Erkenn (2007) 67:183–200
DOI 10.1007/s10670-007-9067-9

ORIGINAL ARTICLE

Mental Causation and the Agent-Exclusion Problem

Terry Horgan

Received: 30 July 2006 / Accepted: 30 April 2007 / Published online: 11 August 2007

Abstract The hypothesis of the mental state-causation of behavior (the MSC hypothesis) asserts that the behaviors we classify as actions are caused by certain mental states. A principal reason often given for trying to secure the truth of the MSC hypothesis is that doing so is allegedly required to vindicate our belief in our own agency. I argue that the project of vindicating agency needs to be seriously reconceived, as does the relation between this project and the MSC hypothesis. Vindication requires addressing what I call the agent-exclusion problem: the prima facie incompatibility between the intentional content of agentive experience and certain metaphysical hypotheses often espoused in philosophy–metaphysical hypotheses like physical causal closure, determinism, and the MSC hypothesis itself. I describe several radically different approaches to the vindication project, one of which would repudiate the MSC hypothesis and embrace metaphysical libertarianism about freedom and determinism. I sketch the position I myself favor–a specific version of the generic approach asserting that the intentional content of agentive experience is compatible with the MSC hypothesis (and with physical causal closure, and with determinism). I describe how my favored approach can plausibly explain the temptation to embrace incompatibilism concerning the agent-exclusion problem.

1 Introduction

Let the hypothesis of the *mental state-causation of behavior* (the MSC hypothesis) be the claim that the behaviors we classify as actions are caused by certain mental states (qua mental)—states like beliefs, desires, intentions, and decisions. In the recent philosophical literature on mental causation, it is commonly claimed that one central reason for wanting to secure the MSC hypothesis is to vindicate our belief in our own agency. A representative example is the following passage from Kim (2005):

T. Horgan (✉)
Philosophy Department, University of Arizona, Tucson, AZ 85721, USA
e-mail: thorgan@email.arizona.edu

> Let us first review some reasons for wanting to save mental causation—why it is important to us that mental causation is real. First and foremost, the possibility of human agency, and hence our moral practice, evidently requires that our mental states have causal effects in the physical world. In voluntary actions our beliefs and desires, or intentions and decisions, must somehow cause our limbs to move in appropriate ways, thereby causing the objects around us to be rearranged. That is how we manage to navigate around the objects in our surroundings, find food and shelter, build bridges and cities, and destroy the rain forests. (p. 9)

Several features of this familiar way of thinking about mental causation vis-à-vis agency are especially noteworthy. First is the explicit assumption that the truth of the MSC hypothesis is a prerequisite of genuine agency. Second is the implicit assumption that the primary philosophical task that must be accomplished, in order to vindicate our belief in our own agency, is to secure this hypothesis about the etiology of behavior. Third is the lack of any acknowledgement or discussion of first-person agentive *phenomenology*: the "what it is like" of voluntarily acting (or of experiencing oneself as voluntarily acting, at any rate).

In this paper, I will argue that we should reconceive in significant ways the philosophical issues that arise concerning the relation between human agency and the MSC hypothesis. Some of my central themes will be the following. First, agentive phenomenology should receive explicit attention. (It has mostly been ignored for the past 60 years in analytic philosophy of mind, and often its very existence has been implicitly denied.) Second, agentive phenomenology provides prima facie reason to think that a prerequisite for genuine agency is that the behaviors we experience as actions are not caused by prior states—not by states of the external environment, not by internal states of the agent, and not by some combination of the two. (The voluntariness aspect of agentive phenomenology is especially important here—and brings into play concerns that are closely related to the traditional metaphysical problem of freedom and determinism. The key idea is that agentive phenomenology presents one's behavior to oneself as caused *by oneself*—and does not present it as being caused by mental *states* of oneself.) Third, agentive phenomenology is therefore in prima facie conflict with various metaphysical hypotheses often embraced in philosophy—e.g., the hypothesis of determinism, and the hypothesis of physical causal closure. (I call this the *agent-exclusion problem*.) Fourth, the agent-exclusion problem also arises concerning the MSC hypothesis itself; thus, the MSC hypothesis is a prima facie *threat* to our belief in our own agency. (This is palpably ironic, given the conventional idea that securing the MSC hypothesis is the key to *vindicating* that belief.) Fifth, the philosophical project of vindicating our belief in our own agency needs to be re-thought, in light of the agent-exclusion problem and in light of the problem's applicability to the MSC hypothesis. Sixth, three very different potential approaches to this vindication project emerge as philosophically serious competitors: (1) a position asserting (i) that the experience of agency is systematically illusory and non-veridical, and hence (ii) that our common belief in our own agency is just mistaken; (2) a position asserting (i) that humans are indeed agents of the sort they

experience themselves to be, (ii) that such agency is incompatible with the MSC hypothesis, and also with determinism and with physical causal closure, and (iii) that these three metaphysical hypotheses are false; and (3) a position asserting (i) that humans are indeed agents of the sort they experience themselves to be, (ii) that the content of agentive experience is compatible with the MSC hypothesis, and also with determinism and with physical causal closure, (iii) that the truth of the MSC hypothesis is a prerequisite for genuine agency (as recent philosophy of mind has typically assumed), (iv) that the MSC hypothesis is true, (v) that the hypothesis of physical causal closure is also true, and hence (vi) that the MSC hypothesis is compatible with physical causal closure.

In my view, position (3) is the right one on these matters, and positions (1) and (2) are mistaken. Thus, the right agenda to be pursuing, with respect to the generic philosophical project of defending our belief in our own agency, is the agenda of providing viable philosophical defenses of the various claims that jointly comprise position (3). That agenda includes not just one, but rather two, principal tasks—not just the commonly-pursued task of arguing that the MSC hypothesis is true even if the hypothesis of physical causal closure is also true, but also the hitherto unappreciated task of arguing that the content of agentive experience is compatible with both of these hypotheses (and with determinism).

I will sketch the specific version of position (3) I favor, whose various constitutive claims I have defended elsewhere. And I will explain how this version of position (3) can explain away—as a subtle and tempting mistake—the tendency to think that the intentional content of agentive experience is incompatible with the MSC hypothesis (and with physical causal closure, and with determinism.)

2 The Phenomenology of First-Person Agency[1]

Lately, George Graham and John Tienson and I 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 (Horgan et al. 2003; Horgan and Tienson 2005; Horgan 2007).[2] We maintain that agentive phenomenology is richly intentional, presenting in experience a self that is an apparently embodied, apparently voluntarily behaving, agent. Since 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

[1] This section is adapted, with some modifications and additions, from Sect. 2 of Horgan et al. (2003).

[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 (cf. Horgan and Tienson 2005). Addressing the phenomenology of agency is part of a larger joint project with Graham and Tienson, arguing that the most fundamental kind of mental intentionality, which we call *phenomenal* intentionality, is phenomenally constituted, is narrow, and comprises not only sensory-perceptual experience but also cognitive states like occurrent thoughts, conative states like occurrent wishes, and agentive experience. See Horgan and Tienson (2002), Horgan et al. (2004, 2007, in press).

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 Sect. 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 as 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, I 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 in 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*.[3] 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

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

a tree in its path is experienced as causing the tree to become uprooted; and so on. Sometimes they even experience their own bodily motions as state-caused by their own mental states—e.g., when one feels oneself shuddering and experiences this shuddering as caused by of a state of fear. 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.[4]

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 state-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 generated by *yourself*.[5]

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.[6] 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

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

[5] The language of causation seems apt here too: 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.

[6] 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).

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 non-deliberative action, with specific purposes coloring conscious experience even when they are not explicitly conscious themselves.[7]

[7] 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.)

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 is so even in situations where one acts under extreme coercion or duress. If a robber points a gun in my face and says "Your money or your life," the phenomenological I-could-do-otherwise aspect is present when I conform to this demand, even though I consider it patently irrational not to do as the robber demands.[8] 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.[9])

Associated with the voluntariness dimension of agentive phenomenology is the following fact about the experience of reasons as motives. Although often one does experience certain conscious reasons (e.g., occurrent beliefs, occurrent wishes, etc.) as playing a state-causal role in relation to one's action, this role is experienced as one's being *inclined* by those reasons to perform the given action; the role is *not* experienced as one's action being *necessitated* by those reasons. Such experiences of one's mental states playing a state-causal motivational role—viz., the role of state-causing an *inclination* toward an outcome—are importantly different from experiences of outright state causation of an outcome itself. For, when one undergoes experiences of state causation of an outcome, typically the cause *is* experienced as necessitating the effect (in the experienced circumstances). The phenomenology of necessary connection between cause and outcome is introspectively palpable in such cases (although it seems to involve some form of necessity other than logical or conceptual). By contrast, it is palpably *absent* in the phenomenology of agency.

Agentive phenomenology is more closely akin to perceptual/kinesthetic experience than it is to discursive thought. (Many higher non-human animals, I take it, have some agentive phenomenology, even if they engage in little or no discursive thought.) Of course, we humans also wield *concepts* like agency, voluntariness, and

[8] This is so even though there certainly are uses of 'could', especially in contexts of moral evaluation, under which it would be correct to say about such a situation, "I could not do otherwise, because of the coercive threat."

[9] 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 it—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.)

the like (whereas it is questionable whether non-human animals do); but thoughts employing these concepts are not to be conflated with agentive phenomenology itself.

3 The Agent-Exclusion Problem

Agentive phenomenology has content that is richly intentional: such phenomenology has satisfaction conditions. Philosophical questions thus arise about the nature of these satisfaction conditions—including questions about whether the intentional content of agentive experience is compatible with various metaphysical hypotheses that are seriously entertained in philosophy. One such hypothesis is *determinism*, which asserts that the laws of nature are such that there are no two nomically possible worlds that are exactly alike up to some moment in time but differ thereafter. A second is the hypothesis of *physical state-causal closure*, which asserts that the state of the world at any moment in time, insofar as it is diachronically determined at all, is diachronically determined by prior physics-level phenomena, on the basis of the fundamental laws of physics. A third is the hypothesis of the mental state-causation of behavior (the MSC hypothesis), which asserts that normally the behaviors that one experiences as one's actions are state-caused by certain psychological states of oneself, such as occurrent wants in combination with occurrent beliefs.

The agent-exclusion problem, as I call it, is the (prima facie plausible) possibility that the intentional content of agentive experience is incompatible with one or another of these metaphysical hypotheses. If indeed there is such content-incompatibility, and if the given hypothesis is also true, then the apparent upshot is that people are not really agents of the sort they experience themselves to be: genuine agency is metaphysically excluded, and agentive experience is systematically non-veridical. (Agent-exclusion could arise in other ways too. Suppose, for instance, that (a) the content of agentive experience is incompatible with determinism, (b) determinism is false, but (c) to the extent that any phenomenon is not state-causally determined, it is just *random*. Arguably, if (a)–(c) obtain, so that human behavior is random to whatever extent it is not state-causally determined, then human behavior never qualifies as genuine action: it never really emanates from the *self* as its source, even though the agent *experiences* it as so generated.)

Why is it prima facie plausible that the content of agentive phenomenology is incompatible with the MSC hypothesis, and with the other two above-mentioned hypotheses too? The virtually ubiquitous aspect of *freedom* in agentive phenomenology figures crucially, as an essential dimension of self-as-source experience. Experiencing one's behavior as produced by oneself is fundamentally different from experiencing it as caused by internal *states* of oneself; and one key aspect of the experiential difference is that one does not experience the behavior as state-causally necessitated. Instead, one experiences the actually-performed behavior as one among a range of alternative behaviors that are genuinely open to oneself—are real alternative possibilities that could be performed instead. But if the behavior that is thus experienced is nonetheless state-causally determined, then it is necessitated

after all—even though it is not *experienced* as necessitated. Hence (one might well think), if the behavior is really state-caused, then it is not a piece of genuine action at all; the phenomenology of agency is illusory and non-veridical. The real source of one's behavior is not really *oneself*, but instead is a *state* of oneself (or a combination of such states). This is a very familiar line of thought in philosophical discussions of freedom and determinism, although here it is being applied directly to agentive phenomenology.

4 Reconceiving the Vindication of Agency

Suppose it turns out that the intentional content of agentive phenomenology is actually incompatible with the MSC hypothesis—because *oneself's* being the source of a piece of behavior turns out to be incompatible with that behavior's having been caused by mental *states* of oneself. Then much recent philosophy of mind thereby will turn out to have been profoundly mistaken in its common assumption that the truth of the MSC hypothesis is a prerequisite of genuine agency. Quite the opposite will be true, viz., that behaviors that are genuine actions cannot be state-caused at all—and a fortiori, cannot be *mentally* state-caused. Securing the MSC hypothesis will turn out to have been entirely the wrong strategy for vindicating our belief in our own agency.

So, in light of the agent-exclusion problem, the philosophical project of vindicating our belief in our own agency needs to be reconceived. It won't do just to set oneself the task of securing the MSC hypothesis. Indeed, if genuine agency is in fact incompatible with the MSC hypothesis—as seems prima facie plausible—then vindicating our belief in our agency would require *refuting* the MSC hypothesis!

Several ways of approaching the project of vindication can be envisioned, each very different from the others. (They are close cousins to the three classic approaches to the problem of freedom and determinism—"hard determinism," metaphysical libertarianism, and compatibilism.) First is the highly pessimistic approach, which asserts that our belief in our own agency is just plain false. It is false because it is incompatible with several metaphysical hypotheses at least one of which is true: the MSC hypothesis, the hypothesis of determinism, and the hypothesis of physical causal closure. In order to be a genuine agent of one's actions, one would have to be a *source* of one's actions in the way one experiences oneself to be. But one is never such a source of one's behavior, because that would require the falsity of all three metaphysical hypotheses. So we are not really agents at all: the content of agentive experience is thoroughly illusory and non-veridical. We must learn to live with this sobering fact.[10] (This

[10] Living with the non-reality of agency would not necessarily mean ceasing to have agentive experience. That is probably psychologically impossible, for normal humans. (It is also psychologically impossible for a normal human to cease having color experience, for example, even though one philosophically respectable view about color asserts that there are no colors and that color experience is systematically non-veridical.)

position is the analogue of "hard determinism" concerning the traditional problem of freedom and determinism.[11])

Such pessimism about agency, I take it, is a last-resort position. Many philosophers, myself included, will find themselves utterly unable to believe it—whether or not they think they have a good handle on how one might actually go about vindicating our belief in our own agency and in the veridicality of agentive experience.

A second approach goes as follows. In order to vindicate belief in human agency, one must (i) make good sense of something like the conception of agency championed by those known as "metaphysical libertarians" concerning the traditional philosophical problem of freedom and determinism, and (ii) argue persuasively that the metaphysical-libertarian conception is actually applicable to much of human behavior (or at least to some of it). Not only is this a very tall order, but it is utterly different from—indeed, is in direct conflict with—defending the MSC hypothesis. Instead, this vindication project is essentially the long-familiar project of articulating and defending libertarian freedom.

Many philosophers, myself included, will find themselves strongly inclined to think that the libertarian project just described is utterly hopeless. For one thing, many of us believe that there is overwhelmingly strong empirical evidence for the hypothesis of physical state-causal closure, and that this hypothesis is incompatible with the metaphysical-libertarian conception of agency. Also, metaphysical libertarians have always had a notoriously difficult time trying even to provide an intelligible and intellectually satisfying *formulation* of the metaphysical doctrine they seek to espouse—let alone arguing that such a doctrine actually applies to humans.[12]

A third approach to the vindication of agency is what I will call the project of *compatibilist vindication*. (As the label suggests, it is closely connected to compatibilism about freedom and determinism.) The project comprises two principal tasks. First is to defend the following *agent-phenomenology compatibility hypothesis* (AP compatibility hypothesis): the content of agentive phenomenology is compatible with the MSC hypothesis—and with physical state-causal closure, and with determinism. Negatively, this means arguing that although claims about putative incompatibility are indeed prima facie plausible, they are nonetheless mistaken. Positively, it means (i) providing, or at least sketching, a compatibilist account of what constitutes the phenomenon of "self as source" of behavior, and (ii) arguing in favor of the account.

[11] Also, various views about morals and moral responsibility could be wedded to the position, analogous to the different views that have been defended by hard determinists. Among the possibilities are (i) that the institution of morality is totally indefensible, or (ii) that some aspects of morality are defensible, but not those involving notions like responsibility, blameworthiness, and praiseworthiness, or (iii) that people can behave morally responsibly and irresponsibly, and can deserve praise or blame for their behavior, even though their behavior does not constitute genuine action.

[12] For two noteworthy recent attempts to address both the formulation problem and the task of arguing that humans really conform to the libertarian conception of free agency, see Kane (1996) and O'Connor (2000).

The second principal task, within the project of compatibilist vindication, is the familiar one in recent philosophy of mind: defending the MSC hypothesis itself. This needs doing because, 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 the libertarian vindication-project is not viable, 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 key burden in defending the MSC hypothesis is to argue for the *mental state-causation compatibility hypothesis* (the MSC compatibility hypothesis): the assertion that mental-state causation is compatible with physical causal closure.

So, insofar as one reconceives the defense of our belief in our own agency in terms of the compatibilist vindication project, it turns out to be true after all that such a defense requires securing the MSC hypothesis—which in turn, for those of us who accept physical causal closure, principally requires securing the MSC compatibility hypothesis. (The main philosophical *worry* about mental state-causation, after all, is the threat that it is excluded by ubiquitous physical causation.) The passage from Kim that I quoted at the beginning is correct, as far as it goes. But more needs doing than that. The full vindication of agency requires not only a defense of the MSC compatibility hypothesis, but also a defense of the AP compatibility hypothesis as well. This lesson needs to be taken to heart in philosophy of mind. What is needed is to make an overall case for what I will call *agentive compatibilism*—the view comprising both the MSC compatibility hypothesis and the AP compatibility hypothesis.

5 A Version of Agentive Compatibilism

In this Sect. I will briefly sketch the version of agentive compatibilism I favor. As regards the MSC hypothesis, my favored approach claims (1) that this hypothesis is true, (2) that the hypothesis of physical causal closure is also true, and hence (3) that these two hypotheses are compatible with one another. It also claims (4) that the two hypotheses are each compatible with state-causal determinism, while leaving open the question whether it is true or false. My recommended approach is also a form of contextualism about causation: it asserts that the concept of cause is governed by implicit, contextually variable, semantic parameters. In typical contexts in which causal efficacy is attributed to mental states qua mental, these implicit semantic parameters operate in such a way that claims about mental state-causation can perfectly well be true—even though the hypothesis of physical state-causal closure obtains. Although there are unusual contextual parameter-settings under which claims about the causal efficacy of mental states are false, the fact remains that such claims are often true under the parameter-settings normally in force when those claims are made. (This is analogous to a parallel contention espoused by contextualists about the concept of knowledge: viz., that although there are unusual contextual parameter-settings for 'know' under which knowledge claims about the

external world are all false—settings that tend to be induced by questions like "How do you know that you're not a brain in a vat?"—the fact remains that external-world knowledge claims are often true under the parameter-settings for 'know' that are normally in force when those claims are made.) I have articulated and defended a specific version of this contextualist approach to the MSC hypothesis in a number of papers (Horgan 1989, 1991, 1993, 1998, 2001a, b, forthcoming). Contextualism about state-causation is an idea that has received far less attention in philosophical discussions of mental state-causation than I think it deserves.[13] In my view, it is the right way to secure the MSC compatibility hypothesis—and thereby is the right way to secure the MSC hypothesis itself, consistently with the hypothesis of physical causal closure.

I maintain that the AP compatibility hypothesis is true too, over and above the MSC compatibility hypothesis—i.e., the content of agentive phenomenology is compatible with the MSC hypothesis, rather than conflicting with it (and is likewise compatible with physical causal closure and with state-causal determinism). I will turn shortly to a sketch of my position on this matter. As a prelude to doing so, let me distinguish two kinds of mental intentionality, which I 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—at any rate, judgmental content involving concepts like freedom or agency.

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

[13] For an overview of philosophical issues of mental causation that gives prominent attention to contextualism, see Maslen et al. (forthcoming). Other contextualist treatments of issues involving mental causation include Menzies (2003), Maslen (2005), and Carroll (forthcoming).

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, 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.

Let me now set forth my favored position about compatibility questions concerning the content of agentive phenomenology. This position, a specific implementation of the AP compatibility hypothesis, comprises the following ten 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; i.e., one cannot reliably tell, just on the basis of careful introspective attention to one's own agentive experience, whether or not the compatibility hypothesis is true. Third, despite the compatibility of agentive phenomenology with the three hypotheses about state-causation, 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 coercion 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 events*. 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 coercion—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 exercised 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.

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 coercion 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 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.

Elsewhere, sometimes collaboratively, I have set forth arguments in support of the various theses constituting this package-deal version of AP compatibilism. Contextualist compatibilism about the judgmental concept of freedom, in a form that acknowledges limit-case parameter-settings that are incompatibilist, is defended in Horgan (1979), Graham and Horgan (1994), Henderson and Horgan (2000), and Horgan (forthcoming). Other aspects of the full package-deal are defended in Horgan (2007, forthcoming). I will not argue for the position here, because of space limitations.

I do recognize that 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 event-causation of one's behavior), one feels *some* tendency to judge that the answer to such compatibility questions is No. If AP 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 MSC hypothesis. I certainly acknowledge that a theoretically adequate AP 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. So let me briefly address this challenge.

The version of AP compatibilism I have described 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*: it is virtually impossible to simultaneously *experience* a single item of one's own behavior both as actional and as state-caused. 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 a 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 behavior *as not* state-caused.) But, psychologically tempting though that inference might be, it is a *non sequitur*.[14]

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 event-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

[14] In Horgan (forthcoming) I argue that there are good evolutionary-biological reasons why agentive phenomenology and state-causal phenomenology are mutually exclusionary. I also argue 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*. These arguments throw further into doubt any inference from the fact that agentive phenomenology and state-causal phenomenology are mutually exclusionary to the conclusion that agentive presentational content has incompatibilist satisfaction conditions.

conditions, and (ii) that the *presentational* content of agentive phenomenology has incompatibilist satisfaction conditions too.

So the version of AP compatibilism 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 (not set forth here) that favor both AP compatibilism and MSC compatibilism, I think a strong case can be made in support of the overall agentive-compatibilist position.

6 Conclusion

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, also 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 MSC hypothesis (and with determinism, and with physical causal closure). Conversely, those who are inclined to believe in the reality of human agency, and who also are inclined to believe in metaphysical hypotheses like the MSC hypothesis and the causal closure of physics, should feel intellectual pressure, by parity of reasoning, to embrace a compatibilist approach not only to the familiar causal-exclusion problem, but also to the agent-exclusion problem. That is, they should feel pressure to embrace the overall position that I call agentive compatibilism, comprising both the MSC compatibility hypothesis and the AP compatibility hypothesis.

A potentially promising way to implement agentive compatibilism—the way I favor—treats both the concept of cause and the concept of agency as governed by contextually variable implicit semantic parameters. This contextual element affects *judgmental* intentional content involving agency and/or state-causation, but does not affect the *presentational* content of agentive experience. Judgmental contextualism about the concept of causation is the key to defending the MSC compatibility hypothesis, thereby resolving the familiar causal-exclusion problem in philosophy of mind. But in addition, judgmental contextualism about the concepts of freedom and agency yields an attractive implementation of the AP compatibility hypothesis too, thereby resolving the agent-exclusion problem. Judgmental contextualism about freedom and agency, and the phenomenological fact that those of one's own behaviors that one experiences as actions cannot also be experienced as state-caused events, together generate a plausible explanation of why it is so easy to form the mistaken belief that genuine agency is incompatible with state-causal determinism, with physical state-causal closure, and with the mental state-causation of behavior.

Acknowledgements This paper elaborates upon and further develops some themes from Horgan (forthcoming), a paper I presented at the 2005 conference on Mental Causation, Externalism, and

Self-Knowledge at the University of Tuebingen. My thanks to the participants of that conference 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, John Pollock, Susanna Siegel, John Tienson, and Mark Timmons for ongoing discussion and feedback, and to two anonymous referees for comments on an earlier version.

References

Carroll, J. (forthcoming). Making exclusion matter less.

Chisholm, R (1964). Human freedom and the self. The Langley lecture (University of Kansas). (Reprinted in J. Feinberg & R. Shafer-Landau (Eds.) (2002), Reason and responsibility: Readings in some basic problems of philosophy, 11th edn. (pp. 492–499). Wadsworth.).

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* (pp. 95–100). Oxford, UK: Oxford University Press.

Graham, G., & Horgan, T. (1994). Southern fundamentalism and the end of philosophy. *Philosophical Issues, 5*, 219–247.

Henderson, D., & Horgan, T. (2000). What is a priori and what is it good for? *Southern Journal of Philosophy, 38*, Spindel Conference Supplement, 51–86.

Horgan, T. (1979). 'Could', possible worlds, and moral responsibility. *Southern Journal of Philosophy, 17*, 345–358.

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* (pp. 73–101). Oxford, UK: Basil Blackwell.

Horgan, T. (1993). Nonreductive materialism and the explanatory autonomy of psychology. In S. Wagner & R. Warner (Eds.), *Naturalism: A critical appraisal* (pp. 295–320). Notre Dame University Press.

Horgan, T. (1998). Kim on mental causation and causal exclusion. *Philosophical Perspectives, 11*, 165–184.

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 & B. Preyer (Eds.), *Reality and Humean supervenience: Essays on the philosophy of David Lewis* (pp. 205–221). New York: Rowman & Littlefield.

Horgan, T. (2007). Agentive phenomenology and the limits of introspection, Psyche 13/2.

Horgan. T. (forthcoming). Causal compatibilism about agentive phenomenology. In T. Horgan, M. Sabates, & D. Sosa (Eds.), *Supervenience mind: Essays in honor of Jaegwon Kim.* Cambridge MA: MIT Press.

Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 520–533). Oxford: Oxford University Press.

Horgan, T., & Tienson, J. (2005). The phenomenology of embodied agency. In M. Saagua & F. de Ferro (Eds.), *A explicacao da interpretacao humana: The explanation of human interpretation*. Proceedings of the conference "mind and action III—May 2001" (pp. 415–423). Lisbon: Edicoes Colibri.

Horgan, T., Tienson, J., & Graham, G. (2003). The phenomenology of first-person agency. In S. Walter & H. D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 323–340). Exeter: Imprint Academic.

Horgan, T., Tienson, J., & Graham, G. (2004). Phenomenal intentionality and the brain in a vat. In R. Schantz (Ed.), *The externalist challenge* (pp. 297–317). Berlin: Walter de Gruyter.

Horgan, T., Tienson, J., & Graham, G. (2007). Consciousness and intentionality. In S. Schneider & M. Velmans (Eds.), *The blackwell companion to consciousness* (pp. 468–484). Oxford: Blackwell.

Horgan, T., Tienson, J., & Graham, G. (in press). Phenomenology, intentionality, and the unity of mind. In A. Beckermann & B. McLaughlin (Eds.), The oxford handbook of philosophy of mind. Oxford: Oxford University Press.

Kane, R. (1996). The significance of free will. Oxford: Oxford University Press.

Kim, J. (2005). Physicalism, or something near enough. Princeton: Princeton University Press.

Maslen, C. (2005). A new cure for epiphobia: A context-sensitive account of causal relevance. *Southern Journal of Philosophy, 43*, 131–146.

Maslen, C., Horgan, T., & Habermann, H. (forthcoming). Mental causation. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), The Oxford handbook of causation. Oxford: Oxford University Press.

Menzies, P. (2003). The causal efficacy of mental states. In S. Walter & H. D. Heckmann (Eds.), *Physicalism and mental causation: The metaphysics of mind and action* (pp. 195–223). Exeter: Imprint Academic.

O'Connor, T. (2000). Persons and causes: The metaphysics of free will. Oxford: Oxford University Press.

Stephens, G. L., & Graham, G. (2000). When self-consciousness breaks: Alien voices and inserted thoughts. Cambridge MA: MIT Press.