience like this: first experiencing an occurrent wish for your right hand to rise and your fingers to move into clenched position, and then passively experiencing a causal process consisting of this wish's causing your hand to rise and your fingers to move into clenched position. Such phenomenal character might be called *the passive phenomenology of psychological state-causation of bodily motion*. People often do passively experience causal processes *as* causal processes, of course: the collision of a moving billiard ball with a motionless billiard ball is experienced as causing the latter ball's subsequent motion; the impact of the leading edge of an avalanche with a tree it its path is experienced as causing the tree to become uprooted; and so on. But it seems patently clear that one does not normally experience one's own actions in that way—as passively noticed, or passively introspected, causal processes consisting in the causal generation of bodily motion by occurrent mental states. That too would be a strange and alienating sort of experience.[3]

How, then, should one characterize the actional phenomenal dimension of the act of raising one's hand and clenching one's fingers, given that it is not the phenomenology of fortuitously appropriate bodily motion and it also is not the passive phenomenology of psychological event-causation of bodily motion? Well, it is the what-it's-like of *self as source* of the motion. You experience your arm, hand, and fingers as being moved *by you yourself*—rather than experiencing their motion either as fortuitously moving just as you want them to move, or passively experiencing them as being caused by your own mental states. You experience the bodily motion as caused by *yourself*.[4]

The phenomenal character of actions also typically includes aspects of *purposiveness*: both a generic what-it's-like of acting *on purpose*, and often also a more specific what-it's-like of acting *for a specific purpose*. The phenomenology of purposiveness can work in a variety of ways.[5] Sometimes, for instance (but not always), the action is preceded by conscious deliberation. In one variant of deliberative action, the process involves settling into reflective equilibrium prior to acting: the overall phenomenology includes, first, the what-it's-like of explicitly entertaining and weighing various considerations favoring various options for action, then the what-it's-like of settling upon a chosen action because of certain reasons favoring it, and then the what-it's-like of performing the action for those very reasons. (Examples range from the weighty, such as deciding which car to buy or which job offer to accept, to the mundane, such as deciding what to order for lunch in a restaurant.) In another variant, the action is preceded by the occurrence in experience of an explicit psychological syllogism: the overall phenomenology includes, first, the what-it's-like of mentally going through a particular piece of practical reasoning, and then the what-it's-like of performing an action because doing so is the upshot of that reasoning. (A familiar example of such an action is a deliberative version of the philosopher's workhorse of belief/desire explanation: at a party you consciously experience a desire for a beer and a perceptually generated occurrent thought about where the beer is located; you consciously form an intention to walk to that location and grab a beer; and then you act, with the explicit purpose in mind of getting yourself a beer.)

Actions are very often performed without prior deliberation, however. Here the tinge of purposiveness, within the phenomenology of doing, is typically more subtle. For example, as you approach your office you pull your keys out of your pocket or purse; then you grasp the office key; then you insert it into the lock; then you twist it in the lock; and then you push the

door open. All of this is routine and automatic: no deliberation is involved. Nonetheless, the what-it's-like of doing these things still certainly includes an on-purpose aspect, and indeed an aspect of doing them for specific purposes both fine-grained and coarse-grained: getting hold of your keys, getting hold of your office key in particular, activating the door lock, getting into your office, etc. In some cases of non-deliberative action, it appears, certain specific purposes for which one acts are explicitly conscious but not salient. In other cases, it seems, certain specific purposes are not explicitly conscious at all, but nonetheless are accessible to consciousness. In still other cases—for instance, specific actions performed during fast-paced sports such as soccer and basketball—some specific purposes for which the agent acts in one specific way rather than another probably are neither explicitly conscious nor even consciously accessible after the fact, because of the way these specific purposes are linked to very short-lived, and very intricately holistic, aspects of the player's rapidly changing perceptual phenomenology. Nonetheless, even here the phenomenology still normally includes the what-it's-like of acting in a specific way *for a specific purpose*, whether or not one finds oneself in a position after the fact to tell what that purpose was. Purposiveness is phenomenologically present in all these types of nondeliberative action, with specific purposes coloring conscious experience even when they are not explicitly conscious themselves.[6]

The phenomenology of doing typically includes another aspect, distinguishable from the aspect of purpose: viz., *voluntariness*. Normally when you do something, you experience yourself as *freely* performing the action, in the sense that it is *up to you* whether or not to perform it. You experience yourself not only as generating the action, and not only as generating it purposively, but also as generating it in such a manner that you *could have done otherwise*. This palpable phenomenology of freedom has not gone unrecognized in the philosophical literature on freedom and determinism, although often in that literature it does not receive as much attention as it deserves. (Sometimes the most explicit attention is given to effort of will, although it takes only a moment's introspection to realize that the phenomenology of voluntarily exerting one's will is really only one, quite special, case of the much more pervasive phenomenology of voluntariness.[7])

In philosophy of mind, on the other hand, there has been a widespread, and very unfortunate, tendency to ignore the phenomenology of doing altogether—and to theorize about human agency without acknowledging its phenomenology at all, let alone seeking to accommodate it. It is time to get beyond this major philosophical blindspot.

2. Some Threats to the Veridicality of Agentive Phenomenology

Let me now briefly set out three distinct, albeit interrelated, reasons to think that the phenomenology of doing might be nonveridical—i.e., that humans might not really be agents of the kind that their own first-person agentive phenomenology represents themselves to be. First is the scientifically highly credible hypothesis of *physical state-causal closure*: the claim that every physical event or state is completely causally determined—to the extent that it is causally determined at all—on the basis of physical laws plus prior physical states, and that the laws of physics are never violated. A fairly plausible line of thought runs as follows:

> If indeed all my behavior, and all my internal mental goings-on too, are state-caused by physical events and states in my brain and central nervous system, then it is never really the case that *I myself* ever do anything; I am never a genuine *agent* of my intentions, decisions, or actions, even though I undergo the phenomenology of such agency. So, if physical causal closure obtains, then the first-person phenomenology of agency is nonveridical, and radically so.

Second is the scientific hypothesis of *state-causal determinism*, which asserts that everything that happens in the world is uniquely determined to happen, given the prevailing laws of nature and given the total state prior of the world. (This claim is orthogonal to the hypothesis of physical state-causal closure. For, the laws of *physics* could fail to be deterministic, and yet state-causal determinism could obtain anyway because the laws of physics and the special sciences could be *collectively* deterministic—with special-science laws doing the determining wherever the laws of physics do not.) A fairly plausible line of thought is this:

> If indeed state-causal determinism obtains, then I never really do anything *freely*, and likewise I never really choose or decide anything freely. But my first-person agentive experience is thoroughly suffused with the dimension of freedom: when I experience behavior as my action, I experience it as something that I *could have refrained from doing*—and likewise for my experience of choosing and deciding. This experiential aspect of freedom is enormously salient, and is virtually always present in what I experience as full-fledged action, decision, and choice. (It is present even when I am being ordered around under extreme threat, even though it might be utterly irrational to refuse such an order. And it is present even when I am shackled or confined, even though the range of alternatives seemingly open to me might be highly constrained by my circumstances.) So, if state-causal determinism obtains, then the first-person phenomenology of agency is nonveridical, and radically so.

Third is the hypothesis of *the psychological state-causation of behavior*, which asserts that the paradigmatic behaviors that people experience as actions are state-caused by mental states of the experiencing subject.[8] Here a plausible line of thought is as follows:

> If the behaviors I experience as my actions are really state-caused by mental states of myself like occurrent thoughts and occurrent wishes (or by combinations of such states), then these behaviors are not really produced by *me myself*. Likewise for mental phenomena I experience as mental acts of mine—experiences as of making decisions, as of making plans, and as of forming intentions. But my agentive experience is thoroughly suffused with the aspect of self-as-source, rather than being the passive phenomenology of psychological state-causation. So if indeed these behaviors and mental phenomena are the product of psychological state-causation, then the phenomenology of first-person agency is nonveridical, and radically so.

3. Compatibilism about Agentive Phenomenology

I seek to formulate and defend a compatibilist position concerning the satisfaction conditions for agentive phenomenology. The kind of compatiblism I am after is one that would simultaneously disarm all three of the lines of argument described in section 2. I.e., I want a position according to which the satisfaction conditions for agentive phenomenology are such that these experiences can perfectly well be veridical even if physical closure obtains, even if causal determinism obtains, and even if the kinds of human behavior experienced as agentive results from mental state-causation. Such a position would be compatibilist with respect to *all* the salient aspects of agentive phenomenology, rather than saving veridicality for some aspects but not for others; in particular, it would include a vindication of the *freedom* aspect of agentive phenomenology.

I will not attempt to provide a short, sweet, exceptionless, tractably articulable, cognitively surveyable, formulation of satisfaction conditions—a formulation that would constitute a *philosophical analysis* of the intentional content of agentive experience. That would be a very tall order. Also, any such analysis that is clearly compatibilist would be philosophically tendentious *for that very reason*, because the analysis would go contrary to the intuitions

underlying the three lines of incompatibilist argumentation described in section 2. And I suspect that any such proposed philosophical analysis of the intentional content of agentive experience would be subject to counterexamples anyway (quite apart from being tendentious about the compatibilism issue). After all, *most* proposed analyses that have been put forth in philosophy have turned out to have counterexamples. Many of us have come to believe, on the basis of such negative inductive evidence, that most philosophically interesting concepts just cannot be given such a conceptual analysis (and don't need one). Why expect things to be different for the content of agentive experience?

In any event, regardless whether or not it would be *possible in principle* to articulate satisfaction conditions for agentive experience in some compact, exceptionless, and tractably surveyable formula, no such formula is required for our purposes here. Compatibilism is a hypothesis *about* those satisfaction conditions, and my objective is to make a case for the hypothesis itself.

Since the issue pertains to a certain facet of *experience*, my argument will proceed partly by appeal to the deliverances of introspection: describing what introspection does (and does not) seem to reveal to *me* about the issue, and urging you to notice from the first-person perspective that the same goes for *you*. But introspection will turn out to have only limited value on this matter—or so I will maintain. This will open the door for other evidential factors to enter the picture, dialectically. I will maintain that compatibilism, as a theoretical hypothesis about the intentional content of agentive phenomenology, provides overall a better explanation of the pertinent evidential data than does incompatibilism—and hence (via inference to the best explanation) that compatibilism is very likely true.

3.1 Two Key Distinctions

Before proceeding, let me set out two distinctions that will prove important in the discussion to follow. First is the distinction between aspects of phenomenology that are *manifest* to introspection, and aspects of it that are not manifest.[9] Many features of the what-it's-likeness of experience are virtually impossible not to notice when one attends introspectively to one's own phenomenology itself. These features are right there before the mind, vividly self-presented in the experience itself. Vivid sensory-presentational aspects of phenomenology are perhaps the most obvious examples: what it's like to visually experience a particular shade of red, or to olfactorily experience a particular smell (e.g., burning rubber), or to auditorily experience a particular sound (e.g., a loud, ringing-buzzing, fire alarm). But other features, though still there in the phenomenology, are not *manifestly* there—are not straightforwardly self-presenting to one's introspective awareness. For example, to most of us it is not manifest that the vast portion of one's visual field is very much out of focus. (One cannot, after all, introspectively attend to this fact by *visually focusing* on other portions of one's visual field other than those one was previously focusing on; for to do *that* is to bring those erstwhile out-of-focus portions *into* focus.) Likewise, to most male subjects in the famous psychology experiment in which one of two identical-seeming photos of a woman's face is chosen over the other in a forced-choice task, it is not introspectively manifest that the woman's pupils are more dilated in the chosen photo than in the other one—even though this difference is indeed phenomenally *present*, and even though it evidently exerts an unnoticed effect on subjects' preferences (supposedly via subliminal responses to pupil dilation as a sign of sexual arousal).

Non-manifestness evidently comes in degrees. For instance, the visual-presentational difference between the two initially identical-looking photos of the female face can *become* introspectively manifest, once the subject is directed to pay attention to the size of the

woman's pupils. Likewise, to some extent at least, one can train oneself to *attend* to the blurry non-focal portions of one's visual field otherwise than by trying to visually *focus* on them. But some aspects of phenomenology might be yet more robustly non-manifest—yet more robustly resistant to becoming directly apparent to introspection. This theme will be important below.

Second is a distinction between two kinds of intentional content, which I will call *presentational* content and *judgmental* content, respectively. (I might perhaps have used the expressions 'non-conceptual content' and 'conceptual content', but there seem to be almost as many different ways of using *that* terminology as there are philosophers who use it.) I will be brief and vague about how to understand this distinction—partly because I think a rough-and-ready construal will serve my present purposes, and partly because I think it is an open philosophical question how best to further elaborate the distinction anyway. Presentational intentional content is the kind that accrues to phenomenology directly—apart from whether or not one has the capacity to articulate this content linguistically and understand what one is thus articulating, and apart from whether or not one has the kind of sophisticated conceptual repertoire that would be required to understand such an articulation. Judgmental intentional content, by contrast, is the kind of content possessed by such linguistic articulations, and by the judgments they articulate. (Here I use 'judgment' broadly enough to encompass various non-endorsing propositional attitudes, such as *wondering whether*, *entertaining that*, and the like.) Dogs, cheetahs, and numerous other non-human animals presumably have agentive phenomenology with presentational intentional content, although it is plausible that they have little or no sophisticated conceptual capacities of the kind required to undergo states with full-fledged judgmental content.

I do not mean to suggest that this distinction is a sharp one. It wouldn't surprise me if the two kinds of content blur into one another, via a spectrum of intervening types of psychological state and/or a spectrum of increasing forms of conceptual sophistication in different kinds of creatures. Also, it may well be that the two kinds of content can interpenetrate to a substantial extent, at least in creatures as sophisticated as humans. It is plausible, for instance, that humans can have presentational contents the possession of which require (at least causally) a fairly rich repertoire of background concepts that can figure in judgmental states. One can have presentational experiences of, for instance, as-of computers, automobiles, airplanes, train stations—all of which presumably require a level of conceptual sophistication that far outstrips what dogs possess.

2. Introspecting Agentive Phenomenology: Lessons and Limits

Our inquiry concerns the content of agentive phenomenology. One important source of data for such inquiry, of course, is the data of introspection. One attends to one's own phenomenology, focusing specifically on certain aspects of the overall what-it's-likeness of one's experience, seeing out what is introspectively *manifest* in the phenomenology. Certain aspects of the phenomenology of doing are indeed manifest—aspects that were described in section 1.2: self-as-source, purposiveness, freedom to do otherwise. But the compatibility issues now under consideration are somewhat abstruse; they are *theoretical* questions concerning the intentional content of agentive experience. Is the content of such experience compatible with the thesis of physical causal closure? Is it compatible with the thesis of universal state-causal determinism? Is it compatible with the thesis that behaviors experienced as actions are mentally state-caused?

Take the first question, for specificity. What does one seem to find introspectively, when one attends to one's agentive phenomenology and reflects on its relation to the hypothesis of

physical causal closure? Well, negatively it seems one can say this much: agentive experience does not represent my doing, to myself, *as* implemented by physical processes governed by physical laws. The question, though, is whether agentive experience represents my doing, to myself, as something that is *not* implemented by physical processes governed by physical laws. Or, at any rate, does it represent my doing, to myself, in a way that *entails* that my doing is not thus implemented?

It seems to me, as I introspectively attend to my agentive phenomenology while reflectively posing this rather abstruse theoretical question about the intentional content of agentive experience, that *no answer is introspectively manifest* to me. (Indeed, it seems introspectively manifest to me that no answer to the question is introspectively manifest to me: the lack of manifestness concerning a first-order question about one's phenomenology can be introspectively manifest itself; this is a higher-order form of introspective manifestness.) It is not introspectively manifest that my agentive experience *is* compatible with the hypothesis of physical causal closure. Likewise, it is not introspectively manifest that it my experience is is *not* compatible with this hypothesis.

So we are acquiring some important introspective data, clearly relevant to our compatibility question concerning agentive phenomenology. But the data is also significantly limited. Since no answer to the question is introspectively manifest, addressing it adequately will require invoking some other forms of data too, and various wider theoretical considerations—and then reasoning abductively with an eye on the whole package of such factors.

Three hypotheses are effectively on the table, concerning our compatibility question: that the answer is yes, that the answer is no, and that the content of agentive experience is indeterminate on the matter. I seek to defend the yes answer. The advocate of this answer can be somewhat cheered by the fact that none of the answers is introspectively manifest, since this means that the yes answer is not in direct conflict with anything introspectively manifest. To be sure, this limited result hardly entails that the yes answer is right; for, the content of agentive phenomenology might yet be incompatible with the hypothesis of physical causal closure, *even though this conflict is not introspectively manifest*. But at least, in defending a yes answer, one is not claiming something that at odds with what seems introspectively manifest about the matter. That would be a *very* tough row to hoe (although in principle there might perhaps be ways to challenge the veridicality of certain instances of the experience of introspective manifestness).

What I have said about the specific compatibility question under discussion evidently extends, mutatis mutandis, to the other two compatibility questions too. No specific answer is introspectively manifest to the question whether agentive experience is compatible with the thesis of causal determinism. Likewise, no specific answer is introspectively manifest to the question whether agentive experience is compatible with the thesis that the behaviors we experience as actions are state-caused by mental states, qua mental. So for these questions too, further considerations need to be garnered in addition to the limited data of introspection, and then the task is to reason abductively in light of the full body of pertinent evidential factors.

3.3. Agentive Phenomenology vs. the Phenomenology of State-Causation of Bodily Motion: The Aspect of Freedom

The range of pertinent data that will need to be considered in addressing our compatibility questions includes various features of agentive phenomenology that *are* introspectively manifest. Of particular importance will be the aspect of freedom that is virtually always

present in the experience of agency: when I experience some behavior as produced *by me*, I experience it as something that I *could have refrained from doing*.

Even in situations where one acts under coercion or extreme duress—say, when a thief is holding a gun in one's face and demanding one's wallet, this *presentational* aspect of freedom is still present—despite the fact that it may well be appropriate to form a *judgment* expressable by saying "I was not acting freely." For, even in such extreme circumstances, one's agentive experience in turning over the wallet to the thief still has the sensory-presentational "I could do otherwise" aspect—notwithstanding one's palpable awareness that it would crazy to do otherwise, one's palpable fear of being shot, and other phenomenologically salient features that are pertinent to the judgment "I am not acting freely."

Compare cases where one experiences one's behavior as action (including cases of agentive phenomenology in circumstances of extreme duress) with cases where one instead experiences the behavior as non-actional bodily motion. Suppose, for example, that a large door swings closed behind me as I stand in the doorway, bumping my arm and thereby causing my arm to move forward. Here the "could have done otherwise" aspect is entirely missing, it seems: to experience the bodily motion *as state-caused* is to experience it as something that *was bound to happen, given the circumstances and the state(s) that did the causing*.

Now ask yourself this question: What would it be like to experience a motion of one's own body as state-caused by certain of one's *mental* states, qua mental? I think we do occasionally undergo such experiences—for instance, experiencing oneself *wincing* as a result of a sudden, unexpected pain, or experiencing one's eye *twitching* as a result of extreme frustration or irritation. (Perhaps you will remember the character Inspector Clouseau, played brilliantly by Peter Sellars in the movie *The Pink Panther* and in a series of sequel-films. Clouseau's boss on the police force undergoes spasmodic eye-twitching every time he becomes exasperated by Clouseau's astounding stupidity.) Now, the striking thing about such cases is that the bodily motions thus experienced—i.e., experienced as caused by one's own mental states, qua mental—are not experienced as *actions* at all.

Could one, though, simultaneously experience a bodily motion both as mentally state-caused and as one's action? When one tries to imagine such an experience, it seems introspectively manifest that it is very hard—indeed, virtually impossible—to do so. One principal obstacle to imagining—or having—such an experience is this. On one hand, the aspect of *freedom*—the 'can do otherwise' aspect—is ubiquitous in full-fledged agentive experience—even (as stressed already) in cases where one acts under extreme duress, such as surrendering one's wallet to the thief who is pointing a gun in one's face. On the other hand, to experience an event *as state-caused*—even as caused by one's own *mental* states, qua mental—is to experience it as something that was *bound to happen* given the circumstances and given the causing-states. It is virtually impossible to imagine experiencing one and the same bodily event both ways simultaneously—experiencing it as an action that one could have refrained from performing, while also experiencing it as a bodily motion that was mentally state-caused and thus was bound to occur given those mental causes.[10] As I will put it, these two potential forms of experience—agentive phenomenology and the phenomenology of state-causation of bodily motion—are *phenomenologically mutually exclusionary*—and manifestly so.

This is a datum, evidentially pertinent to the issue I am investigating. Prima facie, it makes trouble for a compatibilist position about the satisfaction conditions of agentive phenomenology: it lends evidential support to the hypothesis that these satisfaction

conditions require of genuine actions that they are not state-caused by mental events, qua mental. For, this incompatibilist hypothesis provides a natural-looking, prima facie plausible, explanation for the phenomenon of phenomenological mutual exclusion. An adequate compatibilism will need to provide an alternative explanation.

3.4. A Compatibilist Proposal

Let me now set forth the compatibilist position I would like to defend. (I will argue for it below.) It comprises the following theses. First, the presentational intentional content of agentive phenomenology has satisfaction conditions that are compatibilist in all three of the ways described above: being an agent of the kind one experiences oneself to be is compatible with physical causal closure, is compatible with causal determinism, and is compatible with the mental state-causation, qua mental, of the behaviors experienced as actions. Second, this compatibility is a non-manifest feature of agentive phenomenology. Third, despite the compatibility, a bodily event that is experienced as one's action cannot also be *experienced* as state-caused, either by non-mental states or by mental states. Fourth, an essential aspect of agentive phenomenology is the presentational aspect of *freedom*, which is phenomenologically present even when one experiences oneself as acting under cooercion or duress. Fifth, an essential aspect of experiences of state-causation, including experiences of one's own bodily motions as state-caused, is the presentational aspect of *inevitability*—i.e., the aspect of inevitability *given the circumstances and the causing states*. Sixth, the two theses lately mentioned jointly explain the phenomenological mutual exclusion described in the third thesis: this exclusion results from the freedom aspect of agentive phenomenology on one hand, and from the inevitability aspect of the phenomenology of state-causation on the other hand. One cannot experience an item of one's own behavior both as inevitable and as something that one could have refrained from doing.

Seventh, at the level of *judgmental* intentional content, the concept of freedom involves a feature that is probably not exhibited by the freedom aspect of *presentational* intentional content—viz., implicit contextual parameters that determine, in context-specific ways, contextually operative standards of satisfaction. For instance, in many contexts the standards operate in such a way that an action performed under extreme cooercion—e.g., with a gun in one's face—do not count as free. I.e., under the contextually operative standards, the *judgment* that such an action is not free is correct. (In other contexts, however, the concept of freedom is correctly used in such a way that its satisfaction conditions coincide with those for the freedom aspect of sensory-experiential intentional content—for instance, when one says "I could have refused to give the gunman my wallet, although that would have been a foolhardy thing to do; thus, I excercised freedom of choice in giving it to him.")

Eighth, the implicit contextual parameters governing the judgmental concept of freedom can take on a limit-case setting in certain contexts of judgment or conversation—i.e., a parameter-setting under which an item of behavior counts as free only if (i) it is not state-causally determined, and (ii) it comes about as a result of metaphysical-libertarian "agent causation" involving the self as a godlike unmoved mover.[11]

Ninth, at the level of judgmental intentional content, the concept of *agency* also becomes susceptible to implicit contextual parameters. For, in forming a judgment about some behavior's being an action, one construes the behavior as *minimally* free—i.e., as a behavior that possessed the 'could have done otherwise' feature, even if doing otherwise was not a reasonable option (say, because the agent had a gun in his face, and was acting as demanded by the gunman). Yet even this somewhat minimal usage of 'free' at the level of judgmental intentional content, tied mainly to 'could have done otherwise' rather than to matters of

cooercion and the like, is still governed by implicit contextual parameters—parameters that still can take on a limit-case setting under which 'could have done otherwise' is incompatible with the agent's behavior being state-causally determined. Thus, there is a limit-case usage of the notion of agency under which an item of behavior counts as *action* only if (i) it is not state-causally determined, and (ii) it comes about as a result of metaphysical-libertarian "agent causation" involving the self as a godlike unmoved mover.

Tenth, the satisfaction conditions for *presentational* agentive intentional content—i.e., for agentive *phenomenology*—coincide with certain non-limit-case, compatibilist, satisfaction conditions for *judgmental* agentive intentional content. The satisfaction conditions for agentive phenomenology do *not* coincide with the incompatibilist satisfaction conditions that accrue to judgmental agentive intentional content when the implicit parameters at work in the judgmental concepts of freedom and agency have extremal, limit-case, settings.

3.5. Evidence for Compabilism about Judgmental Agentive Content

Ideology is a term I like for broadly empirical inquiry into matters of content—both judgmental content and presentational content (cf. Horgan 1993a, Horgan and Graham 1994, Henderson and Horgan 2000). Ideology is a multi-discipline enterprise, whose philosophical dimension is continuous with relevant work in disciplines like cognitive science, linguistics, and sociolinguistics. Insofar as ideology relies primarily on armchair-accessible data, it can be effectively practiced by philosophers. Philosophical thought experiments, like Putnam's Twin Earth case, really *are* experiments: they generate empirical data for ideological theorizing, much as a linguist's intuitive grammaticality-judgments constitute data for syntactic theorizing.

The kinds of armchair-obtainable data that are pertinent to philosophical ideological inquiry appear to be fairly diverse (more so than in the linguistics case), with some kinds being more directly analogous to grammaticality judgments than others. The types of data that can figure in philosophical ideological reflection include the following:

1. Intuitive judgments about what is correct to say concerning various concrete scenarios, actual or hypothetical.
2. Facts about conflicting judgments or judgment-tendencies, concerning the correct use of certain concepts in various actual or hypthetical scenarios.
3. Facts about standardly employed warrant-criteria for the use of various concepts.
4. Facts about the key sociolinguistic purposes served by various terms and concepts.
5. General background knowledge, including untendentious scientific knowledge.

Facts of all these kinds can go into the hopper of wide reflective equilibrium whereby ideological claims are defended in philosophy. One makes a case for a certain ideological hypothesis—for instance, the contention that the meaning of natural-kind terms depends in part on the language-users' environment—by arguing that it does a better job, all things considered, of accommodating the relevant data than do any competing ideological hypotheses. Such reasoning is broadly empirical: inference to the best explanation, in which empirical data of all the kinds 1-5 are potentially relevant.

I propose to argue in this way in support of the package-deal compatibilist proposal I set forth in section 3.4. That proposal pertains both to judgmental agentive content and to presentational agentive content. The argument will proceed in three stages. First, in the present section, I will focus on judgmental agentive content, garnering data of types 1-5 and arguing that much of it directly supports causal compatibilism. (Here I will be drawing on

prior collaborative work on the freedom/determinism issue with George Graham and with David Henderson.) Second, in section 3.6, I will garner further data concerning presentational agentive content and concerning interconnections in human cognition between the two kinds of content, and I will argue that much of this data too directly supports causal compatibilism. Third, in section 3.7, I will offer a proposed account of some residual data—data involving incompatiblist judgment-tendencies that causal compatibilism needs to treat as mistaken.

I acknowledge at the outset that some considerations I will marshal lend more direct support to some aspects of my package-deal compatibilist approach than to others. Certain specific features of the approach, including the hypothesis of implicit contextual parameters at the level of judgmental agentive content, enter largely by way of helping provide a plausible explanation of residual data.

For simplicity of exposition, I will largely confine my specific attention to judgments about *freedom* (in the case of judgmental content), and to the freedom aspect of agentive phenomenology (in the case of presentational content). Also, I will largely confine my remarks about compatibility to the question of compatibility with *causal determinism*. But I think it will be fairly clear how to generalize these remarks to other aspects of agency (e.g., self-as-source, purposiveness) and to other compatibility issues (e.g., compatibility with the mental state-causation of bodily motion).

So, turning initially to judgmental agentive content, I begin by mentioning some pertinent data of each of the five types lately mentioned. Type 1: One has strong intuitive judgments, with respect to specific scenarios of various kinds, about the question of freedom; for many such scenarios one intuitively judges that the agent decides or acts freely, whereas in some scenarios one intuitively judges that this is not so. Cases of extreme cooercion, such as being threatened with a gun to the head, are of the latter kind; whereas more typical cases, where the agent decides or acts on the basis of uncoerced beliefs and desires, are of the former kind. Intuitive judgments ascribing freedom arise for many scenarios that are described in a way that says nothing at all, one way or the other, about whether or not decisions and actions in the scenario are governed by deterministic causal laws.

Type 2: It is notorious that one's judgment-tendencies are pulled in different ways, with respect to the problem of freedom and determinism. On the one hand are judgment-tendencies of the kind just mentioned, to classify certain paradigmatic cases as instances of freedom; the crucial features are lack of coercion or threat, lack of the influence of drugs or alcohol, and the like. These judgment-tendencies take no notice at all of the question whether the agent's decisions and actions are governed by deterministic laws of nature. On the other hand is a judgment-tendency that almost invariably arises when the problem of determinism and freedom is posed explicitly—viz., a tendency to think that no decision or action is *really* free if it is the result of deterministic causal processes within the agent.

Type 3: The evidential standards that are normally employed, in one's practice of classifying people's decisions and actions as freely taken (and also in the corresponding default assumptions one routinely and implicitly makes about oneself and others) are standards under which attributions of free decision and free action (and default assumptions to the same effect) are regarded as epistemically warranted *independently* of the availability or lack of availability of any evidence for or against causal determinism with respect to human decision and action.

Type 4: Attributions and default assumptions of human freedom figure centrally in practices of responsibility-attribution that are integral to a host of interpersonal and social-institutional relations and structures—including personal friendship, moral praise and blame, and legal sanctions against law-violators. These relations and structures are fundamental in human

society, whether or not human decisions and actions are the result of deterministic causal processes within the agent.

Type 5: There is overwhelmingly good scientific evidence that human decisions and actions result from cognitive processes that are physically implemented by electrochemical interactions between neurons in the brain and central nervous system. Also, there is no strong scientific evidence—at least none that educated laypersons know about—to suggest that the neural activity that directly subserves human deliberation, decision, and action is subject to any significant degree of causal indeterminacy. So it seems fairly likely that quantum-level indeterminacies, if such there be, "cancel each other out" much as they presumably do in televisions, computers, and other electronic gadgetry—rather than getting amplified in a way that significantly affects the behavior of entire neurons and entire neural assemblies.

Consider now the competing ideological hypotheses of compatibilism and incompatibilism, in light of the data just mentioned. Which hypothesis comports better with the data? Leave aside type-2 data for the moment, and consider the other kinds. Regarding data of type 1, a plausible empirical assumption is that one's intuitive judgments about the presence or absence of freedom, in various concrete scenarios one considers, normally are fairly straightforward products of one's own conceptual/semantic competence with the notion freedom, and hence are normally *true*. Compatibilism accommodates such judgments straightforwardly, by allowing freedom attributions to be true under paradigmatic attribution-conditions. Incompatibilism, however, holds the truth of freedom-attributions hostage to a very demanding additional condition, over and above the conditions that are clearly satisfied in ordinary paradigmatic cases of free decision and free action—viz, the requirement of causal indeterminism in the processes generating the decisions and actions. Ceteris paribus, one ideological hypothesis is better than another if the former accommodates the attributional practices of competent users of the relevant concept(s) better than the other. So in this respect, compatibilism does better than incompatibilism.

Data of type 3 reinforces these considerations. A plausible empirical assumption is that the epistemic standards one employs, when one makes confident intuitive judgments that various decisions and actions are free—and when one confidently adopts and maintains the default presumption that most ordinary decisions and actions by oneself and others are free—are *appropriate* epistemic standards, given the ideological workings of the concept freedom itself. For, the use of grossly inappropriate epistemic standards, in the confident intuitive deployment of a concept, typically reflects a deficiency in one's conceptual/semantic competence with that concept. Compatibilism accommodates the epistemic standards governing normal application of the concept freedom, since it treats these standards as appropriate. Incompatibilism, on the other hand, entails that such standards are much too lax; under suitable epistemic standards, freedom-attributions are only warranted when one has good evidence that human decisions and actions are not deterministically generated. Ceteris paribus, an ideological hypothesis that accommodates the epistemic standards normally accompanying a concept's deployment is better than an ideological hypothesis entailing that those epistemic standards are seriously deficient. So in this respect too, compatibilism does better than incompatibilism.

Data of type 4 also favors compatibilism. Human concepts emerge pragmatically, in ways that serve the purposes for which those concepts are employed. In general, therefore, concepts do not have satisfaction conditions built into them that are so demanding that they thwart the very purposes the concepts serve. Attributions and presuppositions of freedom play a key role in personal relations and in social institutions that are integral to civilized human society; this is so whether or not human decisions and actions are governed by

deterministic causal laws. But for all we now know, causal determinism might very well obtain vis-à-vis human decisions and actions—or anyway, virtual determinism, with any micro-level indeterminacies making only a negligible difference to more macro-level phenomena like the behavior of neurons and neural assemblies. So it would be purpose-thwarting for the concept freedom to possess satisfaction conditions requiring that human decisions and actions are not causally determined (since such conditions could very well fail to be satisfied). Compatiblism therefore accords better than incompatibilism with the central purposes which the concept freedom actually serves.

Data of type 5 reinforces these latest considerations. From the epistemic vantage point of the educated layperson concerning what is known about the neural-level etiology of human decisions and actions, it appears reasonably likely that people's decisions and actions normally result from electrochemical brain-activity that is roughly as deterministic as that of a pocket calculator or a digital computer. That is, it appears reasonably likely that there simply does not occur the dramatic kind of causal indeterminacy of thought and action that incompatibilism requires. All the more reason, then, to think that it would be purpose-thwarting for the concept freedom to presuppose such indeterminacy.

When these kinds of considerations are fed together into the hopper of wide reflective equilibrium, they reinforce one another epistemically in such a way that their combined epistemic weight is very powerful. The compatibilist hypothesis about the judgmental content of freedom attributions simply accords better with the relevant data than does the incompatibilist hypothesis. In particular, compatibilism explains why our ordinary intuitive attributions of free choice and free action, and the accompanying justification-standards we rely upon in making such attributions, operate in a way that is essentially orthogonal to the issue of causal determinism. They do so because the question whether a given decision or action was freely taken *is* orthogonal to the issue of causal determinism.

There remains, of course, the task of dealing with considerations of type 2. A fully adequate ideological treatment of the concept freedom owes a credible account of why people's intuitions can be easily and naturally pulled in the direction of thinking that freedom and determinism are incompatible, and why explicitly posing the compatibility question tends to do so. Incompatibilists can explain *this* tendency fairly straightforwardly, as the putative product of our conceptual/semantic competence with the notion freedom. Compatibilists, however, have the burden of explaining why the tendency arises so strongly and naturally even though it is allegedly mistaken. I will return to this matter in section 3.7.

6. Evidence for Compatibilism about Presentational Agentive Content

It is very plausible that agentive phenomenology is present in the mental lives not just of humans but also in those of certain non-human animals. Thus, in seeking out evidence for compatibilism about it, becomes natural to think in terms of data analogous to Type 4 above but involving *evolutionary-biological* point and purpose (rather than sociolinguistic point and purpose). Are there plausible hypotheses to be had concerning the likely evolutionary benefits, in terms of survival and flourishing and successful reproduction, of agentive phenomenology?[12]

Indeed so. Whether or not there exists, or even could exist, any such thing as what metaphysical libertarians about freedom call "agent causation," there are certain unproblematically real distinctions among phenomena that are well tracked by agentive phenomenology and event-causal phenomenology respectively—distinctions the accurate tracking of which is bound to provide survival/reproduction benefits. Some survival-important features of a creature's ambient environment will be ones that are susceptible to

causal influence by suitable bodily motions by the creature itself, motions that can be internally generated by the creature's inner motion-control mechanisms. (Consider a bear, for instance. In an appropriately fortunate ambient environment in which there is a bush nearby with edible berries on it, there are potential bodily motions available to the bear that would have the effect of transferring some of those berries from the bush itself to the bear's stomach. For such potential bodily motions, the anticipatory-freedom phenomenology of "I can" (vis-à-vis those potential bodily motions) will be beneficial to the bear, as will the ongoing free-agency phenomenology experienced by the bear during feeding.) Other survival-important features of a creature's ambient environment will involve event-causal goings on that are *not* susceptible to causal influence by the creature's potential bodily motions, but that need attending to (and responding to) if the creature is to survive and flourish. (For instance, if a bear sees a huge boulder rolling down the mountainside in his direction, this ought to be registered by the bear as an state-causal sequence that (i) cannot be *influenced* by certain bodily motions, and (ii) will be big trouble if the bear's body remains where it currently is.)

So, certain aspects of a creature's environment are appropriately registered cognitively as ones that are susceptible to influence by certain potential patterns of behavior by the creature, patterns that in fact can be internally generated by the creature's motion-control mechanisms. Other environmental features are appropriately registered as inevitable state-causal phenomena, and furthermore as apt to have dire consequences unless the agent's own body moves in certain situationally appropriate ways. It will be hugely important, in myriad ways, for this distinction to be registered distinctly and saliently within the agent's phenomenology. Furthermore, in situations where various potential bodily motions would be situationally helpful to an a creature, but where some such behaviors can be generated by the creature's internal motion-control mechanisms but others cannot—say, propelling its body across a certain crevasse—it will hugely important for this difference too to be very salient phenomenologically. Hence the enormous importance, for a wide range of non-human creatures, of agentive phenomenology—with its accompanying range of potential behaviors the agent experiences itself as *able* to perform. And hence too the enormous importance, for such creatures, of a vivid phenomenological difference between agentive phenomenology on one hand (with its ever-present "I can do otherwise" aspect), and on the other hand the phenomenology of state-causation (with its contrasting inevitability aspect).

This is an evidential consideration falling under a category analogous to type 4 above, except that it involves evolutionary-biological point and purpose and it pertains to a kind of intentional content (viz. presentational) that is plausibly present in the mental lives of numerous non-human animals who lack the capacity for conceptually sophisticated judgment. Agentive phenomenology serves vitally important evolutionary-biological purposes. Furthermore, its working as well as it does depends heavily on its aspect of freedom and its exclusiveness from the phenomenology of state-causation. Evolution thus gave its progeny something very useful, when it instilled in them the phenomenology of agency. But although this usefulness does depend on the features we have discussed, would it be enhanced if the satisfaction conditions for the phenomenology of agency required substantially more than the accurate tracking of kinds of differences lately mentioned? Would it be enhanced if agentive phenomenology had satisfaction conditions requiring genuine agency to be a matter of a creature's being a metaphysical-libertarian uncaused cause?

Here it becomes important to remind ourselves about a key fact about us humans, with our sophisticated capacities for judgment and for introspection, a fact I noted earlier—viz., that *answers to compatibility questions about agentive phenomenology are not introspectively*

manifest. Given that we humans (with all our judgmental and introspective sophistication) cannot even tell, by introspection, whether or not agentive phenomenology is compatible with state-causal determinism (or with physical causal closure, or with mental state-causation of behavior), evolution simply would not have been instilling evolutionary-biological advantage had she built incompatibilist requirements into the satisfaction conditions for agentive phenomenology. On the contrary: agentive phenomenology would work just as well, benefit-wise, if its satisfaction conditions were substantially weaker—if they were just a matter of accurately and effectively tracking the kinds of metaphysically non-tendentious distinctions noted above. So, since there is no evolutionary-biological advantage in an incompatibilist agentive phenomenology rather than a compatibilist one, and since agentive phenomenology itself presumably confers enormous evolutionary-biological benefits upon the types of creatures in which it occurs, it is very likely that actual agentive phenomenology simply does not have incompatibilist satisfaction conditions. Non-manifest incompatibilist requirements, as a dimension of the intentional content of agentive phenomenology, would be entirely pointless from the perspective of evolutionary proper function.

This consideration in favor of compatibilism can be supplemented with the following observations. Agentive phenomenology works hand-in-glove epistemologically with first-person judgments about one's own agency. Humans routinely judge of themselves that they are agents of the kind they *experience* themselves to be—with agentive experience serving as the epistemic basis for first-person agency-ascribing judgments. (As stressed already, even when the judgmental notion of freedom is deployed under contextual parameters that count certain behavior as unfree—e.g., behavior that occurs under extreme cooercion—one still experiences such behavior as actional, and it still has the *presentational* aspect of freedom, the 'I could do otherwise' aspect.) Given this epistemological link, the evidential considerations mentioned in section 3.5 in support of compatibilism about freedom-ascribing judgments also favor compatibilism about the evidential basis of such judgments—in the first-person case, compatibilism about agentive phenomenology.

So there is considerable data, of various complementary kinds, supporting compatibilism about the intentional content of agentive phenomenology. So far, so encouraging. But a fully adequate defense of compatibilism should also provide a plausible and satisfying treatment of recalcitrant data—data that prima facie points toward incompatibilism. To that task I turn next.

7. **Explaining Recalcitrant Data**

Of course the philosophical problem I have been addressing would not be a problem at all were it not for the fact that there are also some pertinent phenomena not directly accommodated by compatibilism about agentive phenomenology—phenomena that initially appear to support incompatibilism. Chief among these, perhaps, is the data of the kind I called "Type 2" above—the fact that one's judgment-tendencies are pulled in different ways about the compatibility questions we have been addressing. When one attends introspectively to one's agentive phenomenology, with its *presentational* aspects of freedom and self-as-source, and when one simultaneously asks reflectively whether the veridicality of this phenomenology is compatible with causal determinism (or with physical causal closure, or with the mental state-causation of one's behavior), one feels *some* tendency to judge that the answer to such compatibility questions is No.

If compatibilism is correct, then this tendency embodies a mistake: the satisfaction conditions of agentive phenomenology do not require the falsity of causal determinism, or of physical causal closure, or of the thesis that human actions are state-caused by mental states, qua

mental. But a theoretically adequate compatibilism should provide a plausible *explanation* of this mistaken judgment-tendency—an explanation of why the tendency arises so strongly and so naturally, once the compatibility issues are explicitly raised.

The version of compatibilism I proposed in section 3.4 has two complementary resources to deploy in formulating such an explanation. First is the fact, already stressed, that agentive phenomenology and the phenomenology of state-causation are *mutually exclusionary*. Although there are good evolutionary-biological reasons why this should be so, and although it can perfectly well be so even if human behavior is indeed state-caused, nevertheless it is virtually impossible to simultaneously experience a single item of one's own behavior both as actional and as state-caused. And it is easy to make the mistake of inferring, on the basis of the fact that one cannot *experience* one's own behavior both as action and as state-caused motion, that no item of behavior can *really be* both a genuine action and an state-caused bodily motion. (It is especially easy to make this mistake if one conflates (i) *not* experiencing one's behavior *as* state-caused, with (ii) experiencing one's behaviour *as not* state-caused.) But, psychologically tempting though that inference might be, it is a *non sequitur*. Again: there are good evolutionary-biological reasons why agentive phenomenology and state-causal phenomenology are mutually exclusionary; yet there is no good evolutionary-biological reason why agentive phenomenology should have incompatibilist satisfaction conditions, especially since such extremely demanding satisfaction conditions, if they really do accrue to the presentational content of agentive exerience, are phenomenologically non-manifest.

The second available explanatory resource is the contextualist element that I claim is operative in *judgmental* attributions of freedom, and thereby in judgmental attributions of agency as well. In contexts of philosophical inquiry about the compatibility of freedom and determinism, the very posing of the philosophical question tends to drive the contextually variable implicit parameters governing the judgmental concept of freedom to a maximally strict setting—an unusual setting, under which the satisfaction conditions for freedom-attributions actually are incompatible with determinism. Likewise, in contexts of philosophical inquiry about the compatibility of the *presentational* content of agentive phenomenology with determinism (or with physical causal closure, or with the mental state-causation of behavior), the very posing of such philosophical questions tends to drive the contextually variable implicit parameters governing the *judgmental* notion of agency to a maximally strict setting—an unusual setting, in which the freedom dimension of agency is understood as incompatible with determinism, and in which the self-as-source dimension of agency is understood as a matter of metaphysical-libertarian
agent causation" as distinct from state-causation. It is easy not to notice the presence and operation of implicit contextual parameters, since after all they are not explicit. Thus, it is easy not to notice that the posing of philosophical compatibility questions tends to drive those parameters toward extremal—and highly unusual—settings. Under such settings, incompatibility claims deploying the judgmental concept of agency are in fact correct. An appreciation of such correctness, together with a failure to notice the underlying dynamics of the implicit parameters, can undergird a tendency to mistakenly believe both (i) that *ordinary* uses of the judgmental concept of agency have incompatibilist satisfaction conditions, and (ii) that the presentational content of agentive phenomenology has incompatibilist satisfaction conditions too.

So the version of contextualism I propose allows for a fairly plausible explanation of the incompatibilist-leaning judgment-tendencies that generate the philosophical problem that has been my focus. When one factors this into the mix, alongside the various convergent forms of

evidence that favor compatibilism, the upshot is a strong—albeit admittedly nondemonstrative—overall case in favor of my compatibilist proposal.

4. Agentive Phenomenology and Mental State-Causation

In agentive phenomenology one does not experience one's behavior *as* state-caused at all, and thus one also does not experience one's behavior as state-caused by *mental* states. Indeed, I have stressed already that experiencing one's behavior as action *excludes* simultaneously experiencing it as state-caused, and thus also excludes experiencing it as state-caused by one's own mental states. This means that an item of one's own behavior that is experienced as an action (and thus is manifestly presented as such) cannot also be *manifestly* presented to oneself as being a bodily motion that is state-caused by certain of one's own mental states. Questions remain, however, about whether certain state-cause requirements might nevertheless be *non*-manifestly part of the presentational intentional content of one's agentive phenomenology. In particular, the question arises whether the satisfaction conditions for agentive phenomenology include, non-manifestly, the requirement that the event experienced as an action is *in fact* state-caused by certain of one's own mental states, qua mental.

If compatibilism about the agent-exclusion problem is right, then it is extremely likely that the answer to this question is Yes. For, whatever else might go into a positive compatibilist account of the self-as-source phenomenon, surely a minimal necessary condition is that people really do behave as they do *because of reasons they have*. And if indeed self-as-source-hood is not a matter of metaphysical-libertarian "agent causation," then evidently the only promising alternative for making sense of such mentalistic 'because'-claims is to construe the reasons that explain behavior as *mental state-causes* of behavior. Thus, for those (myself included) who accept the hypothesis of physical causal closure, the full vindication of our belief in our own agency requires a solution to the more familiar causal-exclusion problem in philosophy of mind—the problem of explaining how mental states could be state-causally efficacious, qua mental, despite physical state-causal closure.

I will conclude by briefly summarizing my treatment of this issue. For concreteness, I will organize the discussion around the problem of causal exclusion—although the same general approach can be harnessed to address other philosophical worries about mental state-causation too, such as worries that arise about the efficacy of mental states whose content is constitutively dependent upon external factors.

The exclusion problem about mental causation can be put this way: Each of the following five statements is prima facie credible, and yet they are jointly inconsistent.

1. Physics is causally closed.
2. Mental properties are real, and are instantiated by humans.
3. Mental properties are causal properties.
4. Mental properties are not identical to physical causal properties.
5. If physics is causally closed, then all causal properties are physical causal properties.

Statement 1, the thesis of the causal closure of physics, is the claim that every physical event or state is completely causally determined—to the extent that it is causally determined at all—on the basis of physical laws plus prior physical states, and that the laws of physics are never violated. Statements 3 and 4 are to be understood as making conditional claims—claims about what mental properties are like, *if* there are any such properties and they are instantiated by humans. Statement 2, then, asserts the implicit antecedent of statements 3 and 4. By 'physical property' I mean, essentially, the kind of property posited in fundamental physical theory—i.e., a physics-level property.

Although each of statements 1-5 has substantial initial credibility, they are jointly inconsistent; so at least one of them must be false. What I originally called "causal compatibilism" is a view that repudiates statement 5 and retains statements 1-4. Causal compatibilism, thus understood, asserts that even though physics is causally closed, and even though mental properties are multiply realizable and hence are not identical to physical causal properties, mental properties are causal properties nonetheless. This position asserts that there is genuine causation and genuine causal explanation at multiple descriptive/ontological levels, and that despite the causal closure of physics, physics-level causal and causal-explanatory claims are not really incompatible with mentalistic causal and causal-explanatory claims.

The form of state-causal compatibilism I favor includes three central ideas. The first is a conception of causal-explanatory relevance for properties, involving systematic patterns of counterfactual dependence. In causal explanation the effect phenomenon *e*, described as instantiating a phenomenon type *E*, is shown to depend in a certain way upon the cause phenomenon *c*, described as instantiating a phenomenon of type *C*. Often the dependence involves the fact that *c* and *e* are subsumable under a counterfactual-supporting generalization—either a generalization that directly links *C* to E, or else a more complicated generalization whose antecedent cites a combination of properties that includes *C*. But in order for the cited properties *C* and *E* to be genuinely explanatorily relevant to the causal transaction between *c* and *e*, it is not enough that *c* caused *e* and *c* and *e* are subsumable under such a generalization. Rather, *C* and *E* must fit into a suitably rich pattern of counterfactual relations among properties.

It is important to understand how this feature is related to the structure of scientific laws. The generality of the fundamental laws of the natural sciences, for example, does not consist merely in their having the logical form, "All *A*s are *B*s." It consists, rather, in the fact that they are systematic in scope and structure, so that a wide range of phenomena are subsumable under relatively few laws. One major source of their systematicity is that (1) the laws cite *determinable* properties, namely magnitude-properties, where the determinants are quantitatively specific instances of these properties, and that (2) the laws contain universal quantifiers ranging over these quantitative determinant-values (in addition to the universal quantifiers ranging over the non-numerical entities in the law's domain). Newtonian velocity, for example, is not a single determinate property but an infinite array of determinate properties, one for each real value of the determinable *V*. The resultant generality of a physical law consists largely in the existence of a whole (typically infinite) set of specific nomically true principles, each of which is a specific instantiation of the law with specific numerical values "plugged in" for the determinant-variables. Rich patterns of counterfactual dependence, of the sort that are a crucial feature of successful causal explanation in science, are reflected by the truth of such sets of specific law-instantiations.

Second: Often several distinct patterns of counterfactual dependence, all subsuming a single phenomenon, will involve different descriptive/ontological levels, for example microphysical, neurobiological, macrobiological, and psychological. Consider, for instance, instances of human behavior, vis-à-vis the level of common-sense intentional psychology, so-called folk psychology. There are robust patterns of counterfactual dependence among the state types (including act types) posited by folk psychology—patterns systematizable via generalizations containing universal quantifiers ranging over suitable determinant-values. These determinant-values are not quantitative, but instead are *propositional* (or *intentional*); i.e., they are the kinds typically specified by 'that'-clauses. Take, for instance, relations between actions and reasons. The intentional mental properties that constitute reasons

(namely belief types, desire types, and other attitude types), in combination with act types, clearly figure in a rich and robust pattern of counterfactual dependence of actions upon reasons that rationalize them, a pattern conforming to the following generalization:

> For any subject *S*, desire-content *D*, and action *A*, if *S* wants *D* and *S* believes that doing *A* will bring about *D*, then *ceteris paribus*, *S* will do *A*.

There are also rich patterns of counterfactual dependence among folk psychological mental properties themselves, again systematizable by suitable *ceteris paribus* generalizations involving quantification over propositional/intentional determinant-values. Wanting, believing, etc. figure in these generalizations as determinable properties, and the generalizations characterize vast (possibly infinite), highly structured, counterfactual-dependence relations among the corresponding determinant properties—a different specific dependence relation for each specific instantiation of the propositional variables in the generalizations.

Third: The closely related concepts of causation and causal explanation are contextually parameterized notions, with an implicit contextual parameter keyed to a specific descriptive/ontological level; I call this the *level-parameter*. The contextually relevant counterfactual-dependence patterns, for purposes of evaluating the truth or falsity of causal and causal-explanatory statements in specific contexts of usage, are those patterns that reside at the level determined by the contextually operative level-parameter. (Thus, the position I advocate is a *contextualist* form of causal compatibilism.)

When we bring together the three key ideas just described, the following picture results. A single phenomenon can perfectly well be subject to a variety of different causal explanations, involving properties from a variety of different counterfactual-dependence patterns at different descriptive/ontological levels. Often various different causal and explanatory claims with respect to a given phenomenon, involving properties from various different descriptive/ontological levels, all will be objectively true, since each is grounded in some objective counterfactual-dependence pattern. But the different kinds of causal and causal-explanatory claims will be tethered to different contexts of causal inquiry—contexts in which the level-parameter has different settings, involving different *kinds* of objective counterfactual-dependence pattern. Which kinds of dependence patterns and generalizations are most germane typically will be a context-relative matter, governed largely by the interests of those doing the explaining and inquiring. Choice of descriptive vocabulary normally will have a very heavy influence on the default settings for the contextually relevant parameters, the operative "score" in the causal-explanation game. If we pose our questions and offer our answers in psychological vocabulary, for instance, then normally the relevant patterns of counterfactual dependence will be ones involving psychological properties, with their associated generalizations—including the generalizations of folk psychology.

Since the notions of causation and causal property are both governed by contextually variable parameters, the properties we may properly cite, when we are tallying an inventory of properties or factors that were causally operative with respect to a given phenomenon, fall within the range determined by the current score in the causal-explanation game. Mental properties fall within the contextually eligible range when the score is set for psychological explanation, whereas the neurophysical properties that realize them fall within the contextually eligible range when the score is set for neurophysical explanation. In a normal context of psychological explanation it is not appropriate to count the neurophysical realizers as causal properties in addition to the mental properties, whereas in a normal context of neurophysical causal explanation it is not appropriate to count the mental properties as causal

properties in addition to the neurophysical ones. In either context, such double-counting goes contrary to the contextually operative score in the causal-explanation game.

On this account, causal exclusion reasoning goes wrong because it wrongly treats the notions of causation and causal property as though they are not governed by an implicit level-parameter, when in fact they are. If one ignores the level-parameter (which is easy to do, since it is not explicit), then it will appear that properties are either causal or non-causal, *simpliciter*—and thus that the causal closure of physics just leaves no room for other properties to be causal. In order to be causal, they either would have to be additional fundamental force-generating properties (going contrary to the causal closure of the physical), or would have to be overdetermining causal properties (going contrary to the lack of an independent causal route to the effect). But if indeed the notions of causation and causal explanation have an implicit level-parameter, then this is just the wrong picture of the matter. The basically mistaken idea is that properties are causal, or not causal, *punkt*. This is something like asking what time it is on earth, rather than asking what it is in a given time zone. Which properties count as causal depends upon the parameters governing engaged causal inquiry.

This is not causal irrealism, or explanatory irrealism. For, the relevant counterfactual dependence patterns are all objectively real. Given a contextually fixed value of the level-parameter, it is a perfectly objective matter that certain properties are causal, and that certain phenomena are causally explainable by appeal to the instantiation of those properties. Causation and causal explanation are perspectival and interest-relative, to be sure. But they are *also* objective, because they involve the way particular phenomena fit it into real, objective, patterns of counterfactual dependence.

5. Conclusion and Dedication

Neglect of agentive phenomenology in philosophy of mind has led to neglect of the agent-exclusion problem; that needs to change. Those who are inclined to resist compatibilist treatments of the more familiar problem of causal exclusion in philosophy of mind—viz., the problem about physical state-causation of behavior allegedly excluding mental state-causation of behavior—should feel intellectual pressure, by parity of reasoning, to resist compatibilist treatments of the agent-exclusion problem too—i.e., treatments that seek to establish the compatibility of the intentional content of agentive experience with the mental state-causation, qua mental, of behavior (and with determinism, and with physical state-causal closure). But that in turn creates pressure either to embrace a metaphysical-libertarian conception of human agency that appears very much at odds with commonly held metaphysical theses (e.g., the thesis of physical state-causal closure), or else to assert that humans are not really agents of the kind they experience themselves to be. I recommend avoiding this bind by going compatibilist about the agent-exclusion problem, and also about the more familiar problem of state-causal exclusion. And I recommend doing so in a way that invokes contextualism about concepts like freedom, agency, and indeed causation itself.

I have not explicitly discussed the work of Jaegwon Kim in this paper. But Kim's important and influential writings in philosophy of mind—in particular, his writings about the causal exclusion problem for mental states qua mental—obviously are very much in the background of my own thinking about exclusion issues. Jaegwon was my teacher and dissertation supervisor at the University of Michigan in the early 1970's, and from then until now his philosophical writings have been a beacon of inspiration for my own work in philosophy of mind and metaphysics. I dedicate this paper to him, with gratitude and respect.[15]

References

Chisholm, R. (1964). Human Freedom and the Self. The Langley Lecture (University of Kansas). Reprinted in J. Feinberg and R. Shafer-Landau, eds., *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, 11th Edition (Wadsworth, 2002), 492-99.

Chisholm, R. (1995). Agents, Causes, and Events: The Problem of Free Will, in T. O'Connor, ed., *Agents, Causes, and Events: Essays on Indeterminism and Free Will* (Oxford), 95-100.

Henderson, D and Horgan, T. (2000). What Is A Priori and What Is It Good For?, *Southern Journal of Philosophy* 38, Spindel Conference Supplement on the Role of the Empirical and the A Priori in Philosophy, 51-86.

Horgan, T. (1979). `Could', Possible Worlds, and Moral Responsibility, *Southern Journal of Philosophy* 17, 345-58.

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

Horgan, T. (1991). Actions, Reasons, and the Explanatory Role of Content, in B. McLaughlin, ed., *Dretske and His Critics* (Basil Blackwell), 73-101.

Horgan, T. (1993a). The Austere Ideology of Folk Psychology. *Mind and Language* 8, 282-97.

Horgan, T. (1993b). Nonreductive Materialism and the Explanatory Autonomy of Psychology. In S. Wagner & R. Warner, eds., *Naturalism: A Critical Appraisal* (Notre Dame), 295-320.

Horgan, T. (1998). Kim on Mental Causation and Causal Exclusion, *Philosophical Perspectives* 11, 165-84.

Horgan, T. (2001a). Causal Compatibilism and the Exclusion Problem. *Theoria* 16, 95-116.

Horgan, T. (2001b). Multiple Reference, Multiple Realization, and the Reduction of Mind. In F. Siebelt and B. Preyer, eds., *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*. Rowman & Littlefield, 205-21.

Horgan, T. and Graham, G. (1994). Southern Fundamentalism and the End of Philosophy, *Philosophical Issues* 5, 219-47. Reprinted in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophy*, Rowman and Littlefield, 1998.

Horgan, T. and Tienson, J. (1990). Soft Laws, *Midwest Studies in Philosophy* 15, 256-79.

Horgan, T. and Tienson, J. (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. In D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford, 520-33.

Horgan, T. and Tienson, J. (2005). The Phenomenology of Embodied Agency. In M. Saagua and F. de Ferro (eds.), *A Explicacao da Interpretacao Humana: The Explanation of Human Interpretation. Proceedings of the Conference Mind and Action III—May 2001*. Lisbon: Edicoes Colibri, 415-23.

Horgan, T., Tienson, J., and Graham, G. (2003). The Phenomenology of First-Person Agency. In S. Walter and H. D. Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Imprint Academic, 323-40.

Horgan, T., Tienson, J., and Graham, G. (2004) Phenomenal Intentionality and the Brain in a Vat. In R. Schantz (ed.), *The Externalist Challenge*. Walter de Gruyter, 297-317.

Kriegel, U. (forthcoming). The Phenomenologically Manifest.
Stephens, G. L. and Graham, G (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. MIT Press.

[1] Here and throughout I speak of 'state-causation' rather than 'event-causation'. More below on my reasons for this choice of terminology; see note 4. States can be short-lived, and often when they are they also fall naturally under the rubric 'event.'

[2] Also important is the third-person phenomenology of agency, the "something it is like" to experience *others* as agents who are acting for reasons. For discussion, see Horgan and Tienson (2005).

[3] For discussion of a range of psychopathological disorders involving similar sorts of dissociative experience, see Stephens and Graham (2000).

[4] The language of causation does seem right here: you experience your behavior as caused by you yourself, rather than experiencing it as caused by *states* of yourself. Metaphysical libertarians about human freedom sometimes speak of "agent causation" (or "immanent causation"), and such terminology seems *phenomenologically* apt regardless of what one thinks about the intelligibility and credibility of metaphysical libertarianism. Chisholm (1964) famously argued that immanent causation (as he called it) is a distinct species of causation from event causation (or "transeunt" causation, as he called it). But he later changed his mind (Chisholm 1995), arguing instead that agent-causal "undertakings" (as he called them) are actually a species of event-causation themselves—albeit a very different species from ordinary, nomically governed, event causation. Phenomenologically speaking, there is indeed something episodic—something temporally located, and thus "event-ish"—about experiences of self-as-source. Thus, the expression 'state causation' works better than 'event causation' as a way of expressing the way behaviors are *not* presented to oneself in agentive experience. Although agentive experience is indeed "event-ish" in the sense that one experiences oneself as undertaking to perform actions *at specific moments in time*, one's behavior is not experienced as caused by *states* of oneself.

[5] The points made in this and the next paragraph, about different ways the phenomenology of purpose can work, are closely connected to the typology of different kinds of phenomenology of doing in Horgan and Tienson (2005).

[6] With respect to successively more fine-grained details of action, specific purposes tend to be progressively less explicit phenomenologically, and progressively less accessible to consciousness—even for actions that result from conscious deliberation. For instance, when you consciously and deliberately decide to get yourself a beer by walking to the fridge in the kitchen and removing a beer from the fridge, the specific purpose in virtue of which your perambulatory trajectory toward the fridge angles through the kitchen doorway, as opposed to taking you directly toward the fridge and smack into the intervening wall, normally will color the phenomenology of your action without becoming explicitly conscious at all. And in some cases, sufficiently fine-grained aspects of one's action might lack even this kind of subtle, non-explicit, phenomenological tinge of specific-purpose phenomenology. For instance, when you grab a can of peas from the grocery shelf, there might be nothing in the phenomenology that smacks even slightly of a specific purpose for grabbing the particular can you do rather than any of several other equally accessible ones. (Indeed, maybe there *is* no specific purpose for grabbing this can rather than any of the others, let alone a purpose that leaves a phenomenological trace.)

[7] This is not to deny, of course, that there is indeed a distinctive phenomenology of effort of will that *sometimes* is present in the phenomenology of doing. The point is just that this aspect is not always present. A related phenomenological feature, often but not always

present, is the phenomenology of *trying*—which itself is virtually always a dimension of the phenomenology of effort of will, and which often (but not always) includes a phenomenologically discernible element of uncertainty about success. (Sometimes the phenomenological aspect of voluntariness attaches mainly to the trying dimension of the phenomenology of doing. When you happen to succeed at what you were trying to do but were not at all confident you could accomplish—e.g., sinking the 10 ball into the corner pocket of the pool table—the success aspect is not experienced as something directly under voluntary control.)

[8] The reason for formulating this thesis in terms of "paradigmatic" behaviors experienced as actions is to allow for the possibility that some such behaviors both (a) have physical causes that are entirely distinct from the brain-states on which the subject's agential experience supervenes, and (b) are not really actions at all even though they are experienced as actions.

[9] On this topic see Kriegel (forthcoming).

[10] Let me stress an important point about my use in this paper of the 'experiencing-as' locution. Experiencing something *as* something (e.g., experiencing an item of one's own behavior as an action) is a matter of the *manifest* aspects of the experience. Thus, to say that one cannot simultaneously experience an item of one's own behavior *as an action* and *as an event-caused motion* is to say that one cannot have an experience in which both aspects are *manifestly* simultaneously present, vis-à-vis a single item of one's own behavior. This leaves open whether, for instance, it is a non-manifiest feature of one's agentive phenomenology that its intentional content has satisfaction conditions requiring an action to be an episode that is state-caused by certain mental states, qua mental. I return to this theme in section 4 below.

[11] Chishom held that libertarian "immanent causation" involves an agent's directly bringing about *some* state that is not itself state-caused, but he allowed that the action as a whole might also include a subsequent state-causal chain. Presumably the immanently brought-about state was supposed to be a brain-state, which then triggered a state-causal sequence leading to muscular motion.

[12] I assume here that phenomenology in general, and agentive phenomenology more specifically, is *causally efficacious*, qua mental, with respect to behavior—a matter I return to in section 4 below.

[13] Typically, I believe, such generalizations will be what are called "soft laws" in Horgan and Tienson (1990). In the case of psychology, for instance, this means that the generalizations will have ineliminable *ceteris paribus* clauses adverting not merely to potential lower-level exceptions resulting from physical breakdown (e.g., having a stroke) or from external physical interference (e.g., being hit by a bus), but adverting to potential psychology-level exceptions as well.

[14] I think that other implicit parameters operate too, typically in an intra-level manner. These are related to the contextually appropriate way of distinguishing between what counts as cause, and what counts instead as "background conditions."

[15] A draft of this paper was presented at the 2005 conference was presented at the 2005 conference on Mental Causation, Externalism, and Self-Knowledge at the University of Tuebingnen. My thanks to the participants at that conference (including Jaegwon Kim) for their feedback, and to Christian Sachse for his commentary. Thanks too to Michael Gill, George Graham, Uriah Kriegel, Keith Lehrer, Cei Maslen, Sean Nichols, Susanna Siegel, John Tienson, and Mark Timmons for ongoing discussion and feedback.