PSYCHE

Reply to Himma: Personal Identity and Cartesian Intuitions

Thomas Metzinger
Philosophisches Seminar
Johannes Gutenberg-Universität Mainz
D-55099 Mainz
www.philosophie.uni-mainz.de/metzinger
metzinger@uni-mainz.de

PSYCHE 12 (4), August 2006

Reply to: Himma, K.E. 2005. The Problem of Explaining Phenomenal Selfhood: A Comment on Thomas Metzinger's Self-Model Theory of Subjectivity, *Psyche* 11 (5).

Keywords: Self, personal identity, Cartesian intuitions, phenomenal self, cognitive agency, attentional agency

In Kenneth Einar Himma's substantial commentary, there are a number of conceptual misunderstandings I want to get out of the way first. This will allow us to see the core of his contribution much clearer. On page 2, Himma writes about the problem of "explaining how it is that a particular phenomenal self (e.g., me) is associated with a set of neurophysiological processes." This philosophical question is ill posed: no one is identical to a particular phenomenal self. "Phenomenal self" must not be conflated with "me." Under SMT, phenomenal selves, in standard situations, are highly specific forms of representational content. They are not particulars in an ontological sense. First, Himma introduces the notion of a "mental subject," without giving any defining characteristics. He then proceeds to make a strong claim about conceptual necessity, presenting it as self-evident without an independent argument: "…it is not conceptually possible for a conscious mental state to occur that is not instantiated by a mental subject" (p.3). I must admit that I do not have this modal intuition, the point is not self-evident to me.

In the next paragraph, Himma begins by stating the apparently uncontroversial fact that "…we have a conscious sense of being phenomenal selves that function as mental subjects" (p.3). Unfortunately, even this point is controversial. Let us take the expression "mental subjects" to mean "subjects of mental states under a standard interpretation," where the "standard interpretation" is just ordinary, everyday folk-psychological discourse. I would then propose to delete the word "phenomenal" in this sentence. Why? Because the *phenomenality* of the phenomenal self is not a content of

conscious experience in standard situation. What is true is that we have a conscious sense of being *selves* functioning as mental subjects. The whole point about the phenomenal transparency of the self-model developed in BNO, however, was to draw attention to the fact that, in standard situations, human beings do *not* have a conscious sense "of being *phenomenal* selves," but simply of being selves.

In the first paragraph of section 2, Himma writes: "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)." It is not clear to me how it could come as a surprise that neurophysiological states turn out to be nothing over and above neurophysiological states. I also think that on page 6, top paragraph, there maybe something of a strawman fallacy, introduced by an equivocation of "epistemic subject" and "phenomenal subject" in the way Himma use the concept "mental subject". Some eliminativists may claim that no such things as "phenomenal subjects" (e.g., subjects of actually existing states of consciousness) exist at all. This, however, would not commit them to the claim that there is no kind of knowledge whatsoever, because *epistemic* subjects do not exist. For instance, scientific knowledge might still exist from an eliminativist's perspective, and the subjects of the process of expanding scientific knowledge might be scientific communities moving through time. Since "subject" is also a well-introduced notion in, say, logics and epistemology, it may be a misconstrual on the part of the opponent to seriously describe his claim as holding that *qua* (epistemic) subject he is actually nothing more than the relevant brain state (as he does in the next but one sentence). At least this is not my own position.

Another misunderstanding can be found in the next to the last sentence of the first paragraph in section 3. The point is not that the self-model is a "*self*-model in the sense that it performs … functional operations *for itself* and presents their outputs *to itself*" (p.7). Here is what the theory really says: the self-model performs certain functional operations *for* ***the system*** *itself* and represents their outputs *to* ***the system*** *itself*. The PSM is neither the subject nor the object of conscious self-representation. The whole point of the theory is to avoid the typical classical fallacies in German idealism that claim an identity of subject and object for the special case of self-knowledge and are then not able to give an account of the epistemicity of the underlying relation anymore. The PSM is not a little man in the head, an agent that performs functional operations. Rather, it is an *instrument* (in a teleofunctionalist sense) developed by the system as a whole to satisfy its needs. It also is not an epistemic agent representing information to itself—instead, it is a vital part of the *system as a whole* that achieves this.

In section 4.1 he presents us with an interesting thought experiment on two functionally isomorphic twins living on two different planets, to which I will return below. Here, the initial misunderstanding recurs, unfortunately this time in a much less benign form. On page 10, Himma writes, "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…" As pointed out above, this is a misconstrual of what the theory says. We are not phenomenal selves. We are systems transiently generating phenomenal selves. And as whole systems we have unique physical properties (space-time positions) that ground our individuality.

To simply begin assuming the existence of "selves" again, and then to ask questions about the strength of their relationship to particular self-models, is simply a *petitio principii* in this context. It assumes that selves in a strong sense exist. This would have to be shown first. The problem recurs a number of times, but it becomes most obvious when Himma refers to Nagel's beautiful, but incoherent neo-Cartesian interpretation of the succession of mental states caused by what he calls the "*View from Nowhere"*, in his book by the same title (cf. BNO: 582ff, 596f). He writes, "…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" (p. 11). This simply introduces a new entity, standing in an ownership relationship to the self-model. Let us follow Nagel and call it the "objective self". What is the empirical fact making the introduction of this additional entity necessary? What are the criteria making a level of description a *relevant* level of description? Himma's thought experiment actually seems to support my own point: the purported fact that one of the phenomenologically isomorphic self-models in our functionally isomorphic twins is *you* while the other is someone else is not an arbitrary fact. It is not a fact at all.

After clearing away some of these misunderstandings, let us turn to Himma's critique of the self-model theory, as presented in section 4. In his new version of a twin-earth experiment for phenomenal selfhood already mentioned above, he presents us with two functionally isomorphic twins living their self-conscious lives on two different planets. First, as Himma clearly sees, such twins would not share all *physical* properties: they would necessarily be located at different points in space-time. However, as both twins are also described as functionally identical, and given the theoretical background of SMT, it follows that they will also possess phenomenologically identical self-models. Because their *phenomenal* content supervenes locally, their PSMs will be equivalent in this respect. It is important to note, however, that the *intentional* content of our twins' mental representations will necessarily differ. Because their epistemic position and their perspective on the physical universe diverge, at least one of them may have a large number of false beliefs about himself (see above for a *Caveat*). And this is the reason why it is not true, as Himma claims on page 9 f., that you and your twin are mentally and physiologically indistinguishable: mental states are individuated by their *intentional* content, by what they represent for the system. True, you and your twin would have exactly the same kind of phenomenal self-experience. You would be phenomenological clones. But you would certainly not have identical self-knowledge. Your twin would have a host of false beliefs about his own physical history, and it does not matter how many of them are conscious, integrated into your PSM, and how many are not. Unconscious mental states are individuated by their intentional, representational content, possibly by their causal role. Conscious mental states like occurrent beliefs are individuated by their intentional, representational content *plus* the first-person characteristics we today call their "phenomenal content"—and how to reconcile these two types of characteristics, how to match up the inner and the outer taxonomy, *is* precisely the reason for the underlying philosophical problem, the epistemical asymmetry.

SMT says that the evolutionary function of phenomenal states consisted in making certain facts globally available to an organism within an internally constructed window of presence. Your phenomenological twin brother could never sign a contract on twin earth, because he would seriously misrepresent his socially constituted personal

identity. He would think he was you, and he would consciously experience himself exactly as you do, but he could never really sign a contract or buy a house. There is a large number of facts about his own history that he cannot make globally available with the help of his conscious self-model, because with regard to him, these facts simply do not exist—he has the wrong kind of history.

I have already pointed out that it is a *petitio* to simply claim that selves as distinct entities exist and are "associated with a stream of experience" (p.10), and that it is a misunderstanding to say that you are "identical" to a *phenomenal* self. Nevertheless, let us assume we were classical Cartesian souls, non-physical substances only contingently associated with the flow of experiences generated by a concrete physical body. As you and your twin can clearly be distinguished on the epistemological level of analysis, it would make a great difference for a substantial self whether it was associated with the twin possessing a much higher degree of self-knowledge, a much larger set of true beliefs about himself and his own history, or to the phenomenal clone on another planet, who only transparently *hallucinates* the possession of self-knowledge. Or would it?

Kenneth Himma has thought very hard and systematically about the self-model theory of subjectivity. For me, the perhaps most important point he makes (p. 15) is that in addition to phenomenal *mineness*, there is a more global phenomenal property, which he calls "me-ness." I fully agree that on our search for the minimal set of necessary conditions, for the *core* of phenomenal selfhood, many other factors than the sense of ownership alone play a role, and that these factors are important. As explained in BNO, I believe that the phenomenal experience of substantiality (as opposed to ownership) has a lot to do with invariance over time: we must further investigate those layers of the PSM that stay rather stable over time and are characterized by a high degree of invariance. I have offered autonomous, internal sources of input, like certain parts of the body image (the "proprioceptive background buzz") and certain aspects of our global emotional state (upper brain stem and hypothalamus) as candidates in BNO. There may be many more if we take a closer look. Himma's property of "me-ness" also has a lot to do with the constitution of autobiographical memory: a fully amnesic subject could well exhibit ownership, but would rarely possess the global phenomenal property Himma is trying to get at.

What is more, the conscious experience of *agency* certainly plays a vital role in constituting phenomenal me-ness, the sense of being a subject of intentions and goal states. In my own theory, I have analyzed agency as a specific subtype or form of ownership, because I think that phenomenal agency appears exactly when certain time slices of the process of assembling specific motor commands and possibly of integrating reafferent bodily feedback are integrated into the conscious self-model (see Metzinger 2006). But now, eminent French philosophers like Elisabeth Pacherie and prominent neuroscientists like Marc Jeannerod are developing an alternative model portraying agency as an entirely different phenomenon than ownership. Of course, I will not enter this discussion here, but it is clear that goal states and intentions may contribute more to what Himma has in mind than the simple bodily sense of ownership alone. To support Himma's point, let me also point out that this is particularly true of the two more subtle concepts I have introduced, namely "cognitive agency" and "attentional agency": I am quite convinced that, for instance, the conscious experience of being able to control and

direct your attentional focus has a much greater role to play in the phenomenology of selfhood than is commonly assumed.

Fourth, and last, if we are interested in longer time windows and in understanding the genesis of Himma's more comprehensive phenomenal property of me-ness, we do not only have to think about neural correlates, but must also begin to think about the *social correlates* of conscious selfhood. Many empirical data seem to show how low-level bodily ownership may be partially hardwired and in full existence at birth (e.g., in the phantom limb experiences of congenitally limb-deficient patients). "Me-ness," however, is something that must clearly be *learned* in the course of social interactions. As a matter of fact, another way of strengthening Himma's point could be by saying that emotional auto-regulation and the different varieties of phenomenally experienced agency (see above) are acquired post-natally as well, in a social context. In this sense there are clearly strong functional differences underlying the conscious experience of ownership vs. the conscious experience of being a *subject* of these states.

References

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