PSYCHE

The Problem of Explaining Phenomenal Selfhood: A Comment on Thomas Metzinger's Self-Model Theory of Subjectivity

Kenneth Einar Himma
Department of Philosophy
Seattle Pacific University
3307 Third Avenue West
Seattle, WA 98119
U.S.A.

himma@u.washington.edu

PSYCHE 11 (5), June 2005

KEYWORDS: Consciousness, self, subject, subjectivity, phenomenal selfhood, dualism, Metzinger, physicalism, functionalism

COMMENTARY ON: Metzinger, T. (2003) *Being No One. The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press xii + 699pp. ISBN: 0-262-13417-9.

ABSTRACT: Thomas Metzinger argues that phenomenal selves *are* appearances produced by the ongoing operations of a "self-model" that simulates, emulates, and represents aspects of the system's states to itself – and not substantial things. In this essay, I explain the nature of phenomenal selfhood and then describe the most important problem that arises in connection with explaining phenomenal selfhood. I then argue that, by itself, the self-model theory of subjectivity lacks sufficient resources to wholly solve this problem and that Metzinger's argument does not justify his ontological conclusions about selves.

In his remarkable book *Being No One*, Thomas Metzinger defends a representationalist and functionalist analysis of the first-person phenomenal experience of being a self.[1] According to Metzinger, the phenomenal self—i.e., the experience of oneself as a conscious subject with a first person perspective—*is* no more than an appearance

produced by the ongoing operations of a complicated information-processing system that simulates, emulates, and represents aspects of the system's states to itself. Phenomenal selves are not substantial things at all on this view; while it is quite natural that we think of our selves as being real substances of some kind, selves are merely representational appearances that result from ongoing computational processes in the brain that satisfy certain conditions and produce what Metzinger terms a "self-model."

In this essay, I wish to evaluate the self-model theory of subjectivity and the strikingly nuanced and detailed analysis offered by Metzinger in support of this theory. To this end, I will begin by explaining what I take to be the nature of phenomenal selfhood; on this analysis, phenomenal selfhood is itself a pre-reflective element of every conscious experience. Next, I describe what I take to be the most important problem that arises in connection with explaining phenomenal selfhood—namely, the problem of explaining how it is that a particular phenomenal self (e.g., me) is associated with a specific set of neurophysiological processes (e.g., the processes that create a self-model in a particular living organism). I distinguish a "hard" and an "easy" issue associated with this problem.

I then attempt to evaluate the self-model theory of subjectivity and argue that Metzinger's theory falls short in a couple of important respects. First, I argue that, by itself, the self-model theory of subjectivity lacks sufficient resources to wholly ground a solution to either the hard or easy problems of phenomenal selfhood (or subjectivity). Second, I argue that Metzinger's theory fails to justify the conclusion that the furniture of the world does not include *substantial* selves.

None of this, however, should be construed as in any way disparaging the value or importance of this truly groundbreaking work. Although an explanation of phenomenal selfhood seems fundamental to an explanation of consciousness, philosophers of mind have devoted comparatively little space to explaining self, focusing instead on problems that presuppose it has already been explained.[2] To my knowledge, Metzinger provides the first comprehensive attempt to articulate and solve the problems associated with explaining the self and produces an analysis that is deep, detailed, nuanced, challenging, and nearly exhaustive in scope. That Metzinger's framework enables us to make sense of many pathological conditions which have eluded traditional theories and frameworks provides a compelling reason, on my view, to think that it will be an important part of understanding phenomenal selfhood and of solving the hard and easy problems of selfhood—even if, by itself, it cannot fully ground such solutions.

But regardless of whether I am correct in thinking Metzinger's work falls short in these respects, Metzinger's work is the state of the art on the topic and provides a standard of excellence that few of us will ever meet. Simply put, *Being No One* is an outstanding philosophical achievement.

1. The nature of phenomenal selfhood

We are conscious subjects with conscious mental states and the two seem related in a conceptually intimate way. It is hard to imagine that something could have a conscious mental state without being a conscious mental subject. The idea that there are, so to speak, free-floating mental states not instantiated by some mental subject seems conceptually incoherent: it seems clear that it is not conceptually possible for a conscious

mental state to occur that is not instantiated by a mental subject. Conscious mental states or events, as a conceptual matter, happen to (or include) conscious subjects or "phenomenal selves."

I think it is also uncontroversial that we have a conscious sense of being phenomenal selves that function as mental subjects. I am always there *qua* phenomenal self in every conscious perception and experience that I have—and this is part of what I experience. For example, my conscious experience of a sunset includes, as partly constitutive of the experience, its happening to me *qua* phenomenal self. It is quite natural, then, to follow Honderich (1995) in thinking that all our conscious mental states have two parts: a "content-part" and a "subject-part." On this characterization, being a particular phenomenal self (which is constituted by the subject-part) and having a particular content are both necessary constituents of any conscious experience.

While "subject" refers to a familiar part of experience, its character and role in conscious mental states are not easy to describe. One property frequently associated with the subject-part of mental states is the property of *mineness*. As Metzinger aptly describes it:

> What justifies treating all these highly diverse kinds of … phenomenal representational content as belonging to *one* entity … [is] the property of *mineness.* Mineness is a property of *a particular form of phenomenal content* that, in our own case, is introspectively accessible on the level of inner attention as well as on the level of self-directed cognition.… Here are some typical examples of how we, linguistically, refer to this particular higher-order phenomenal quality in folk-psychological contexts: "I experience *my* leg subjectively as always having belonged to me" (Metzinger 2003: 302; emphasis added).

On this altogether intuitive characterization, the subject-part of my experience either confers upon the content-part of my experience a felt sense of belonging to me (that is, an experienced sense of being mine that is immediately and pre-reflectively accessible) or constitutes this content-part as belonging to me.

It is important to note that while talk of the property of mineness is quite helpful in identifying this particular feature of consciousness, it does not fully describe the subject-part of mental experience as typically experienced. While I certainly experience a felt sense of ownership over the contents of my mental experiences, I also experience something more basic than that. The contents of my experience include a felt sense of a *me* that is the subject (or bearer) of those contents. As a phenomenological matter, it is hard for me to even imagine what it would be like to have an experience of something's being mine without simultaneously having an experience of a me to whom that something belongs as mine.

Moreover, the idea of a felt sense of mineness without a felt sense of me-ness seems suspect from a logical standpoint. The property of mineness is a relational property that obtains between two relata. The idea that some entity or property is mine presupposes a *me* to whom that entity or property belongs—and this is no less true of

mental content than of any other entity or property: the claim that mental content is mine presupposes a me to whom that content is appropriately related. Thus, it is not clear that it is even logically possible for someone to experience a sense of mineness without also experiencing a sense of me-ness. If so, then subjectivity cannot be explained as just a felt sense of being mine because it logically presupposes the more basic felt sense (or phenomenal experience) of being a *me*.

Indeed, it seems that the relationship between the subject-part of my experiences and me is considerably more intimate than conveyed by the idea that the subject-part confers just a sense of mineness on the content-part. It is not—and could not be—just that the subject-part of an experience confers a sense of mineness upon the content-part of the experience; it is rather that the subject-part of an experience is what I take to be *me qua* bearer of conscious mental states. At any moment in time, I take myself *qua* phenomenal self to be identical with a subject-part that seems to remain constant throughout the changes that occur over time in the content-part of my experience. In this sense of the term, the phenomenal self that accompanies this body is identical with the subject-part that is present in all my experiences.

None of this, however, should be construed as implying any substantive conclusions about the nature of the self or subjectivity; though the term "phenomenal self" functions grammatically as a noun and hence purports to designate entities of some kind, this term, as I use it, carries no ontological baggage whatsoever. It is intended to do no more than pick out a particular element of experience. While it is consistent with the claim that the subject-part of an experience is, comprises, or includes a substantial Cartesian ego capable of existing independent of the body, it is also consistent with Metzinger's view that the self is nothing more than a phenomenal appearance. To use this term to pick out the subject-part of a conscious mental state is no more theoretically loaded than to use the term "content" to pick out the content-part of a conscious mental state. As I use the term, then, "phenomenal self" is consistent with both a dualist account of the self as substantial soul and with any particular physicalist account, including Metzinger's, that denies the ontological independence of minds from bodies and explains consciousness in terms of the causal properties of physical entities, states, and processes.[3]

2.The problem of explaining phenomenal selves

Physicalism comprises an ontological theory and a theory of mind. As an ontological theory, physicalism asserts that physical entities are the only substances in the world. Since there are thus no ontologically independent entities that are essentially incorporeal or immaterial, it follows that human beings are entirely physical in nature and composition. But, as a theory of mind, physicalism holds that all mental states, properties, and processes can fully be explained in terms of the causal properties of neurophysiological states, properties, and processes (even if such states turn out to be nothing over and above neurophysiological states). Every fact about human consciousness, then, can be explained entirely in terms of physical facts and causal laws.

Given that phenomenal selfhood (i.e., conscious subjectivity) is a crucial aspect of consciousness, the physicalist must give some sort of explanation for the existence (or emergence) of the phenomenal self. Not surprisingly, this is an assumption that is uncontroversial among empirical researchers in consciousness, who uniformly regard

subjectivity or selfhood as something that requires neurophysiological explanation. As neuroscientists Josef Parvizi and Antonio Damasio put this important point:

> [There are] two closely related but separable problems in the investigation of consciousness. The first is the problem of understanding how the brain engenders the mental patterns we experience as the images of an object.… The second problem of consciousness concerns how, in parallel with creating mental patterns for an object, the brain also creates a sense of self in the act of knowing. The solution for this second problem requires the understanding of how each of us has a sense of "me"; of how we sense that the images in our minds are shaped in our particular perspective and belong to our individual organism (Parvizi and Damasio, 2001: 136-137).[4]

As Parvizi and Damasio conceive the problem, consciousness cannot be fully explained without a causal explanation for both the content-part (i.e., the "image of an object") and the subject-part (i.e., the "sense of 'me'") that, as a conceptual matter, constitute a conscious mental state.

But merely identifying the neurophysiological processes, functions, states, and operations that produce the phenomenal self cannot fully explain phenomenal selfhood. Showing that subjectivity supervenes upon these processes leaves unanswered a "hard" problem of subjectivity that requires an explanation of *how* it is that these processes produce this particular element of conscious experience. A showing that phenomenal selfhood is associated with a particular set of processes is an important step towards providing a full physicalistic explanation for phenomenal selfhood, but it is nonetheless solves only an "easy" problem of consciousness.

As it turns out, the problem of providing a full physicalistic explanation for phenomenal selfhood requires solving a much more intimate philosophical problem—one that is "hard" in character. As Thomas Nagel describes this problem:

> It isn't easy to absorb the fact that I am contained in the world at all. It seems outlandish that the centerless universe, in all its spatiotemporal immensity, should have produced me, of all people—and produced me by producing TN [i.e., Thomas Nagel]. There was no such thing as me for ages, but with the formation of a particular physical organism at a particular place and time, suddenly there *is* me, for as long as the organism survives. In the objective flow of the cosmos this subjectively (to me!) stupendous event produces hardly a ripple. How can the existence of one member of the species have this remarkable consequence (Nagel, 1989: 55)?

If, as the physicalist believes, a complete explanation of consciousness can be given entirely in terms of the causal properties of physical objects, events, and processes, then it follows that a physicalistic explanation of phenomenal selfhood, which is a particular phenomenal element of conscious mental experience, constitutes a physicalistic

explanation of how a particular physical body causally gives rise to a particular mental subject.

One qualification is needed here. The assumption that there is a real problem here presupposes the falsity of both identity theory and eliminativist materialism.[5] Insofar as the statement of the problem distinguishes two entities, mental subject and body (which would include any neurophysiological correlates of subjectivity), it presupposes that there is something to which each refers—an assumption denied by elminativists; as far as eliminativists are concerned, there is no problem of subjectivity at all—because there are no such things as subjects. Moreover, if some version of identity theory is true, then the mental state of being a subject is nothing over and above the relevant neurophysiological state. Since *qua* subject, I am nothing more than the relevant brain state, there is no more to be said about explaining how my body brings me into existence than there is to be said about how collecting the numbers 1, 2, and 3 brings the set {1, 2, 3} into existence—because the two expressions involved in describing the problems refer to the very same thing. On the assumption that eliminativism or identity theory is true, the Nagel passage fails to state a genuine problem.[6]

On any other physicalist view, however, the problem of explaining phenomenal selfhood is equivalent to the problem of explaining how a particular body produces a particular mental subject.[7] Giving a full physicalistic explanation of phenomenal selfhood requires explaining how it is that a particular set of neurological operations instantiated by a particular body causally gives rise to a particular phenomenal self. Put in the first-person singular terms favored by Nagel, the problem is to explain how it is that the particular body that was born at a particular set of points in space-time (i.e., the first one born to my mother) brings *me* into existence as a phenomenal self—and not someone else. To be successful, a full physicalist account of selfhood, then, must explain how the set of mereological simples arranged in the form of my body brings *me* into existence *qua* phenomenal self.

It is crucial to note here that Metzinger is neither an eliminativist nor an identity theorist. Most obviously, Metzinger's conception of the self as "a phenomenal appearance" and "felt sense of mineness" is inconsistent with the eliminativist view that there is nothing to which phenomenological terms refer. Further, while Metzinger sometimes suggests that the self is nothing over and above a functioning "self-model," his phenomenological analysis seems inconsistent with the identity theorist's view that there is no ontological distance between mental states and physical states.[8] For this reason, he is most plausibly construed as taking the position that the self is nothing but a phenomenal appearance that is the causal result of the relevant neurophysiological functional processes—which presupposes that the appearance is ontologically distinct from those processes (though not that the appearance is a substantial entity of some kind). If this is correct, the Nagel passage above states a problem that Metzinger's theory must address in some way to be wholly successful.

It is important to be clear about the character of this problem. One might be tempted to think that, even on other physicalist theories, it requires nothing more than an explanation of why I am identical with myself and am not identical with you. Thus conceived, the problem of explaining phenomenal selves is no more a problem than the problem of explaining why any particular thing is what it is and not something else.

Interpreted this way, the problem is simply to explain why the following formula is necessarily true: (x)[(x = x) & (y)(x ≠ y _ x ≠ y)]]. If, on this line of interpretation, this is a problem, it is a problem that afflicts every theory—and not just a physicalist theory of mind.

But the demand for an explanation of how a particular collection of atomic, subatomic, or molecular material arranged body-wise gives rise to *me qua* phenomenal self is not a demand for an explanation of the trivial claim that I am me and not someone else. Phenomenal selfhood is a particular element of conscious experience that has a felt quality; as such, it is as much in need of theoretical explanation as any other element of conscious experience. Indeed, Nagel's articulation of the problem suggests that it is conceivable that the phenomenal self associated with my body might very well have been associated with another body (so that I might have been the subject associated with a stream of content-parts produced by that other body)[9]—which, of course, is not true of the claim I might not have been identical with myself. It should be clear, then, that the problem of explaining how it is that my self is associated with a particular body is a problem quite different from the problem of explaining why any particular logical truth is true and not false.

3. Metzinger's physicalist explanation of phenomenal selfhood

Metzinger's view that the phenomenal self *is* nothing more than the ongoing operations of a complicated information-processing system is grounded in two conceptual entities that, taken together, provide a model of subjective phenomenal experience. The first is the phenomenal self-model (PSM), which incorporates "the content of the conscious self: your current bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processing" (Metzinger 2003: 299). According to Metzinger, a PSM comprises a number of computational processes that make system-related information (e.g., information obtained from the sense organs) available in an integrated form. The PSM is a self-*model* in that its operations *simulate* and *emulate* abstract properties and states of its own internal information processing. It is a *self*-model in the sense that it performs these functional operations *for itself* and represents their outputs *to itself*. Otherwise put, the subject and object of the PSM are the same.

The second requisite conceptual entity is the phenomenal model of the intentionality relation (PMIR), which provides a functionalist model of the experienced subject-object relation that forms the basis for the perspectival dimension of self. A PMIR depicts a relationship between the system, which is transparently represented to itself, and some (possibly internal) object in the world. For example, the PMIR currently operative in your body would depict, among other things, your state of being someone who is currently reading a critical analysis of Metzinger's views on phenomenal selfhood. PMIRs are usefully thought of as arrows pointing from self-model to the object component.

Both conceptual entities are necessary to fully model consciousness. According to Metzinger:

> Full-blown conscious experience is more than the existence of a conscious self [which is modeled by the PSM], and it is much more than the mere presence of a world. It results from the dynamic interplay between this self and the world, in a lived, embodied present (Metzinger 2003: 417).

Thus, while the instantiation of a PSM "forms the central necessary condition for a conscious *first-person perspective* to emerge on the representational as well as on the functional level of description" (Metzinger 2003: 299), it is not sufficient: it is "the existence of the PMIR [that] generates full-blown consciousness" (Metzinger 2003: 417). Full-blown consciousness, Metzinger concludes, requires "the generation of a world-model, the generation of a self-model, and the transient integration of certain aspects of the world-model *with* the self model" (Metzinger 2003: 427).

Metzinger's analysis provides a powerful framework for understanding the functional and representational characteristics of both normal and pathological subjective experience. Consider, for example, how this framework contributes to explaining the condition of patients who, despite showing all the functional signs of having lost their sight, continue to insist that they can see:

> Under the present theoretical model, there are two possible routes of interpretation.… [T]he object component of the second-order, cognitive phenomenal model of the intentionality relation (PMIR) (in this case, the transparent model of oneself *as a person no longer seeing*) [could] simply [be] absent. Information concerning the deficit simply does not exist. This could happen when it is impossible for the post lesional brain to *update* its phenomenal self-model.… [Or] there could exist an updated self-model in the patient's brain, but this new model could functionally not be *globally available for attention*. Deficit-related information would then be active within the system as a whole, but it could never become subjective information, because, for functional reasons, it cannot be represented under a PMIR (Metzinger 2003: 430).

Whereas such cases seem impossible to reconcile with traditional frameworks that presuppose one cannot be mistaken about the contents of one's mind, they are easily and elegantly explained within Metzinger's framework. Metzinger's models thus define an analytical framework that can be reconciled with various conditions that undermine traditional frameworks.

Metzinger takes his analysis to provide further support for physicalism. On his view, the conceptualization of the self as a system that instantiates a PSM and PMIR is sufficient to warrant an ontological claim about the status of selves: "The phenomenal property of selfhood as such is a representational construct; it truly is a *phenomenal* property in terms of being an appearance only" (Metzinger 2003: 563).[10] The property of "mineness" that unifies the various elements of conscious experience as the experience of a single self is itself nothing more than a phenomenal appearance.

If Metzinger is correct, his analysis provides an additional reason to reject the dualist view that every person is a causal composite of one mental substance (a substantial self that is usually called a "soul") and one bodily substance. If selves are purely phenomenal in the sense of being no more than "appearances," then it follows that "no such things as selves exist in the world" (Metzinger 2003: 563). Selves and subjects are simply the insubstantial outcomes of these processes and hence are not *substantial* entities, like souls or substantial minds capable of existing independently of bodies, that would count as part of the furniture of the world. A self would be simply be a phenomenal representation and, as such, an insubstantial appearance, but not the sort of substantial entity, like an atom, that would count as a real entity properly included in an ontological inventory of what there is in the world.

4. Critique of the Self-Model Theory of Subjectivity

In this section of the essay, I argue that Metzinger's nuanced and insightful analysis, despite its obvious merits, is problematic in certain critical respects. First, I argue that, while the self-model theory is a plausible piece to explaining phenomenal selfhood, it is, at best, an incomplete explanation; by itself, the self-model theory lacks sufficient resources to fully ground a solution to either the easy or hard problem of subjectivity. Second, I argue that it fails to refute dualism. In both cases, then, more is needed to do the work that Metzinger believes his theory can do.

4.1 The Easy and Hard Problems of Subjectivity

At the outset, it is worth noting that Metzinger's framework is limited with respect to its explanatory power in one important respect. No theory of the self that ultimately explains the existence of self in terms of models that emerge from various computational processes can be fully successful without identifying the neural correlates of the various processes. And though he believes the neural correlates of these models will be identified at some point, Metzinger concedes, as he must, that "not much is presently known about the neural underpinnings of the transparent self-model in humans" (Metzinger 2003: 340).

This, of course, should not be taken as a criticism of Metzinger's analysis. The problems of subjectivity are sufficiently complex that a full solution to these problems will evolve gradually in discrete pieces. Some of these pieces will be conceptual; some will be philosophical; and some will be empirically grounded in observations about neurophysiological states, operations, functions, and processes. The fact that Metzinger does not identify the neural coordinates of the self-model, by itself, should not be construed as a weakness in his remarkably detailed and sophisticated analysis.

Indeed, Metzinger's analysis is plausibly construed as providing a conceptual framework that defines an adequacy constraint on neurophysiological explanations of subjectivity. Thus construed, the idea is that any successful neurophysiological explanation of subjectivity must include all the necessary elements to play the functional and representational roles played by a self-model as Metzinger describes it. Insofar as a neurophysiological explanation lacks sufficient resources to give rise to a PSM and PMIR, it falls short as a physicalistic explanation of conscious subjectivity.

But if Metzinger's analysis succeeds in providing a functionalist analysis of the role of phenomenal subjectivity, it nonetheless lacks sufficient resources to fully ground a physicalist account of subjectivity. To see this, it would be helpful to consider a twin-earth thought experiment. Suppose that the earth has a perfect twin that is distinguishable from earth only in terms of spatial and physical properties not known to anyone on either planet. In this world, then, you have a twin who, though composed of different material than you, is in every other known respect, mentally and physiologically, indistinguishable from you. You and your twin are genetically indistinguishable at every moment in time. The two of you are physiologically isomorphic in the following respect: at every moment in time, the two of you have materially distinct but otherwise qualitatively indistinguishable atoms and molecules arranged according to the same blueprint, and those materials are always in exactly similar physical states. Thus, for example, your respective brains and brain states are spatially distinct but otherwise qualitatively indistinguishable at every moment in their lives.

Likewise, you and your twin's mental states and characteristics track each other at every moment in your lives. You and your twin are exposed to exactly similar—though obviously not the *same*—sensory input at all times, and your brains respond to this input in qualitatively indistinguishable ways.[11] You and your twin have exactly similar long- and short-term memories at every moment in time. You and your twin have exactly similar personality and emotional characteristics at every moment of your lives. Indeed, even if we assume that we have libertarian free will, you and your twin will always instantiate exactly similar volitions at the same moments in time.[12] And all this is true regardless of whether these mental states, events, and characteristics supervene upon physical states, events, and characteristics—though it is quite reasonable, of course, to think that they do.

You and your twin, then, are mentally and physiologically indistinguishable at every level of description (again, apart from the spatially distinct materials that compose your bodies)—including a description of the self-models of you and your twin. At every moment in time, your PSM and PMIR are exactly similar to those of your twin in every respect that Metzinger believes is causally relevant to explaining your phenomenal selves. From the standpoint of the self-model theory, then, you and your twin are utterly indistinguishable at every moment in time.

Nevertheless, there remains one crucial difference between you and your twin: one of these phenomenal selves is *you* and the other is not. You are the phenomenal self associated with a stream of experience that arises from one of these two perfectly similar bodies with perfectly similar histories and not the other. That is, the phenomenal self that you identify as *you* is paired with the phenomenal content that arises from one of these bodies and not the other.

From the standpoint of the self-model theory, it is utterly arbitrary that *you* (or your phenomenal self) are the subject of a stream of content brought about by physical stimulation of one of those bodies and not the other. Since the constituents of your self-model are perfectly isomorphic to the constituents of your twin's self-model, you and your twin are indistinguishable with respect to *every property and operation causally relevant under Metzinger's analysis* and hence should be indistinguishable with respect to every phenomenal feature that is the causal outcome of those properties and operations.

Thus, if you and your twin are perfectly similar in respect of all causally relevant properties and operations, it is completely arbitrary that *you* are the phenomenal self associated with one of these self-models rather than the phenomenal self associated with the other.

And this is true even if the self is nothing more than a phenomenal appearance. As long as there is some ontological distance between these elements of experience and the underlying brain processes, the association of your self with the self-model produced by one of these qualitatively indistinguishable bodies has no properly physicalistic explanation. Even if your self is conceived as nothing more than a phenomenal appearance (or felt quality), it is utterly arbitrary that phenomenal appearance that you experience as being you is associated with the self-model produced by the functioning of one of these bodies instead of the other.[13]

Indeed, the self-model theory, by itself, lacks sufficient resources to fully ground a solution to even the easy problem of subjectivity. If you and your twin agree on all properties relevant in causally explaining consciousness but disagree in some way with respect to the subject-part of your experiences, then those properties *cannot* ground a causal explanation for the subject-parts of *either* of your experiences. That is, those properties cannot ground a causal explanation of why one body/model, rather than the other, gives rise to you—instead of someone else. Accordingly, the twin-earth thought experiment seems to show that a neurophysiological explanation of subjectivity will require more than identifying the neural coordinates of the relevant models comprising phenomenal selfhood. If this is correct, the self-model theory of subjectivity cannot fully ground a solution to the easy problem of subjectivity.

In response, one might argue that, by definition, you are the phenomenal self (or appearance of self) associated with the self-model produced by your body and your twin is the phenomenal self (or appearance of self) associated with the self-model produced by your twin's body. On this line of response, the twin-earth thought experiment poses no deep problem for the self-model theory because there is simply no other logically possible outcome. By definition, for all phenomenal selves (or appearances of selves) *A*, *A* is the phenomenal self (or appearance of self) associated with *A*'s body. Thus, it is simply trivially true that you are the phenomenal self (or appearance) associated with your body and your twin is the phenomenal self (or appearance) associated with your twin's body.[14]

This, however, misunderstands the difficulty that the twin-earth thought experiment poses for the self-model theory. The point here is not to demand an explanation for the claim that, for any phenomenal self *A*, *A* is the phenomenal self associated with *A*'s body. I think it is fair to say that this claim is necessarily true—though the nature of the modality is not entirely clear to me.[15]

Rather, the point is to demand an explanation for the fact that *you* are the self brought about by neurophysiological states defining your self-model instead of by the perfectly similar neurophysiological states defining your twin's self-model. In other words, the issue, as Nagel might describe it above, is why one of these self-models is *yours* while another perfectly similar self-model is someone else's. Since the two models and corresponding neurophysiological states and operations are physically and

nomologically indistinguishable at every relevant level of description, it is completely arbitrary from the standpoint of the self-model theory that one of these self-models is (or produces) *you* while the other is someone else.

Here it is absolutely crucial to note that the difference between you and your twin's phenomenal self is not simply a difference in number; if ordinary intuitions are any indication, the difference between your two selves involves a difference that theories of mind are obligated to explain. The termination, by one or another means, of my phenomenal self means that *I* (in the most intimate sense) no longer exist as a conscious mental subject and cannot instantiate any mental content; in the sense most meaningful to me, the end of my phenomenal self results in my *death*.[16] But the termination of my twin's phenomenal self does not have this unhappy result for me: while the termination of my twin's phenomenal self results in my twin's death, it has nothing to do with my own continuing sentient existence.

Accordingly, if ordinary intuitions are correct, the difference between my phenomenal self and my twin's is of tremendous significance. Despite the fact that my twin and I have qualitatively indistinguishable self-models, my continuing existence as a conscious subject of experience—as a locus of awareness—depends on, so to speak, the survival of one particular phenomenal self. The fact that my twin's self and my self's conscious *mental* lives are different in this important *phenomenal* respect but have indistinguishable self-models is, I think, a serious problem for the self-model theory of subjectivity.[17]

Of course, if the self-model theory is correct, then these ordinary intuitions about the self are false—something that Metzinger also believes. But it will not suffice to simply assert that the self-model theory is correct because this simply begs the question against those very intuitions.[18] What is needed is a good independent reason to think that these common intuitions are false. As far as I can tell, there is nothing in the admittedly remarkable analysis of *Being No One* that would provide a non-question-begging reason to reject these very common, and quite stubborn, intuitions.

What this argument purports to show, then, is this: phenomenal selves (construed as conscious subjects of experience) cannot be explained by any physicalist theory that goes no further than identifying the neural coordinates of the PSM and PMIR. The twin-earth thought experiment shows that it is not the case that your phenomenal self is causally explained by a full description of the neural coordinates of the self-model theory of subjectivity. As far as the laws of nature are concerned, it could have been the case that you were the phenomenal self associated with the neural coordinates of your twin's self-model and that your twin was the phenomenal self associated with the neural coordinates of your self-model.[19] If this is correct, then Metzinger's analysis cannot ground a solution to either the hard or easy problem of explaining the existence of phenomenal selves.

4.2 Dualism and the Self-Model Theory of Subjectivity

There are two versions of the dualist view that human persons are a causal composite of two analytically distinct kinds of substance: bodily substances and mental substances (or souls). The first, classical dualism, denies that the existence of souls necessarily depends on the existence of physical bodies; on this line of analysis, souls are capable in principle

of existing independently of bodies. The second, emergentist dualism, holds that minds are mental substances that "emerge" from neurophysiological processes and operations.[20]

As noted above, Metzinger believes that his self-model theory shows that no form of dualism is true, but his theory lacks sufficient resources to support this conclusion. If the self-model theory could fully explain how the phenomenal self is wholly produced by the ongoing operations of a self-model, that would be a decisive reason to reject classical dualism. After all, if the hypothesis that mental substances exist independently of physical substances is not needed to explain phenomenal selfhood (which is presumably the one job that the dualist hypothesis must do if classical dualism is correct), then it can be rejected (under Ockham's Razor) as explanatorily superfluous.[21] Thus, if the self-model theory could solve the hard and easy problems associated with explaining phenomenal selves, then it would justify the rejection of classical dualism.

But, as we saw in the last section, the self-model theory solves neither problem. First, even if the neural correlates for every one of my conscious states could be identified and described, this simply provides a map from the set of my brain states to the set of my conscious states. Such a map cannot explain why these brain states bring *my* phenomenal self, rather than some other, into existence. At bottom, the hard ontological problem of explaining selves is a philosophical problem—and not an empirical problem.

Second, the twin-earth thought experiment shows that phenomenal selfhood cannot be mapped onto any particular set of neural coordinates that would satisfy the self-model theory; since you and your twin agree, at every moment, on all possible sets of neural coordinates but have phenomenally distinct (as opposed to just numerically distinct) selves, it follows that phenomenal selfhood does not supervene on any neurophysiological processes that would satisfy the self-model theory.[22] By itself, the self-model theory cannot solve either problem and hence does not justify any sort of sweeping ontological conclusion about the existence of selves.

To buttress his case against dualism, Metzinger takes on one of its most influential arguments. On this first-person line of analysis, I could have been someone else in the sense that the self that is the subject of my experiences could have been paired with some other body. If so, then it is false that phenomenal selves are created by neurophysiological processes and hence physicalism is false.

In response, Metzinger argues that the "contingency intuition [that I could have been someone else] is not even based on a phenomenal possibility" (Metzinger 2003: 597). He argues that I can imagine, for example, being Immanuel Kant only in a limited sense: the best you can do is "phenomenally simulate him as subject (Metzinger 2003: 597). This, as I understand it, is based on two reasons. First, since selves are "fictitious" entities, there is no way to perform the thought experiment of being a different self. Second, if selves are more accurately characterized in terms of self-models, the "subject component remains opaque" (Metzinger 2003: 596).

These arguments do not add enough to the self-model theory to justify thinking that the ontology of the world does not include selves. For starters, both of the reasons offered by Metzinger beg the question against the dualist by assuming that the self-model theory is true and shows that selves are "fictitious" entities. More importantly, it is false, as far as I can tell, that we cannot perform the thought experiment of being Immanuel

Kant. To imagine being Immanuel Kant, I simply have to be able to imagine that the "particular form of phenomenal content" (Metzinger 2003: 302) that is the subject of my mental experience is paired with the stream of mental content associated with Kant's body and brain. If it is false that I could have been Immanuel Kant, it will be for nomological reasons—and not because the thought experiment is incoherent and fails to describe a *logical* possibility.[23]

But, as it turns out, the twin-earth thought experiment suggests it is not just logically possible that I had been Immanuel Kant; it is also *nomologically* possible. The twin-earth experiment seems to show that, as far as the laws of nature are concerned, my phenomenal self might have been associated with another body: if it is utterly arbitrary as far as the laws of nature are concerned that my phenomenal self[24] is associated with my body rather than my twin's, then, as far as those laws are concerned, my phenomenal self might very well have been associated with my twin's body.

And if that is true of my twin's body, then it is also true of some very different body: as far as the laws of nature are concerned, my phenomenal self might have been associated with a very different body with very different experiences and properties—say, Kant's. In that case, of course, my memories, personality traits, tastes and preference would all be different, making me a different "person" in the sense relevant for the theory of personal identity. But I *qua* phenomenal self would be the subject of a stream of experience that arises from the operation of that other body. In that nomologically possible world, the phenomenal self that is the subject of experiences arising from my body in the actual world would be the subject of experiences arising from that other body. There is no non-question begging reason to reject this as a coherent possibility.

Part of the problem here is that the self-model theory of subjectivity is conceptual in nature. Although Metzinger develops his models with an eye towards various empirical phenomena, his methodology is largely conceptual. The models he provides are, on his own characterization, theoretical entities that "may form the decisive conceptual link between first-person and third-person approaches to the conscious mind" (Metzinger 2003: 9).[25] The various models and the self-model theory to which they give rise are the fruits of a methodology that is self-consciously conceptual in character.

But one cannot solve substantive ontological problems just by doing conceptual analysis. For example, the fact that we call a particular arrangement of mereological simples arranged in the form of a chair "an object" does not imply that the ontology of the world includes chairs in addition to the mereological simples arranged in the form of chairs.[26] The issue of whether the world includes composite material objects like chairs is a deep and difficult philosophical issue that cannot be solved just by moving concepts around. Whether there *really are* chairs in the world does not depend in any simple way on our conceptual practices with respect to words like "chairs" and "objects."

Nor has conceptual analysis solved many substantive problems in the philosophy of mind. Physicalists, for example, are no closer to understanding how mental states cause physical states in virtue of having rejected the dualist claim that mind is a substance.[27] The only conceptual theories that, by themselves, would solve the mind-body problem do so at the cost of falsifying much ordinary talk about mental states: the

identity theory, for example, "solves" the problem of how mental states cause physical states by conceptually identifying the two (mental states *are* brain states), but renders problematic much of what we commonly predicate of mental states (e.g., the property of being pleasant is not sensibly attributed to brain states). Similarly, eliminative materialism "solves" the problem by simply denying that we have mental states, which amounts to a surrender and not a solution of any sort. As long as we conceptualize mental states as non-spatial and non-extended, we will face prohibitive conceptual difficulties explaining how such states can cause brain states that are spatial and extended. Denying substance dualism, which is partly a conceptual move (i.e., mental entities are "states" but not "substances"), does nothing to solve difficult problems like this.

4.3. Avoiding Substance Dualism

None of the foregoing, of course, should be thought to imply that substance dualism is true. Both Nagel's statement of the problem and my analysis of the twin-earth example assume, as noted above, that eliminative materialism and the various versions of identity theory are false. Moreover, it might very well be that, as a matter of brute fact, there is a certain randomness involved in the generation of these phenomenal appearances (typically referred to as selves). The claim that phenomenal selfhood cannot be mapped onto any particular set of neural coordinates that would satisfy the self-model theory does not imply that selves must be substantial entities.

As a logical matter, this means that, on the assumption that my argument succeeds in doing what I think, Metzinger can avoid dualism in one of at least two ways. First, he could simply adopt some version of identity theory. While I find each of these views sufficiently unattractive from the standpoint of what seems to me obviously true about my mental experience to think that substance dualism, with all its difficulties, is a far more appealing approach, there are a number of philosophers who would certainly disagree. Though I suspect that identity theories do not command a majority of physicalists, many do find these views quite congenial.

Second, he could simply take the position that phenomenal selfhood is, at least in part, a brute fact about consciousness that cannot be explained in terms of nomological laws. By my lights, this is not a particularly attractive option because, at least at this point in the game, it is as poorly motivated as eliminativism: it is more a surrender than anything else. But it is a logically viable move.

There might be one more way to avoid dualism. Earlier, I argued that phenomenal selfhood is more than a felt sense of *mineness*; it is a felt sense of *me-ness* that is the subject (or bearer) of all other conscious mental states. Moreover, I suggested that the loss of this felt sense of me-ness results in the extinction of the particular subject it defines; if, for example, this felt-sense in me were to be extinguished, I would no longer be the subject of any conscious mental states. The experience of being a self, then, involves more than just an experienced sense of ownership over the contents of one's perceptions, thoughts, and so on; it involves an experienced sense of being a subject of those states – and is identical with that sense.

The idea that my existence *qua* conscious subject needs philosophical explanation seems to depend on this configuration of views about phenomenal selfhood. If there is nothing more to phenomenal selfhood, then just some sort of free-floating sense of

ownership over the contents of conscious mental states, there is no element of experience that I could identify as being me *qua* subject. And if there is no such element, then there is no need to explain the existence of that element.

Accordingly, Metzinger could simply deny my views about selfhood. This does not strike me as a particularly plausible position partly because it is inconsistent with what seems obvious to me from every one of my conscious mental states and partly because the idea of a sense of mineness without a sense of me-ness strikes me as logically incoherent. But in the absence of a better argument than I have for these views, I have given little reason to think that it is not a logically coherent position.

While there are thus any number of ways to avoid dualist commitments here, the fact that at least one of these is needed means that the self-model theory of subjectivity and supporting arguments, by themselves, cannot do the work that Metzinger believes they do. In particular, there is just not enough here to justify the rejection of substance dualism, in part, because there is nothing here that would provide some decisive reason to think that any one of these maneuvers is correct. As far as I can tell, there is nothing in *Being No One* that provides a compelling reason to think that identity theory is true or that phenomenal selfhood is nothing more than sense of mineness.

Again, none of this should be construed as denying the importance or quality of Metzinger's work in *Being No One*. *Being No One* defines the state of the art on a difficult problem that has largely, and inexplicably, been neglected by philosophers of mind. Indeed, the problem of phenomenal selfhood is not implausibly characterized as the most basic problem in the philosophy of mind, as it seems clear that no conscious mental state can be fully explained without explaining the phenomenal self whose content it is. If Metzinger's self-model theory does not solve this problem, it represents a plausible starting point for doing so in the following sense: while no theory that does no more than explain subjectivity in terms of self-models can succeed, no theory that does not explain subjectivity in terms of such models can succeed either. For this reason, one can expect Metzinger's self-model theory of subjectivity to have a significant and lasting influence on efforts to solve the problems associated with explaining phenomenal selfhood. It deserves no less.

Notes

[1] Portions of this essay have appeared in a short review of Metzinger (2003) which appeared in *Metapsychology* (Christian Perring, editor); available from http://mentalhelp.net/books/. I am grateful to *Metapsychology* for granting permission to reproduce those portions here. I am also indebted to Timothy Bayne and Dorothée Legrand for their constructive criticism.

[2] While philosophers of mind and cognitive scientists focus more on explaining conscious content, one cannot fully explain how it is possible to be aware of content without explaining how it is possible to be aware; and this requires explaining how it is possible to be a conscious subject of that content. For this reason, the problem of explaining subjectivity (or phenomenal selves) is the more basic problem.

[3] Indeed, it is, strictly speaking, consistent with the eliminativist claim that there do not

exist any mental states and hence no selves as I have described them.

[4] It is worth noting that the experienced sense that Parvizi and Damasio take to be basic is "a sense of *me*" and not a just a sense of mineness.

[5] I am grateful to Dorothée Legrand for pointing this out to me.

[6] As it turns out, there are no so-called "hard problems" of consciousness on the assumption that either eliminativism or identity theory is true. The problem of explaining *how* neurophysiological states produce mental states presupposes that such entities are distinct—something that both identity theorists and eliminativists deny.

[7] This is not true of substance dualism. If, as even the substance dualist would have to concede, the brain plays some role in what I have called the subject-part of conscious mental experience, the substance dualist will have to solve the mind-body problem in order to fully explain the subject-part of mental experience. But doing so will not obviously tell us anything about how it is that a particular mind-substance is associated with a particular body.

[8] Metzinger also frequently speaks of these appearances being produced by the normal operations of the neural correlates of a self-model. If, as seems reasonable, one thing cannot cause itself, then it follows that the appearances are not identical with these operations or states.

[9] Again, this presupposes the falsity of both eliminativism and identity theory.

[10] Passages like this militate (decisively, on my view) against construing Metzinger as embracing any general form of identity theory.

[11] Since you are presumably billions of light years apart, you could not be exposed to, for example, exactly the same sun. But the sun to which you are exposed is qualitatively indistinguishable from the counterpart sun to which your twin is exposed.

[12] While this is unlikely if libertarianism is true, it is possible. The state of affairs in which two persons always agree in their volitions is not inconsistent with libertarianism.

[13] This is subject to one exception, which is discussed in Section 4.3 below.

[14] From here on, I will drop the awkward "(or appearance of self)." However, the term "self" should be construed to include the possibility that self is nothing more than a phenomenal appearance.

[15] My intuition is that the modality is conceptual. It seems reasonable to think that, as a conceptual matter, a physical object X is A's body if and only if A is the subject associated with X. That is, it is part of the very concept of some physical object X being A's body that A is the subject associated with X. It should be noted that the dualist intuition that a subject A could, so to speak, inhabit some other body than the one A actually inhabits is consistent with this conceptual claim. Let A be a subject, X be A's body, and Y be some body distinct from X. The relevant dualist claim, then, is that it is possible that X is not A's body and that Y is A's body. This is consistent with it being necessarily true that A is the subject of A's body; for the claim is just that it is not necessarily true that X (which is, in fact, A's body) is A's body.

[16] Indeed, insofar as one hopes for a life after death, one is hoping for the continuing

existence of one's existence *qua* subject of mental experience (and also for a continuing stream of enjoyable mental content).

[17] This seems to be a general problem with functionalist explanations of mental states. While it is reasonable to think that a full explanation of most, if not all, mental states will include functional elements, a purely functionalist analysis tends to leave the phenomenal elements of mental states under-explained. A functionalist analysis of pain, for example, will tell us much that is right about pain, but it cannot tell us anything about the felt hurtfulness of pain—which, on my view, is the element that is most in need of explanation. What ultimately causes the problem for Metzinger's view is that my and my twin's experiences of being selves are indistinguishable in terms of functions but are distinguishable in terms of phenomenal content.

Of course, not everyone agrees with the above assessment of functionalism. One might deny the existence of these felt qualities or deny that they need explaining. While there is no knock-down refutation of this position, the counterintuitive quality of these denials is, on my view, a very good reason to reject them. I am, for example, far more certain of my having experienced such felt qualities than I am about anything one could say in defense of the claim that these qualities do not exist. But there are many theorists who do not find such reasoning persuasive.

[18] Here is another way to develop these common intuitions. In a famous thought experiment, Derek Parfit asks whether you would survive being "transported" from Earth to Mars by a machine that (1) destroys your body on Earth; (2) sends to Mars a perfect blueprint of the arrangement of all the atomic constituents making up your body; and (3) creates a body on Mars out of new atomic materials that satisfies perfectly the blueprint transmitted from Earth (Parfit, 1984: 199-201). The body that "arrives" at Mars will be exactly like the one that "departed" at Earth.

While it is tempting to think that you have survived this transmission, the following variation of the above thought experiment shows that this cannot be correct. Suppose that you enter the transporter, but it malfunctions in the following way: it sends a perfect blueprint of your body to Mars where a perfect copy of your body—your "Replica"—is constructed out of new materials, but fails to destroy your body on Earth. According to Parfit: "Since I can talk to my Replica, it seems clear that he is *not* me. Though he is exactly like me, he is one person, and I am another. When I pinch myself, he feels nothing (Parfit, 1984: 204). But if your Replica is not you when the machine malfunctions, it simply cannot be you when it functions the way that it is supposed to. The difference between the two of you is, of course, that your Replica is "someone else" in an important sense: the phenomenal self that is the subject of its mental states is not you. It seems clear that, in the first case, *you*, construed as a phenomenal self or subject of conscious experience, have died in the sense that most matters to you: you are no longer capable of instantiating mental content.

[19] It is worth noting here that it is not logically possible that you are not the self associated with your body. Had your phenomenal self been paired with your twin's body, *that* body would be *yours* by definition. Again, the problem described above does not involve explaining some trivial truth.

[20] See, e.g., Hasker (1999).

[21] This is not, however, true of emergent dualism. The emergentist believes that a mental substance is causally produced by the ongoing operations of the brain and thus accepts that mentality supervenes upon physicality; unlike classical dualism, emergentism does not claim that these mental substances exist (or are capable of existing) independently of physical substances. Thus, assuming that the self-model theory solved hard and easy problems of subjectivity, an emergentist could adopt the self-model theory as an explanation of how these mental substances are causally produced by self-models. As far as I can tell, Metzinger lacks a compelling reason to think that the causal results of such models are mere appearances – and not substances. While I do not find emergentist views particularly plausible, I think there is an important point here: Metzinger's analysis assumes, rather than shows, that phenomenal selves are mere appearances.

[22] A physicalist can, of course, avoid this conclusion by rejecting the underlying presupposition that subjectivity is a phenomenal element of experience, but this seems intuitively implausible and a somewhat *ad hoc* response to this portion of the analysis.

[23] This thought experiment would be logically coherent even if eliminativism or identity theory were true provided that we understand these theories as making claims about only possible worlds that have the same nomological features and laws as this one. If we understand them as making a claim about all logically possible worlds (which I take to be a non-standard and quite implausible interpretation), then it is false that I could have been anyone else.

[24] It should be recalled that I intend the concept-term "self" to be construed as being compatible with the assumption that it is purely phenomenal in character – a mere appearance. Obviously, the concept-term itself is agnostic between dualism and physicalism.

[25] Elsewhere he describes them as "conceptual prototypes" (Metzinger 2003: 107), "working concepts" (Metzinger 2003: 208), and "conceptual devices" (Metzinger 2003: 303).

[26] For a discussion of the issue, see, e.g., van Inwagen (1990).

[27] See, e.g., Himma (2005).

References

Hasker, W. (1999). *The emergent self*. Ithaca, MY: Cornell University Press.

Himma, K.E. (2005). What is a problem for all is a problem for none: Substance dualism, physicalism, and the mind-body problem. *American Philosophical Quarterly*, 42, 81-92.

Honderich, T. (1995). Consciousness, neural functionalism, real subjectivity. *American Philosophical Quarterly*, 32.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Nagel, T. (1989). *The view from nowhere*. Oxford: Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Parvizi, J. and Damasio, A. (2001). Consciousness and the brainstem. *Cognition*, 79, 135-160.

van Inwagen, P. (1990). *Material beings*. Ithaca, NY: Cornell University Press.