PSYCHE

Reply to Gallagher: Different conceptions of embodiment

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: Gallagher, S. 2005. Metzinger's Matrix: Living the Virtual Life with a Real Body, *Psyche* 11 (5).

Keywords: Embodiment, disembodiment, Out-of-body epxierences (OBEs), Cotard syndrome, phenomenal transparency

Let me begin by thanking both the Dorothée Legrand, the editor of this special issue, and Timothy Bayne, general editor of PSYCHE for their great and sustained efforts to make this debate possible. Everyone who has ever done this type of service to the philosophical community knows how much work it *really* is – I am therefore more than grateful to both of them, and I am certain that the same is true of all my critics and commentators as well. Of course, I am also deeply indebted to all of the commentators themselves for making this debate possible, and for giving me the opportunity to learn from their interesting and substantial criticism. However, after some thought, and because their contributions explore a considerable range of quite diverse topics, I have finally decided to not organize my replies on the following pages along thematic lines and in a single piece, but to reply to each author individually. I hope that, for the majority of readers, this makes my replies more accessible.

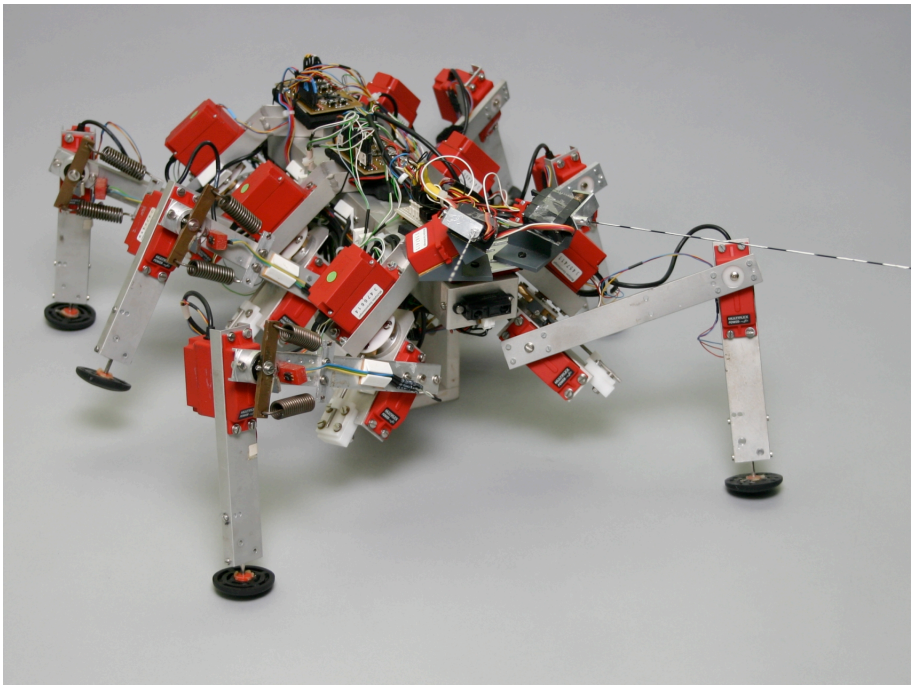
Gallagher is right in pointing out that scientific realism is an implicit background assumption of BNO, and that I did not give an independent argument for it. He is also right in saying that science does not *demonstrate* the existence of certain entities, but that it assumes those entities in a process of explanation and theory formation. However, it is not true that science, as Gallagher writes (p.2), "simply" assumes the reality of certain things: such assumptions are embedded in the context of an attempt to find the *minimal* set of ontological assumptions one has to make relative to a set of explanatory goals and relative to a specific data set in a certain domain. This parsimonious spirit is also the

spirit of SMT, which can be seen as a search for the minimal conditions under which a phenomenal self and a consciously experienced first-person perspective can emerge.

Gallagher takes the notion of "full-blown, pre-reflective embodiment" as the starting point for his commentary, and this is certainly a good idea. However, his selection of the first two isolated quotations from BNO suggests that he does not fully understand that, in the context of embodiment, I am mostly concerned with representational content and the *phenomenology* of embodiment. Of course, I also make claims about functional properties realized by the unconscious, implicit parts of the human self-model and about how causal interaction between the phenomenal and non-phenomenal layers of our self-model helps in implementing intelligence and evolving new functional properties. However, my main goal has clearly been to understand the role of the PSM in achieving embodied cognition—whatever that may mean.

"Embodiment," unfortunately, has long become a trendy buzzword. Probably precisely because of its implicit Cartesian connotations in an explicitly anti-Cartesian approach, its semantic vagueness, and the spatial-mental imagery it evokes, it is now used by many different authors in many different ways. For some, "embodiment" is something that has to do with robotics, for others, it is something that has to do with "existing under the gaze of the other". Although Gallagher himself certainly is on the side of those interested in conceptual clarity, he himself doesn't offer a definition of the term "embodiment", at least neither in his commentary nor in his recent (2005) monograph. Let me try to quickly develop a minimal conceptual platform. We need at least *some* conceptual clarification. Therefore, before we go on, let me introduce three new working concepts: "first-order embodiment," "second-order embodiment," and "third-order embodiment."

"First-order embodiment" (1E) is aimed at and can be found, for instance, in biorobotics and in all "bottom-up approaches" to artificial intelligence. The basic idea is to investigate how intelligent behavior and other complex system properties, which we previously termed "mental," can naturally evolve out of the dynamical, self-organizing interactions between the environment and a purely physical, reactive system that does not possess anything like a central processor or "software" and no explicit computation. For researchers in 1E, the relevant questions are: How could the very first forms of pre-rational intelligence emerge in a physical universe? How could we *acquire* a flexible, evolvable, and coherent behavioral profile in the course of natural evolution? How is it possible to generate intelligent behavior without explicit computation? Here is an example of 1E, the tripod gait as exhibited by the walking machine Tarry II:



For more information, see http://www.tarry.de/index_us.html

"Second-order embodiment" (2E) can develop in a system that satisfies the following three conditions: (a) we can successfully understand the intelligence of its behavior and other "mental" properties by describing it as a *representational* system, (b) this system has a single, explicit and coherent self-representation of itself *as being an embodied agent*, and (c) the way in which this system uses this explicit internal model of itself as an entity possessing and controlling a body helps us understand its intelligence and its psychology in functional terms. Some advanced robots, many primitive animals on our planet, and possibly sleepwalking human beings or patients during certain epileptic absence seizures (as discussed in BNO) could be examples of 2E.

"Third-order embodiment" (3E) is the special case (indeed the *very* special case) in which a physical system not only explicitly models itself *as* an embodied being, but also maps some of the representational content generated in this process directly onto conscious experience. That is, 3E means that in addition, you consciously *experience* yourself as embodied, that you possess a specific type of what, in BNO, I call a "phenomenal self-model" (PSM). Human beings in ordinary wake states, but also orangutans swinging from branch to branch at great height, could be examples of 3E: they have an online model of their own body as a whole that has been elevated to the level of global availability and integrated within a virtual window of presence. They are consciously present as bodily selves.

The general framework emerging from this threefold distinction is that human beings permanently possess 1E and 2E: a considerable part of our own behavioral intelligence is achieved without explicit computation and results directly from physical properties of our bodies, such as the genetically determined elasticity of muscles and

tendons, or the degrees of freedom realized by the special shape of our joints. Moreover, certain parts of our unconscious self-model, such as the immune system and the elementary bioregulatory processes in the upper brain stem and the hypothalamus, are continuously active. Another candidate for an important aspect of the unconscious self-model, a representation of global properties of the body, is the body schema (a concept which Gallagher has critically, and helpfully, discussed over the years, most recently in his 2005 monograph). Having an unconscious body schema is clearly a new, biological form of intelligence: having a body schema means having 2E. Only episodically, during wakefulness and in the dream state, do human beings realize 3E. (Even Gallagher's own everyday notion of "enactive embodiment" (see p. 8 of commentary) does not refer to anything that can be found in deep sleep or even in dreams—it is something only existing in waking consciousness and in lucid dreams.) It is important to understand that BNO is mainly about the relationship between 2E and 3E.

Let us now map these distinctions onto Gallagher's four types of disembodiment.

(1) Cartesian disembodiment is just a theoretical position, and it has nothing to do with individual phenomenology as such. Descartes' account has nothing to say about 1E, 2E, or 3E.

(2) The Cotard patient clearly has 1E. The hypothesis put forward in BNO is that, during severe psychotic depression, he lacks a specific layer of representational content in terms of 2E, and that the phenomenological profile of his conscious self-model lacks important dimensions. He is "emotionally disembodied," and the life process itself is not reflected on the level of 3E anymore. But, because he still has a representation of himself as a spatially extended entity possessing sensors and effectors, as a potential agent in the world, he clearly still has 2E and 3E. Gallagher is not correct in assuming that these patients only "think" that their body is dead or that they suffer from a purely cognitive deficit, like a failure to "recognize" their body as a "lived" body. Very obviously, Cotard patients have a major distortion on the level of 3E – and this includes much more than a distortion in their web of beliefs, it is a distortion of non-conceptual, emotional and proprioceptive layers in their PSM as well. As Dan Zahavi (2006: 145) writes, it "might be wrong to interpret the delusions as if they were simply strongly held ordinary beliefs that happen to be false." True, Cotard patients are often cognitively inconsistent. Nevertheless, the utterances of these patients do not rest on a purely cognitive delusion, but on a massive reconfiguration of the emotional self-model (see Gerrans 2000).

(3) Out-of-body experiences (OBEs) are clearly cases of 3E, because, at least in a majority of cases, they include some sort of ethereal double, a conscious self-model of a spatially extended and perceiving agent. Then, there seem to be rare borderline cases where the phenomenal property of selfhood is only instantiated in terms of what, in BNO, I termed "attentional agency" and "cognitive agency"; the location of the self is only instantiated as an unextended point in visual space, which forms the origin of a visual perspective. Phenomenologically, a thinker of thoughts and an entity actively directing its attention is preserved. Please note that as long as there is a perspectivally organized visual space, even an extensionless point, for instance the geometrical *origin* of the visual perspective, will have to count as a form of

spatial phenomenal content. I think this is enough to categorize even "bodiless" OBEs as cases of 3E.

(4) Brain-in-a-vat disembodiment, just like (1), is just a theoretical possibility. As such, it is a thought experiment and not a statement about individual, real-world phenomenology. In terms of the conceptual distinctions introduced above, there is an absence of 1E. In terms of 2E, the epistemic status of almost all self-representational content activated in the isolated brain would change dramatically: most of it would now be *misrepresentational* content. However, since phenomenal content, the way things *appear* to you, supervenes locally (see also Metzinger 2004), our effectorless brain in a vat could enjoy 3E in the absence of both 1E and 2E. Or would it? The available empirical evidence from research on dreams and hallucinations generally, but also from whole-body illusions caused by direct brain-stimulation of the right angular gyrus (see, e.g., (Blanke, Ortigue, Landis, and Seeck 2002; for a more detailed hypothesis concerning the temporo-parietal junction, see Blanke, Landis, Spinelli, and Seeck 2004) certainly makes it overwhelmingly plausible that the phenomenal experience of embodiment could continue. However, there is an interesting philosophical point here: the self-model activated in a brain in a vat no longer fulfills condition (c), because it does not help us understand the intelligence of this system in functional terms, as a function of *using* its self-model. Or does it? It would certainly still possess a Millikanian "proper function" in the sense that it played a specific causal role in the biological/evolutionary *history* of our poor brain in the vat, namely for its biological ancestors. In its new, ecologically invalid, situation, and given the host of false beliefs about itself, would it still be appropriate to describe it as a *representational* system?

Due to limited space, I cannot enter into an extended discussion of whether a brain in a vat can still be described as "intelligent" at this point, or as an *epistemic* subject at all. Could we say that it now has a large number of false beliefs *de se*? This would depend on the further details of the respective thought experiments and on assumptions about the necessity of external relations for the realization of intelligence. Therefore, my interim conclusion at this point will be that a brain in a vat is only a "weakly representational" system, which may not even possess 2E in any stronger, philosophically interesting sense—in a sense that allows us to hold on to a *representationalist* theory of consciousness. Nevertheless, as appearance as such is neither knowledge nor intelligence, there may be no principled conceptual obstacles to the claim that a brain in a vat could actually have 3E: Having 3E would then have to be identical to some complex, but local *functional* property of our isolated brain. And a fully reductive, domain-specific identity claim should be tenable.

There are some minor empirical difficulties and conceptual misunderstandings in Gallagher's commentary. Of course, the patient with unilateral hemineglect does not lack a "body image," as Gallagher (p. 5) claims. She only lacks *part* of a body image, because this part of the self-model is no longer attentionally available. It is also a misunderstanding, as Gallagher (p. 5) writes, that in situations where "we lack any explicit experience of our own body—it becomes transparent, in just the way that Metzinger claims the self-model becomes transparent." The way the terminology is

introduced in BNO, transparency is a property of phenomenal representations only. In this sense, Gallagher could only refer to the "experience of our own body". If, as Gallagher interestingly points out, the body image actually becomes unconscious or "implicit" during intentional action or periods of attentional distraction, then it is neither transparent nor opaque (see also Metzinger 2003b). In the terminology proposed in BNO, the bodily self-model is still *available* for attention in these cases (and therefore counts as conscious in a weaker sense), but is not currently accessed by attentional processing. The prediction my theory makes is that there are actually many situations in which we "become absorbed" or "lose ourselves," in which large parts of the PSM are transiently shut down as it were, but in which availability is preserved in the absence of ongoing access (for a recent empirical study supporting this idea, see Goldberg, Harel, and Malach 2006). In the terminology of the Kiel school of *Neue Phänomenologie*, which may be more to Gallagher's taste, currently accessed and processed parts of the bodily self-model are termed "body islands."

But it really looks like Gallagher wants to introduce a new use of the term "transparency"—as a property of physical bodies themselves. This could not be a physical property—bodies certainly do not become invisible by being successfully "enacted" (whatever the precise meaning of this term may be). So Gallagher could also envision a new *functional* property of our bodies. In BNO (p. 177, 294) I pointed out how the *brain* can be analyzed as functionally blind to itself (because it has no internal sensory perception whatsoever, and therefore, as the incessantly active medium of conscious experience, cannot be directly sensed at all). Could a whole body or person become functionally blind to itself in this sense? Yes, transiently it could—if, as in absorption, distraction, or maximally congruent, successfully sensorimotor interaction with it environment it shuts down the relevant layers of the self-model.

Gallagher in describing the lived body as "the body I live" (p. 7, bottom paragraph) introduces a distinction between himself and his body, the relationship between the two being that he himself, Shaun Gallagher, "lives it." There is a person, and an object, an "it". The real body, of course, could be replaced by a functionally isomorphic system. Then he himself, Gallagher, could again "live it." It is just like a horse and its rider: if a suitable horse would be available, the rider could in principle change to it. The self-model theory is free of these underlying Cartesian intuitions: it does *not* say (which is one of the most frequent, recurrent misunderstandings, also in other commentaries of this special issue) that "you are" simply a self-model (p. 8). When you refer to yourself using "I," you refer to the system as a whole, including your brain, body, self-model, history, and social context—but you do so in a very special, displaced manner: by using the content of your PSM as an intermediary in the act of self-reference, most of the time without noticing this fact.

References

Blanke, O., Ortigue, S., Landis, T., and Seeck, M. (2002). Stimulating illusory own-body peceptions. *Nature* 419: 269–270.

Blanke, O., Landis, T., Spinelli, L., and Seeck, M. (2004). Out-of-Body experience and autoscopy of neurological origin. *Brain* 127: 243–58.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford; New York: Clarendon Press.

Gerrans, P. (2000). Refining the Explanation of Cotard's Delusion. *Mind & Language* 15/1: 111-122.

Goldberg, I.I., Harel, M., and Malach, R. (2006). When the brain loses its self: Prefrontal inactivation during sensorimotor processing. *Neuron* 50: 329-39.

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, 2: 353-393.

Metzinger, T. (2004). Appearance is not knowledge: The incoherent strawman, content-content confusions and mindless conscious subjects. Invited commentary for Alva Noë and Evan Thompson: "Are there neural correlates of consciousness?" *Journal of Consciousness Studies*, 11/1: 67-71.

Zahavi, D. (2006). *Subjectivity and Selfhood*. Cambridge, MA: MIT Press.