Splendor and misery of self-models

Conceptual and empirical issues regarding consciousness and self-consciousness

An interview with
Thomas Metzinger

By Jakub Limanowski & Raphaël Millière

Citation: Metzinger, T., Limanowski, J. & Millière, R. (2018). Splendor and misery of self-models: conceptual and empirical issues regarding consciousness and self-consciousness. An interview with Thomas Metzinger. *ALIUS Bulletin*, 2, 53-73

Thomas Metzinger
metzinge@uni-mainz.de
Department of Philosophy
Johannes Gutenberg-Universität, Mainz, Germany

Jakub Limanowski
j.limanowski@ucl.ac.uk
Wellcome Centre for Human Neuroimaging
University College London, UK

Raphaël Millière
raphael.milliere@philosophy.ox.ac.uk
Faculty of Philosophy
University of Oxford, UK

Hello Thomas. Are you a self right now?

Hello Raphaël and Jakub! Hmmm…. If I really was "a" self, how would I *know*? Could I ever catch myself, epistemically, in my *substantiality?* And if I wasn't, how could *I* know?

In *Consciousness Explained*, Dan Dennett has famously criticized what he calls Philosophers' Syndrome, which consists of "mistaking a failure of imagination for an insight into necessity" (Dennett, 1991, p. 401). You have yourself expressed similar concerns regarding "armchair" philosophy of mind, and have always favored the analysis of empirical cases over thought experiments. Furthermore, you have dedicated a good deal of attention to pathological or otherwise non-ordinary states of consciousness, from autoscopic phenomena, depersonalization and somatoparaphrenia to dreaming, meditation and full-body illusions—among many others. In your opinion, what place should the discussion of empirical data about so-called altered states of consciousness have in philosophy of mind?

There are many ways in which it can be useful. For example, it changes our theoretical intuitions. Intuitions are phenomenal states that guide our thinking, and they are millions of years old—shaped in the world of our ancestors, determining the attractor landscapes of our brains today, setting priors. They have the form "I *just know* x", without us having any idea or introspective access to the causal history of this knowledge—in intuitive "insight" we suddenly know something, but we really

have no idea where this knowledge comes from. What many don't see is that there is a distinct phenomenology of knowing, and also a phenomenology of certainty (of knowing that one knows). The phenomenal signature of "knowing" is characterized by the phenomenology of direct accessibility of knowledge (which may be preceded by the initial phase of the phenomenology of ambiguity), which sometimes has a character of immediacy. Systematically and rationally investigating altered states of consciousness has the great advantage of exploding many of your theoretical intuitions, and very efficiently. It makes you open-minded. Against the background of a serious interest in the issues and a good academic training it may simply be the most fruitful general heuristic available. If you had a metric to compare the *fecundity* of armchair phenomenology or old-school analytical philosophy of mind to empirically-informed, interdisciplinary philosophy of cognitive science, then what would you think the results over the last three decades would be?

> "Philosophers should design and propose their own experiments."

A frequent epistemological fallacy consists in ascribing epistemic status to phenomenal states because of the phenomenal signature of knowing. Jennifer Windt and myself, in a German paper, called this the "E-fallacy" or "E-error" (Metzinger & Windt, 2014). I would like to point anybody interested in this point to section 3.1 of our introduction to the Open MIND project (http://open-mind.net). Just because something *feels* like an insight doesn't mean that *is* an insight, to put it in an oversimplified way. But a large part of academic philosophy in the last century has consisted in exactly this—"making our intuitions explicit". However, the great danger in cultivating so called "first-person methods" (perhaps as an antidote to intuition-mongering, as exemplified by some forms of old-school armchair philosophizing) is that people have dramatic experiences of deep, ineffable "insights" or subjective "certainty", and then re-iterate the E-fallacy.

There are not only logically possible worlds, but also phenomenally possible worlds: each world that can be simulated on the phenomenal level, relative to a certain class of systems; the possibility of its simulation depending on the functional architecture. However, the functional architecture of our brains has not evolved to help us generate metatheoretical knowledge—therefore we should always be very careful with modal intuitions about what is necessary or possible. So-called "xPhi" or "experimental philosophy" uses statistical estimation of phenomenological reports of others to search the landscape of possible phenomenal worlds in different populations and comparing the intuitions of lay persons to the academic ones, and

it is very stimulating and leads to interesting results. However, the method of interdisciplinary constraint satisfaction (Weisberg, 2006) that I have tried to develop not only uses empirical bottom-up constraints in domain-specific theory-formation, but ideally also shapes the epistemic aim and the process of experiments themselves. Philosophers should design and propose their own experiments! It is more like a methodological experiment of an intuition-free and actively interdisciplinary oriented philosophy of mind.

There is a downside to this: All our results will be preliminary, and highly domain-specific (e.g., only applicable to human minds). Thirty-five years ago, I was fascinated by Hilary Putnam and the project of classical machine functionalism, the "truly philosophical" project of a fully hardware-independent "universal psychology" where we could aim at saying what consciousness, cognition, and so on, really are in *all* possible beings that instantiate them, no matter how they are physically realized (if at all). Now I have a slightly more modest and sober attitude—that is perhaps another downside of looking into the messy details of real-word embodiment with an open mind.

Another, the third, downside is that you become aware of your own psychological vulnerability and your own mortality in a much more acute way, and that many of the relevant recent empirical discoveries are sobering and unattractive on an *emotional* level. Because, as I have come to think, a very strong and mostly unconscious motive for many people to become interested in the philosophy of mind and related areas in the first-place is to discover something that is uplifting, emotionally thrilling and entertaining, causing phenomenal experiences of "meaningfulness", and which helps them develop sustainable psychological strategies for mortality-denial and self-deception, you will face a lot of resistance by the philosophical establishment and parts of the public.

In *Being No One* (Metzinger, 2003), you argue that the folk psychological view of selfhood is thoroughly misguided, insofar as no such things as selves exist in the world. More precisely, you claim that what we traditionally call "the self" is nothing like a mind-independent substance, but a special kind of representational content, namely the content of a sophisticated mental model—a "self-model"—reducible to neurophysiological processes. Yet you maintain that there is such a thing as an experience of selfhood, construed as the dynamic content of the phenomenal self-model, or 'PSM' for short. On this view, the PSM is simply the part of the self-model whose content is phenomenally conscious. This is consistent with your antirealist stance on selfhood insofar as the PSM is a mental process rather than a substance. Moreover, you argue that self-conscious systems such as human beings *identify* with the content of their PSM. Importantly, you claim that the PSM is phenomenally *transparent*, meaning that the various stages underlying this cognitive model are not available for introspective attention. According to you, this transparency constraint

entails that we cannot be aware of our PSM *as a model*: this explains why we have the illusion of being 'selves' in the folk-psychological sense, i.e. substantial, holistic entities. Indeed, having a PSM entails the instantiation of a phenomenal property described as the "primitive, prereflexive feeling of conscious selfhood" (Metzinger 2003, p. 565). However, in a recent talk given at the Sense of Self conference in Oxford (http://senseofselfoxford.wordpress.com), you argued that "if one takes the phenomenology seriously, there really is no such thing as a determinate subjective quality of 'selfhood'" (quoted from the abstract). On the face of it, there seems to be a significant tension between these two claims. Did you revise your hypothesis regarding the existence of a phenomenology of selfhood in PSM-endowed organisms, or does this apparent tension result from a misunderstanding?

The 2017 Oxford talk was entitled "MPS reloaded" and took a critical look back at a paper which I co-authored with Olaf Blanke in *Trends in Cognitive Sciences* in 2009, and which has been cited more than 400 times (Blanke & Metzinger, 2009; see also Metzinger, 2008). One central aim of this paper was to isolate a minimal model of self-consciousness, the phenomenal property of "minimal phenomenal selfhood", which we defined as "transparent spatiotemporal self-location". One major result of the investigation was that the phenomenology of agency is *not* part of MPS, another one was that (in asomatic out-of-body experiences and bodiless dreams) an extensionless point in space can suffice as the locus of identification. We claimed that having MPS is a necessary condition for developing a strong, cognitive or attentional, first-person perspective (1PP), that is of developing what today I would call an "epistemic agent model" (or EAM) (Metzinger, 2013a, 2017a, 2018; see also my new essay on mind-wandering for AEON: http://bit.ly/2DAckUu). We also claimed that spatiotemporal self-location, self-identification (through phenomenal transparency), and a weak 1PP in the purely geometrical sense of an egocentric frame of reference are necessary and sufficient for MPS. Many people seem to have agreed with this general conceptual framework.

I now think that one subtle mistake I may have made is the uncritical assumption that the property called "MPS" is *phenomenally determinate*. In Oxford, I illustrated the problem by dubbing it the "Refrigerator Light Problem": You believe that whenever you close the refrigerator door the lights go out. However, whenever you try to verify your belief and carefully peep into it, the lights automatically go on. In the talk, I discussed first-person methods like classical mindfulness meditation and the status of introspective reports of the type "Whenever I effortlessly come to rest in a clear, emotionally neutral, thoughtless state, I experience MPS". If somebody claims that they introspectively know that MPS is a distinct quality, which can be instantiated in isolation, they face the problem that any attempt at introspective validation automatically creates a much more elaborate phenomenal structure, including an EAM. If you try to find out what the "rock bottom" level of self-

awareness is by willfully directing your attention inwards, then you create a sense of effort and the phenomenal quality of attentional agency. There is no introspective knowledge of MPS as such. What overlooked is that MPS may actually be phenomenally indeterminate.

Phenomenal indeterminacy is $\neg(F(a) \vee \neg F(a))$, i.e., "neither-nor", relative to phenomenal content, as in the sentence "Raphaël neither *instantiates* the phenomenal property called 'MPS' nor he does *not instantiate* the phenomenal property called 'MPS'". This is not the same as $F(a) \wedge \neg F(a)$, i.e., contradiction as in "Raphaël *instantiates* the phenomenal property called 'MPS' and he does *not instantiate* the phenomenal property called 'MPS' *at the same time*" and it also not the same as phenomen*ological* indeterminacy: "Raphaël retrospectively reports that there was no phenomenal fact of the matter regarding minimal phenomenal selfhood" or "Raphaël retrospectively reports that there was a phenomenal fact of the matter, which cannot be expressed in natural language".

I think we must reject introspective authority: we can assume that there is a determinate phenomenal fact of the matter, but at this point in time we do not (scientifically) know it. Blanke & Metzinger (2009) were right, and in the future science may show that "transparent spatio-temporal self-location" is conscious, and *determinate* with regard to the sense of self. However, *individual subjects themselves* are interestingly limited in their access to "minimal selfhood": We find ourselves in a very special epistemic situation with regard to minimal self-consciousness (that was one of my main points in the talk). The instantiation of MPS is an epistemically elusive conscious experience: *a phenomenal fact that is unknown to the subject*. MPS is a 3PP-determinate phenomenal fact, but, currently, epistemically indeterminate. In its minimality, MPS is 1PP-*indeterminable*: we assume that there is a determinate phenomenal fact of the matter, but we are in principle unable to know it $(F(a) \vee \neg F(a))$. Yes, Blanke & Metzinger (2009) were right, and in the future science will show that "phenomenally transparent spatio-temporal self-location" is an objective fact, and *determinate* with regard to the sense of self. But they overlooked that there will always be 1PP-indeterminability: 3PP-knowledge (involving excellent predictive power, etc.) can be had, but this knowledge will never be 1PP-validated, because attentional and/or cognitive agency necessarily activates an EAM.

I have always been interested in how exactly the phenomenal self "bottoms out" (Metzinger, 2014), and also what is the relevant layer in the human self-model that creates the transition from a weak to a strong first-person perspective *above* MPS (namely, the EAM). If you would like a more poetic description, 1PP-indeterminability of MPS can perhaps be read as the "groundlessness" of self-consciousness.

I think there is an additional interesting discovery, which I tried to draw attention to on the excellent meeting you organized. I call it "indeterminacy blindness": Human beings are completely unaware of the fact that they are introspectively blind to truly *fundamental* and philosophically relevant phenomenal facts, namely, the indeterminate *origin* of their very own 1PP. If this is true, then all autophenomenological reports about the "innermost core of the conscious self" are highly dubious and necessarily theory-contaminated. If I am right in my two claims about 1PP-indeterminability and indeterminacy blindness, then I think this is a philosophically interesting feature of human self-consciousness that might warrant further research.

> "I think we must reject introspective authority: we can assume that there is a determinate phenomenal fact of the matter, but at this point in time we do not (scientifically) know it."

A key concept in your Self-Model Theory of Subjectivity (SMT) is the notion of *phenomenal transparency* of conscious mental representations, which means that only the content of such representations is accessible to consciousness—not the fact that they are representations. As mentioned above, you propose that the experience of being a self arises from such a phenomenally transparent part of a system's self-model. This content (i.e., the *phenomenal self-model*) may be used to represent the subject component in a subject-object relationship, while also representing this relationship—what you call a phenomenal model of the intentionality relation. An interesting criticism of SMT was put forth by Josh Weisberg, who worried that the theory "makes too much of the system phenomenal" (Weisberg, 2006). Weisberg instead proposes, very much in the spirit of higher-order theories of consciousness, that to become conscious, the phenomenal self-model needs to be integrated into a nonconscious model of the intentionality relation (NMIR). Has your conception of SMT changed in response to such thoughts?

Of all the critical reviews of BNO, Weisberg's is probably my favorite one—very intelligent, careful and constructive. I have not looked into this issue for a long time, but in a 2003 paper co-authored with Vittorio Gallese and entitled "The emergence of a shared action ontology: Building blocks for a theory", we showed there exist unconscious functional precursors of what can later also be phenomenally represented as a goal, an acting self or an individual first-person perspective (Metzinger & Gallese, 2003). Empirical evidence demonstrates that the brain models movements and action goals in terms of multimodal representations of organism-object-relations and there is empirical evidence for mirror neurons as specifically coding organism-object relations on various levels of abstraction. The motor system

constructs goal-states (successfully terminated actions), action models, and intending selves as basic constituents of the world it interprets by assigning a single, unified causal role to them. I must confess that I have not thought about this a for a long time and have not followed the empirical literature. My intuition is that the PMIR is anchored and dependent on competing, unconscious MIRs, which in turn evolved out of the need to dynamically model whole organism/object-relationships like grasping. My proposal is that first the brain had to model spatial/motor relationships (for example as observed in conspecifics), then it used this basic schema to represent *semantic* relations like "reference" and *epistemic* relations like "attending to a perceptual object" or "grasping an abstract object".

Have you ever thought about the concept of "grasping a *concept*"? It is perhaps the essence of high-level cognition, of human thought itself. It may have to do with simulating hand movements in your mind but in a much more abstract manner. I once looked into this and found out that humankind has apparently known this for centuries, intuitively or because our ancestors had a much more fine-grained introspection than we do today: "Concept" comes from the Latin *conceptum*, meaning the "fruit of the womb" or "a thing conceived," which, just like our modern "to conceive of something," is rooted in the Latin verb *concipere*, "to take in and hold." At this time, the capacity of a woman to successfully "hold the fruit of the womb" was not something self-evident, not something that could be taken for granted, because many more pregnancies failed than today. As early as 1340, a second meaning of the term had appeared: "taking into your mind." If we go back to the original meaning, then infecting other people with memes via philosophical discussion it like trying to make them pregnant with your own ideas—making them "hold" what you take to be your own intellectual fruit in their own brains, by something we like to call "rational argument" forcing them to "take in and hold" what you (perhaps mistakenly) experience as your *own* insights, hopefully later giving birth to something beautiful.

Surprisingly, there is a representation of the human hand in Broca's area, a section of the human brain involved in language processing, speech or sign production, and comprehension. A number of studies have shown that hand/arm gestures and movements of the mouth are linked through a common neural substrate. For example, grasping movements influence pronunciation—and not only when they are executed but also when they are observed. It has also been demonstrated that hand gestures and mouth gestures are directly linked in humans, and the oro-laryngeal movement patterns we create in order to produce speech are a part of this link. By the way, such empirical data are good examples of something that philosopher of language and cognition should know.

Broca's area is also a marker for the development of language in human evolution, so it is intriguing to see that it also contains a motor representation of hand

movements; here may be a part of the bridge that led from the "body semantics" of gestures and the bodily self-model to linguistic semantics, associated with sounds, speech production, and abstract meaning expressed in our cognitive self-model, the thinking self. I think Weisberg was absolutely right when demanding the phenomenological notion of a "model of the intentionality relationship" (which in my recent writings has somewhat morphed into the EAM, or "epistemic agent model") must be grounded in unconscious mechanisms and an evolutionary story. But, again, I must admit that I have not monitored empirical research in this area for a long time.

You have recently edited an open access volumes that are largely drawing on the so-called predictive processing framework and discuss its philosophical implications (Metzinger & Wiese, 2017, http://predictive-mind.net). The predictive processing framework has appealed to philosophers, however, it has been interpreted both in representationalist terms (Hohwy, 2013; Gładziejewski, 2016) or along enactivist ideas (Bruineberg, Kiverstein, & Rietveld, 2016; Gallagher & Allen, 2016). The active inference formulation of predictive processing (Friston, 2009) has been proposed to dissolve this tension: on the one hand, active inference fundamentally assumes inference on representations in hierarchical generative models in the brain—thus appealing to representationalist accounts. On the other hand, active inference is all about reaching the best possible (i.e., least surprising) situation of myself in and as part of my world, and hence representations arise from interaction with the world—thus appealing to enactivist ideas. Since your initial SMT is a purely representationalist account, do you think the active inference (predictive processing) framework may indeed resolve some issues that philosophers have been arguing about for a while now—or do you think this is a too ambitious claim? How much do you think theoretical neuroscience and philosophy can mutually enrich each other?

Oh, they can certainly enrich each other—but it may need a new generation of philosophers who not only know neuroscience and cognitive science, but also mathematics. Personally, I have a very relaxed attitude about the concept of "representation". Many people take me as a realist, and sometimes also assume some caricature concept of "representation", but actually I am more of an instrumentalist. We live and work in a certain period in the history of science, and it is important to never forget that concepts are *historically plastic* entities, they move through time, just like scientific communities do. They are instruments used by communities of epistemic subjects, they serve a purpose for a certain time, eventually you have to throw them away. I have seen a lot of changes from early machine functionalism to the "computer model of mind" and on to connectionist representation (Paul Churchland's *A Neurocomputational Perspective* and Andy Clark's *Microcognition* were important books in my own intellectual biography), dynamicism and EEEE. I think that running neural models described as having properties like "integrated

likelihood" or "model evidence" in Bayesian statistics can still count as representational processes, and that the representational *level of analysis* continues to be very useful and fecund. But that doesn't commit one to realism, it is just a theoretical tool that works for a certain time—maybe we can dissolve it all into measures of entropy or something else soon.

I still remember when, ages ago, Francisco Varela invited Dave Chalmers and me to Paris, and after my talk on self-models he said to me: "I think in principle your whole story is absolutely right, but with that representationalism it is all false and you will *never* get anywhere!" Maybe so. The two of us had more in common than we ever had a chance to explore, that is for sure. But with all the trendy-sexy stuff today, I wonder if he would have liked it. "Enactivism" obviously refers to a relation: Some *A* "enacts" some *B*. Can someone tell us in a non-circular way what that *A* and that *B* actually are?

You have devoted much of your time to ethical problems implied by cultural or technological advances. Recently, you have discussed the potential implications of technological advances in artificial intelligence. One of the appealing features of your SMT is that one can in principle derive empirically testable predictions about when an (artificial) organism or system would experience a first-person perspective and, ultimately, phenomenal selfhood (Blanke & Metzinger, 2009). In light of your recent modifications of your initial proposal—do you think a collaborative effort of philosophers and cognitive and computer scientists could in fact lead to a form of "Turing test" for first-person perspective and experience of selfhood in artificial systems? And, if this was the case, what would your advice to AI developers be—do you think we should be (more) worried by these recent developments?

I have indeed been recently working on ethical issues raised by technological advances such as Virtual Reality (Madary & Metzinger, 2016) and Artificial Intelligence (Metzinger, 2017b). As you may or may not know, I have demanded a moratorium for synthetic phenomenology for quite a number of years now. I think we should always try to minimize the overall amount of suffering in the universe, and recklessly creating artificial consciousness would carry a high risk of *increasing* the overall amount of suffering in the universe. The last time I have done so was in the very short piece I wrote for the EU, asking for the development of Global AI Charter. Here is an excerpt from the forthcoming collection *Should we fear the future of artificial intelligence?* (reproduced here with permission of the STOA Panel of the European Parliament):

A Moratorium on Synthetic Phenomenology

> It is important that all politicians understand the difference between "artificial intelligence" and "artificial consciousness". The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective,

> because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. "Synthetic phenomenology" (SP; a term coined in analogy to "synthetic biology") refers to the possibility of creating not only general intelligence, but also consciousness or subjective experiences on advanced artificial systems. Future artificial subjects of experience have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing negative states like suffering. One potential risk is to dramatically increase the overall amount of suffering the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.
>
> **Recommendation 7**
>
> The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements. This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness (for recent examples see Dehaene, Lau & Kouider 2017, Graziano 2017 and Kanai 2017).
>
> **Recommendation 8**
>
> Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund, and coordinate relevant interdisciplinary research projects (comprising philosophy, neuroscience, and computer science). Specific relevant topics are evidence-based conceptual, neurobiological, and computational models of conscious experience, self-awareness, and suffering.
>
> **Recommendation 9**
>
> On the level of foundational research there is a need to promote, fund, and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness, and subjectively experienced suffering.

Your question of a test for phenomenality is right on target, because this is what the applied ethics of AI needs. I have tried to isolate the to isolate the four central necessary conditions for suffering in some freely available papers (for example: Metzinger, 2013b, 2016): The C-condition (having a phenomenal model of reality), the PSM-condition (a self-model), the NV-condition (the ability to represent negative valences—for example via homeostatic cost functions folded into the self-model, representations of decreasing functional coherence or low levels of self-control), and the T-condition (transparency, Mother Nature's most evil trick: forcing organisms to *identify* with negatively valenced states). Nobody knows if they are sufficient, and we have no theory of consciousness. From this it follows that there

is an ethics of risk too: We should take great care to always err on the side of caution, and this principle holds for future AI systems as well as for animals. In any case, we should work hard at an evidence-based theory of suffering that is as hardware-independent as possible. We need such a theory, else in the mid-term we will be unable to move forward with AI in an ethically responsible way.

> “Recklessly creating artificial consciousness would carry a high risk of *increasing* the overall amount of suffering in the universe.”

But if you look at what “The First Postbiotic Philosopher” already said in 2009, we could also introduce a *much* stronger criterion for artificial persons who claim to have phenomenal states:

> The Metzinger Test for consciousness in nonbiological systems demands that a system not only claim to possess phenomenal experience and a genuine inward perspective but also comprehend and accept the theoretical problem of subjectivity, and that it demonstrate this by participating in a discussion on artificial consciousness. It has to put forward arguments of its own and convincingly defend its own theory of consciousness. (Metzinger, 2009a, p. 201-202)

In more recent work, you have refined your previous account of MPS, introducing the notion of the *phenomenal unit of identification* or UI for short (Metzinger, 2013c). You define the UI as the relatively invariant phenomenal property (or set of phenomenal properties) with which a given subject self-identifies at a given time, generating “the distinct experience of ‘I *am* this!’”. In ordinary cases, the target properties of self-identification would most likely be “the integrated contents of our current body image”, accompanied by “the subjective quality of ‘agency’ in the control of bodily actions” (p. 5), because we are embodied agents and identify with a body over which we have global control. In bodiless dreams and asomatic OBEs, however, not only do subjects lack the experience of identifying with a body (describing themselves as “pure consciousness”, “balls of light” or “points in space”), but they can also lack agency and motor control. Thus, in some altered states of consciousness, the UI can be something else than the experienced body image, namely either: (a) the experienced origin of the visuospatial perspective as an “extensionless point in space”, which you call the *minimal* UI—the simplest possible unit of identification; or (b) the unity of consciousness, or “awareness as such”, which you call the *maximal* UI—the most general phenomenal property available for identification. The latter, you speculate, might happen in some asomatic OBEs and in deep meditative states in which subjects self-identify with “pure consciousness”. Apparently against your original account, you conclude that MPS is constituted by self-identification with *at least* a minimal UI (and not necessarily with a body), which

merely requires spatiotemporal self-location. Can you explain to us how the introduction of the minimal UI concept has changed the original MPS proposal, and what its potential benefits may be for addressing empirical questions? For instance, one may wonder to what phenomenal property the minimal UI corresponds to in asomatic OBEs and bodiless dreams. Presumably, there is no special phenomenal property of being an "extensionless point in space": a disembodied experience may simply be an experience of a visual scene which lacks any bodily awareness. The assumption that there is an extra feeling of being a disembodied point in space is at least controversial -– subjects might describe their experiences in such a way simply because this is the easiest way to describe an experience of disembodiment. Do you think there is room for a more deflationary take on such experiences?

The original motivation was to describe more clearly *what* exactly it was that was manipulated in those early experiments trying to create full-body illusions. Very often misreported, they do *not* create classical OBEs in the strong sense of involving a perceptually impossible external perspective (Metzinger, 2005a, 2009b). The UI-concept is determined by the following set of empirical constraints:

- Explicit embodiment is not a necessary condition for the UI
- *Minimal* spatiotemporal self-location is a sufficient condition for the UI
- The UI and origin of visuospatial perspective can be dissociated
- The UI can be located outside of phenomenal body model
- The UI and the origin of visuospatial perspective can be dissociated
- The UI can be *smeared* in phenomenal space
- The UI can be *duplicated* in some subjects

I think that one the most important future research targets is that the identification dimension of MPS has to be analytically grounded in a computational model, and Jakub Limanowski has the merit of having come up with some of the best work in this nascent field. One thing that I was hoping was that the "unit of identification" could be a much clearer concept for computational modelers than "the pre-reflexive self" or something like this.

Likewise, the notion of maximal UI raises a couple of questions. By your definition, when the UI is maximal and equated with "awareness as such", it seems that no phenomenological distinction can remain between the subject herself and anything she might experience. While you suggest that certain states induced by meditation might be examples of such a maximal UI, other researchers have suggested that meditation can induce conscious states in which phenomenal self-consciousness is entirely missing (Ataria, Dor-Ziderman, & Berkovich-Ohana, 2015; Dor-Ziderman, Berkovich-Ohana, Glicksohn, & Goldstein, 2013). On the face of it, self-reports of such states are consistent with this claim (although they certainly must be interpreted with caution): "It was emptiness, as if the self fell out of the picture.

There was an experience but it had no address, it was not attached to a center or subject" (Dor-Ziderman et al., 2013, p. 6). (Incidentally, if reports of expert meditators should not be taken at face value, then the same skeptical point could be made about reports of bodiless dreams or asomatic OBEs.) What do you make of such reports (if you consider them reliable enough); do you think they indeed suggest a lack of MPS (i.e., phenomenal selfhood) altogether and thus contradict a notion of maximal UI according to which the subject would literally experience everything as being identical to *herself*?

Bodiless dreams or asomatic OBEs still have an epistemic agent model, for example a "seeing self" that can control its focus of visual attention. Phenomenologically, the UI will be the sense of effort going along with mental action: the phenomenology of identification latches onto this effortful sense of control.

Autophenomenological reports necessarily presuppose autobiographical memory. Autophenomenological reports about states of "non-dual awareness" also create a "performative self-contradiction": A performative self-contradiction arises when the propositional content of a statement contradicts the presuppositions of asserting it. If *you* weren't there, why do you have an autobiographical memory of the episode? If it was *timeless*, why do you know how long it lasted? If the was no *self-location in space*, why do you know where it happened? I have been thinking about this for quite a while, as I have a long-standing interest in states of this type. It may well be possible that many of the so-called "spiritual" people underestimate what they are talking about, at least if they were to take their own beliefs about such "zero-person perspective" episodes seriously. They have nothing to do with *you*, because *you* cannot directly cultivate them, *you* cannot even prevent them. If they appear, they have nothing to do with *you*. If that is correct, your nervous system may have already realized such states in the past and *you* do not know it. They are not even episodes, because if they are timeless there is a strong sense in which they have been there all along and pervade every moment of your mundane temporal experience. Conceptually, instantiating an EAM plus MPS clearly seems to be a presupposition for autophenomenological reports. So, I think what these advanced practioners reports must be some sort of hybrid state in which the autobiographical self-model must still have been "recording" as it were.

I think one of the strengths of the new conceptual instrument of an UI is that one can ask new questions more precisely: can the phenomenology of identification and the phenomenology of selfhood be dissociated?

- For example, could there be a maximal UI that is non-selfy? Do we know conceivable and empirically plausible phenomenologies of unification with the world as a whole, which are more like an "all-pervading emptiness that has awoken to itself", i.e. more on the Buddhist side than on the Advaita Vedanta

side?

- For example, could there be a minimal UI that is non-selfy? This could for example be a phenomenology of *hacceity*. Maybe, if we do the phenomenology seriously and properly, what we really have never is MPS, but only a conscious THIS-here-now. A haecceity is a non-qualitative property responsible for individuation and identity. A haecceity is not a bare particular in the sense of something underlying qualities. It is, rather, a non-qualitative property of a substance or thing: it is a "thisness" as opposed to a "whatness".

I think there very clearly is a phenomenology of numerical singularity, namely, the subjective experience of *mere particularity*. But if that is the case, are we perhaps misdescribing exactly this phenomenology as minimal phenomenal *selfhood*, when there really is no such thing as a self there? If the phenomenology is indeterminate, then all reports are necessarily theory-contaminated. If you think of your beautiful Dor-Ziderman quote above—would the subject ever have used the word "emptiness" if there hadn't been twenty-five centuries of Buddhist philosophy for which exactly this concept was absolutely central?

But the notion of an UI also allows you to describe empirical results in a more differentiated way. Robotic re-embodiment studies demonstrate that the UI can be dissociated when given two explicit body representations as candidates for subjective self-location (Aymerich-Franch, Petit, Ganesh, & Kheddar, 2016). But do we need to speak of two *selves* in such cases, and would that even be logically coherent?

My own empirical claim that if we were to apply 1PP methods to MPS, we would very likely get some statistical distribution of certain types of reports, of which I have just presented two classical examples. Autophenomenological reports cannot determine the metaphysical status of MPS, because, for example, you cannot decide between MPS" readings and "*haecceitas*"-readings. Phenomenal indeterminacy for MPS seems to be a fundamental epistemological problem, that is why I brought it up at your Oxford conference.

In *Being no one*, you argue that there are two ways in which a conscious system could lack the phenomenology of selfhood: (a) by having a phenomenal world-model without a phenomenal self-model, or (b) by having a "fully opaque" phenomenal self-model, i.e. a phenomenal "system-model" which seamlessly accesses all stages of its own information processes (Metzinger 2003, p. 565). You acknowledge that the first case probably applies to "many simple organisms on our planet", while the second case may loosely coincide with the Buddhist notion of "enlightenment", although it is unclear whether it is nomologically possible, at least for humans. Do you believe that either of these forms of "selflessness" can be (at least temporarily)

exemplified by human subjects, for instance in radically altered states of consciousness induced by psychopathologies or psychoactive drugs?

Absolutely. Full ego-dissolution plausibly occurs in serious cases of depersonalization disorder or organic brain diseases, and I can only recommend your own paper on causal etiologies based on pharmacological stimuli (Millière, 2017)—it is perhaps the best, most well-researched, and most careful discussion of the empirical literature out there. Possibility (b) also seems quite obviously something that has happened to human beings for millennia and in many different cultures. My own attempt to approach what, if I remember correctly, I have called "system consciousness" in BNO (as opposed to "self-consciousness") is of course highly dubious, because it is relative to a certain level of description and a specific functional analysis. The way I used the concepts of "transparency" and "opacity" was as properties of phenomenal representations and, as indicated above, such concepts are historically plastic entities. Nevertheless, if my central conceptual point—namely, that for conscious self-representations transparency necessarily leads to the phenomenology of identification—still holds, then it is obvious how this specific phenomenology can gradually be dissolved by leaving the content of the conscious self-model as it is. There could be many stages, for examples for whom MPS is still robust and fully transparent, but in which the phenomenology of agency on the mental level (that is, the cognitive and attentional EAM) has disappeared, because introspective attention has penetrated into the fine-grained functional mechanisms underlying it. But again, please note the functional analysis I have developed for opacity and de-identification rests on notions like "earlier processing stages" and "vehicle properties" versus "intentional properties". Especially the last two concepts might soon begin to look as artifacts of old-school armchair philosophizing—for example, I think we may perhaps find better conceptual tools in the predictive processing framework.

To add to the previous question, several authors—philosophers and scientists alike—have argued that in a predictive processing framework, the self-model results from active (Bayesian) inference and the brain's implied prediction error minimization about which sensory signals are "the most likely to be me" across exteroceptive, proprioceptive and interoceptive domains (Apps & Tsakiris, 2014; Limanowski & Blankenburg, 2013; Seth, 2013). On this view, the brain's self-model is just a special part of its world-model. In cases in which information processing is heavily disturbed (e.g. by a pharmacological agent), it may be the case that persistent prediction errors are transmitted to higher levels of the system's generative model, resulting in an update of normally very stable predictions regarding the self and world. For example, couldn't it be the case that the phenomenon known as "drug-induced ego dissolution", described as a (reversible) loss of self-awareness at high doses of psychedelic drugs such as LSD (Letheby & Gerrans, 2017; Millière, 2017), is best construed as a breakdown of the conscious self-model itself—resulting from

an (temporary) update of hyperpriors regarding the distinction between self and world? One might argue that this would be an instance in which the first way of being "selfless" mentioned above could temporarily apply to human beings. Put in terms of your view on (phenomenal) self-models, do you think there can be a re-instantiation of a PSM after its complete opacity (e.g., induced by drugs or pathological conditions), or must there always be a part of the self-model that is conscious and transparent? If so, would this part correspond to the minimal UI or might it also be conceived as the maximal UI? Finally, how much weight do you assign nonconscious self-models, the biophysical "grounding" of selfhood in bodily background processes, in such altered "selfless" states and the re-instantiation of a perceived "ego"?

Very interesting questions, much too deep for a short interview! First, "psychedelic" means "mind-manifesting" and one of the most intriguing aspects of such states is perhaps that it makes you prior- and hyperprior-landscape itself a potential object of manifest, explicit conscious experience, simply because this landscape becomes extremely flexible and malleable, highly context-sensitive. Second, these states of consciousness hold the potential to simply make normal people who haven't thought about all these things much very concretely and directly aware of the fact that it is *literally* true that conscious experience is a model. For many subjects, it is the first and only experience ever to approximate global opacity. Now, if that even happens on the level of self-consciousness, then it obviously is quite a dramatic affair, because, if you will, it leads to a Husserlian "bracketing" of the certainty of one's very own existence.

> "Psychedelics states hold the potential to simply make normal people very concretely and directly aware of the fact that it is *literally* true that conscious experience is a model."

About the philosophical problem of a "performative self-contradiction" as related to pharmacologically induced non-egoic states I simply have to say that I have no solution and am thinking about it. Probably the answer is that full blown dissolutions are not remembered (perhaps on the unconscious levels of the bodily self-model), and that everything that people report are just graded phenomenologies, slightly incomplete mystical experiences. Maybe the memory traces are also only constructed when *leaving* such states (remember Dennett's "cassette theory" of dreaming? see Dennett, 1976). In any case, I think multisensory integration leading to MPS and the bodily self-model is an automatic bottom-up process, and it may exactly be what "rescues" the subject in a scientific experiment with ego-dissolving psychoactive substances. At some point, the Here-Now-model

becomes so good that the veil of transparency drops and everything becomes real again.

On a related note, you have raised skeptical concerns regarding reports of alleged "selfless" conscious states, arguing that they "generate a performative self-contradiction" (Metzinger 2003, p. 566). Indeed, you write, "how can you coherently report about a selfless state of consciousness by referring to your own, autobiographical memory?" (Metzinger, 2005b, p. 23). Is it not conceptually possible that the brain may store a conscious experience lacking self-consciousness in episodic memory, and then retrieve the stored memory later in an illusionary autobiographical mode of presentation? In other words, the apparent contradiction in such reports might come from the structure of memory retrieval, rather than the memory itself. Furthermore, descriptions of "drug-induced ego dissolution", for instance, frequently underline the inadequacy of the first-person pronoun to report such experiences, although it is hard to avoid using it for grammatical reasons (e.g. "There existed no one, not even me… so would it be proper to still speak of 'I', even as the notion of 'I' seemed so palpably illusory?", Millière 2017, p. 14).

I have already touched upon this topic in a previous answer, but I think you are floating a very interesting idea here, namely that there could be different phenomenal data-formats and that sincere autophenomenological reports could refer to memories that have become available under an egocentric inner mode of presentation. Would they then be false memories? What I find most intriguing about your proposal is that if something like this can happen in a larger time-window, then it could also happen in a much shorter time frame. Perhaps all experience is originally selfless, and is transformed into first-person experience by continuously creating false memories of the type you describe, via ultrafast forms of "illusionary memory retrieval"?

> "As a philosopher, I am very skeptical about all this loose talk concerning 'first-person methods' and 'first-person data'."

You have argued that "first-person data do not exist" (Metzinger 2003, p. 591), because there is no scientific procedure to settle introspective disagreements. Furthermore, you have suggested that there might not even be any "empirical fact of the matter" regarding some phenomenological disputes, because of "the possibility of phenomenal indeterminacy" (Metzinger 2013, p. 4). On the other hand, your work frequently appeals to subjective reports of altered states of consciousness (e.g. dreaming, out-of-body experiences or thought insertion) and to what you call "paradigmatic autophenomenological reports". What epistemological status do you attribute to such reports? Do you endorse the view that part or totality of phenomenal consciousness is indeterminate?

A part certainly is, and indeterminacy is an important research target for the future. Total indeterminacy would be then end of all knowledge. I am a strange person: As a philosopher, I am very skeptical about all this loose talk concerning "first-person methods" and "first-person data", but as a consciousness researcher I have certainly tried many of these methods—probably even a bit more rigorously than all the people who publicly advertise them to promote ideological forms of anti-reductionism or for purposes of academic virtue signaling and reputation management. But it is exactly because I have done a bit of this in my personal life that I am very much aware of the risk of "theory-contaminated reports", to name just one example. Most of the people exploring altered states of consciousness have extremely strong motives and metaphysical background assumptions, they look for something, otherwise they would not have the courage or discipline it takes. 1PP-loose-talk certainly has a nice, politically correct ring to it (Diversity! No evil reductionism! FINALLY taking inner experience seriously! Everybody can claim what they have always wanted to claim!) and it is the best strategy to get applause from many different types of audience. Stressing the importance of first-person methods makes everybody believe you are a good person, it is good for your career. But as I have explained in publications, the whole concept of "data" is overextended here, the original usage refers to something very different. Second, from a philosophical perspective, the *really* interesting methods are "zero-person methods"—but they are extremely difficult to talk about in any coherent manner.

But of course, we can get very far by refining interview methods and simply taking the reports themselves as data, doing careful semantic evaluation and statistics. Reports, neural correlates, and computational models can get us much further than we may often believe.

References

Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*, 85–97. https://doi.org/10.1016/j.neubiorev.2013.01.029

Ataria, Y., Dor-Ziderman, Y., & Berkovich-Ohana, A. (2015). How does it feel to lack a sense of boundaries? A case study of a long-term mindfulness meditator. *Consciousness and Cognition*, *37*, 133–147. https://doi.org/10.1016/j.concog.2015.09.002

Aymerich-Franch, L., Petit, D., Ganesh, G., & Kheddar, A. (2016). The second me: Seeing the real body during humanoid robot embodiment produces an illusion of bi-location. *Consciousness and Cognition*, *46*, 99–109. https://doi.org/10.1016/j.concog.2016.09.017

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13*(1), 7–13. https://doi.org/10.1016/j.tics.2008.10.003

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. https://doi.org/10.1007/s11229-016-1239-1

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Dennett, D. C. (1976). Are Dreams Experiences? *The Philosophical Review*, *85*(2), 151–171. https://doi.org/10.2307/2183728

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Dor-Ziderman, Y., Berkovich-Ohana, A., Glicksohn, J., & Goldstein, A. (2013). Mindfulness-induced selflessness: a MEG neurophenomenological study. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00582

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Gallagher, S., & Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 1–22. https://doi.org/10.1007/s11229-016-1269-8

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. https://doi.org/10.1007/s11229-015-0762-9

Graziano, M. S. A. (2017). The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *Frontiers in Robotics and AI*, *4*. https://doi.org/10.3389/frobt.2017.00060

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Kanai, R. (2017). We Need Conscious Robots. How introspection and imagination make robots better. *Nautilus*, 47. http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Letheby, C., & Gerrans, P. (2017). Self unbound: ego dissolution in psychedelic experience. *Neuroscience of Consciousness*, *3*(1). https://doi.org/10.1093/nc/nix016

Limanowski, J., & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00547

Madary, M., & Metzinger, T. K. (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00003

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, Mass.: A Bradford Book.

Metzinger, T. (2005a). Out-of-Body Experiences as the Origin of the Concept of a 'Soul '. *Mind and Matter*, *3*(1), 57–84.

Metzinger, T. (2005b). Precis: being no-one. *Psyche*, 1–35.

Metzinger, T. (2008). Empirical Perspectives From the Self-Model Theory of Subjectivity: A Brief Summary with Examples. In R. Banerjee & B. K. Chakrabarti (Eds.), *Models of Brain and Mind: Physical, Computational, and Psychological Approaches*. Elsevier.

Metzinger, T. (2009a). *The EGO Tunnel: The Science of the Mind and the Myth of the Self* (1 edition). New York: Basic Books.

Metzinger, T. (2009b). Why are out-of-body experiences interesting for philosophers?: The theoretical relevance of OBE research. *Cortex*, *45*(2), 256–258. https://doi.org/10.1016/j.cortex.2008.09.004

Metzinger, T. (2013a). The Myth of Cognitive Agency: Subpersonal Thinking as a Cyclically Recurring Loss of Mental Autonomy. *Frontiers in Psychology*, *4*, 931.

Metzinger, T. (2013b). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.), *Robotik und Gesetzgebung* (pp. 247–286). Baden-Baden: Nomos.

Metzinger, T. (2013c). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Consciousness Research*, *4*, 746. https://doi.org/10.3389/fpsyg.2013.00746

Metzinger, T. (2014). First-order embodiment, second-order embodiment, third-order

embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In L. Shapiro (Ed.), *The Routledge Handbook of Embodied Cognition* (pp. 272–286). London: Routledge.

Metzinger, T. (2016). Suffering. In K. Almqvist & A. Hagg (Eds.), *The Return of Consciousness*. Stockholm: Axel and Margaret Ax:son Johnson Foundation.

Metzinger, T. (2017a). The Problem of Mental Action: Predictive Control without Sensory Sheets. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predicitive Processing*.

Metzinger, T. (2017b). Benevolent Artificial Anti-Natalism (BAAN). *EDGE Essay*. https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan

Metzinger, T. (2018). Why is Mind Wandering Interesting for Philosophers ? In K. C. R. Fox & K. Christoff (Eds.), *The Oxford Handbook of Spontaneous Thought: Mind-wandering, Creativity, Dreaming, and Clinical Conditions*. Oxford University Press.

Metzinger, T., & Gallese, V. (2003). The Emergence of a Shared Action Ontology: Building Blocks for a Theory. *Consciousness and Cognition*, *12*(4), 549–571.

Metzinger, T., & Wiese, W. (2017). *Philosophy and Predictive Processing*. MIND Group. Retrieved from https://predictive-mind.net/

Metzinger, T., & Windt, J. M. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Horvath, & J. Kipper (Eds.), *Die Experimentelle Philosophie in der Diskussion* (pp. 279–321). Berlin: Suhrkamp.

Millière, R. (2017). Looking For The Self: Phenomenology, Neurophysiology and Philosophical Significance of Drug-induced Ego Dissolution. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00245

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Weisberg, J. (2006). Consciousness Constrained: A Commentary on Being No One. *Psyche*, *12*(1).