The No-Self Alternative

Article · January 2011

DOI: 10.1093/oxfordhb/9780199548019.003.0012

1 author:

Thomas Metzinger
Johannes Gutenberg-Universität Mainz
91 PUBLICATIONS **3,001** CITATIONS

SEE PROFILE

CHAPTER 11

THE NO-SELF ALTERNATIVE

THOMAS METZINGER

WHENEVER popular or academic debates about 'the self' flare up again, we can often observe an embarrassing fact: Just because—quite obviously, and in many cultures—there is a folk-metaphysical and a folk-phenomenological *concept* of 'the self', and just because someone has put this concept back on the agenda, many participants automatically assume that an *entity* like 'the self' must actually exist and that a relevant and well-posed set of scientific and theoretical questions relates to this entity. However, there seems to be no empirical evidence and no truly convincing conceptual argument that supports the actual existence of 'a' self. Nothing forces us to make this assumption. Therefore, many debates of this type are threatened by a certain shallowness right from the very beginning and, due to their endorsement of an unwarranted existence assumption, run the risk of triviality. As it turns out, the 'no-self alternative' may not be an alternative at all—it could simply be the default assumption for all rational approaches to self-consciousness and subjectivity.

In the first part of this short chapter, I will differentiate a number of possible claims regarding the non-existence of 'the self' and will also try to at least sketch one typical argument for each thesis. In the second part, I will offer some new ideas on why all such arguments will always remain counterintuitive for many of us.

I am indebted to Jennifer M. Windt and Adrian J. Smith for a number of a number of helpful critical comments and for their help with the English version of this chapter.

1. Anti-Realism about 'the Self'

Ontological anti-realism about 'the self'

The standard way to be an anti-realist about selves is to deny that selves are *substances*:

(ARS$_0$) The self is not a substance.

A substance is an entity that can 'stand in existence' all by itself, even if all other existing entities were to disappear. It is 'ontologically self-subsistent', because it can sustain its own existence. It endures over time and is an ontologically *fundamental* entity, because it belongs to the basic building blocks of reality. Of course, there is a highly differentiated universe of metaphysical theories about 'substancehood'. But, given this simple working definition, the thesis of ARS$_0$ amounts to the claim that selves are not self-subsistent entities: they do not endure over time and they do not belong to the basic building blocks of reality.

There are two well-established ways in which one can argue for this sort of ontological anti-realism about 'the self'. The first strategy is to confine discussion to the target phenomenon of self-consciousness and related folk-phenomenological discourse, including its unwarranted metaphysical assumptions. The second strategy is to deny the existence of *all* substances or individuals on theoretical grounds, *tout court*. Here the obvious traditional example is the anti-substantialist metaphysics of Buddhist philosophy (see Siderits 2003, 2007, Chapter 12 below; for examples of recent discussions, cf. Albahari 2007; Westerhoff 2009). In addition, current debates in the theory of science—particularly, in the philosophy of physics—have long made it obvious that our standard views about individuals, intrinsic properties, and relations are obsolete. Instead, an ontic form of *structural realism* seems to be the most promising candidate in modern metaphysics, proposing that relations are all that exist. Said relations do not hold between objects with intrinsic properties, but even the relata themselves can be decomposed into sets of relations, with structure always being more ontologically fundamental than objects or 'substances' (for a paradigm example, see Ladyman and Ross 2009). In a certain sense, this would lead us into an 'unfounded universe', because the ultimate reality would then be the nomological structure of the world, where even the identity and individuality of objects would depend upon this (relational) structure.

Just as current physics is fully compatible with the view that quantum-objects are non-individuals, the corresponding metaphysical underdetermination of cognitive neuroscience certainly makes it tenable that mental, psychological, or phenomenological entities like 'the self' are not proper individuals at all: they simply possess no clearly specifiable identity criteria. In some contexts, classical objects, as well as 'selves', are useful heuristic posits, but nonetheless ontologically dispensable entities. If this is correct, we should rather search for an alternative metaphysics that renders the predictive success of our empirical theories intelligible without committing us to theoretical terms referring to individual entities in the world (like 'the self'), or to the truth of certain statements in which these terms appear.

Let us now confine discussion to a sketch of the *first* strategy mentioned above, considering only the target phenomenon of self-consciousness and the related folk-phenomenological discourse. Why should we assume the existence of selves in the first place? The main problem, according to ARS_0, is that in our widespread and naively realistic manner of speaking about 'the self' we introduce an individual, making an existence assumption which typically is not backed up by independent argument.

What are selves? If not 'nuggets of reality', what else could they be? Selves could be unobservable entities, perhaps conceptual fictions—although in an interesting sense they are also phenomenological 'everyday objects'. As robust elements of the phenomenal ontology applied by our brains (cf. Metzinger and Gallese 2003), they play a leading role in the everyday phenomenology of our experiential *Lebenswelt* and the pre-scientific folk-psychology it gives rise to. How are we to interpret this fact on the level of philosophical metaphysics? Let us briefly look at the three main theoretical options.

If we wanted to establish selves as proper individuals, our first option could be to take a sober and straightforward approach: we could try to be old-fashioned scientific realists, and, for example, identify them with their position in space and time, turning them into single, countable entities by means of the physical properties they share with no other physical object. With Leibniz (1646–1716) we would then say that individuality simply amounts to distinguishability, adopting the Principle of the Identity of Indiscernables. However, only *bodies* can be fully individuated in this way—but selves are not to be taken as bodies or biological organisms *simpliciter*.

If we wanted to find a metaphysical representation of our pre-scientific, phenomenological intuitions about selfhood, then the second theoretical option would be to posit some special sort of *haecceitas* in the sense of Duns Scotus (1266–1308): a property of primitive 'thisness' or self-identity, a transcendent property of selves that grounds their intrinsic individuality as well as their numerical identity. For example, there could be a uniquely instantiable, non-qualitative property of (say) being identical with Shaun Gallagher, responsible for the irreducible individuality and the numerical identity of the very person who is the editor of this *Handbook*. The haecceity approach has a certain prima facie appeal, because it may actually capture something phenomenologically important, whereas, metaphysically, it is less convincing. On the one hand, introducing a 'primitive thisness' constituting individual substances is, of course, a pure hypostatization. It is a philosophical move that does not explain anything but just introduces a further, unobservable property without argument or potential empirical evidence. On the other hand, I believe that we may have overlooked an important phenomenological fact, one that could usefully serve as a constraint on a more comprehensive theoretical account. There is a distinct *phenomenology of singularity*, a non-sensory phenomenology of 'thisness'—for example, in the phenomenology of meditation, but also in bodily self-consciousness. If we look closely enough, we can discover the phenomenology of primitive 'thisness' in our own subjective experience. It is particularly distinct in certain non-conceptual layers of self-awareness (their content has a non-perceptual but nevertheless 'demonstrative' character), and it certainly is a feature that requires careful attention in the phenomenology of self-consciousness. But phenomenological structure *per se* will never determine metaphysics.

The third option then would be to develop a theory of 'the self' as a mere collection of properties. A typical example of anti-substantialist and anti-individualist approaches to 'the self' are so-called 'bundle theories'. I have already mentioned Buddhist philosophy above; the most prominent Western representative, perhaps, is David Hume (1711–76). Hume would have said that we typically solve the conflict between experienced sameness across time and the succession of change, between the phenomenology of identity and that of diversity, by conjuring up a substance: 'the imagination is apt to feign something unknown and invisible, which it supposes to continue the same under all these variations; and this unintelligible something it calls a substance, or original and first matter' (T 220). Conceptually, bundle-theorists analyze substances as some sort of, possibly complex, relation between properties (which, however, could in turn either be conceived of as universals or as tropes, i.e. individual property instantiations). Selves would just be collections of properties.

The open question now becomes what principle *exactly* is responsible for the establishment of the complex relation just mentioned—what turns all the features comprising the self into a coherent whole? Empirically, one could say that selves (or other entities previously described as substances) are just collections of properties, which, as a matter of fact, we happen to mentally represent *as* individual entities. From the perspective of present-day cognitive neuroscience, this would be a scientifically plausible strategy: Our brains segment scenes and constitute multimodal, consciously perceived perceptual objects (e.g. one's own body as a whole) not by attaching properties to some more basic entity, but by a dynamic, bottom-up process of self-organization called 'feature-binding' (see e.g. Singer and Gray 1995). All technical details aside, what is new today is that science offers conceptually clear models of functional mechanisms which could parsimoniously explain the *integration* of individual property-representations into a unified self-representation. This theoretical model requires no transcendental subject to stand behind the appearance of 'a' self as consciously represented, because it gradually emerges out of the self-organizing interaction between a large number of simpler components. This possibility—the appearance of ordered structures without external interaction or a well-defined and highly specific initial state—simply was not available to thinkers in the past; it is a novelty in the history of ideas. Therefore, dynamical self-organization is a new theoretical option for the bundle theorist, in metaphysics as well as in phenomenology (see Metzinger 1995, 2003; Parfit 1982 and Chapter 18; Thompson 2007).

To sum up, the general point about substances, individuals, and identity criteria is that none of the currently available scientific data determine our metaphysics in such a way as to make the assumption of the existence of 'a self' necessary. Moreover, the principle of parsimony demands that we try to find a *simpler* metaphysical representation of our current knowledge about self-consciousness, of its causal history and its constitutive conditions, than what we have traditionally assumed to be the self-as-substance.

Before proceeding to a short sketch of epistemological, methodological, and semantic variations on the no-self alternative, let us briefly pause to look at the wider context in which such discussions take place. The first strategy mentioned above often emphasizes simply that no empirical evidence could ever ground a substantialist metaphysics of selfhood. This more modest approach would refrain from making any general metaphysical claims about the possibility of substances *per se*, but would only demonstrate the irrationality of positing a special sort of *individual* substance, namely, 'the self'. Although our very own naturally evolved cognitive structures (our inbuilt 'naïve physics') almost seem to make it a functional necessity for us to use substance-concepts like the notions of an enduring particular or an individual substance, which then work as carriers of properties, nothing in the brain or the self-conscious biological organism as a whole could even remotely count as a substance in any philosophically interesting sense. We just don't find a substantial self anywhere in the world and nothing on the level of scientific facts determines our metaphysics in this way.

What we do find, however, is the *phenomenology* of substantiality, on the level of introspective experience: subjectively, we often experience ourselves exactly as self-subsistent, enduring entities forming non-exchangeable and irreducible parts of reality. Moreover, the deeper core of our theoretical problem might lie in the more subtle fact that our phenomenal experience of selfhood not only expresses an aspect of 'reality' (i.e. the factual *realness* of the self), but also an aspect of 'metaphysical necessity' (i.e. the *impossibility* of non-existence, across all conceivable scenarios). I will return to these points below, and for two reasons. First, all deflationary or so-called 'weaker' theories of the self (for examples see Ghin 2005; Legrand 2005) typically miss the mark, by failing to explain what could ground the phenomenology of the self-as-substance and what the causal history or the biological function of this illusion of substantiality could have been. Second, the phenomenal self is *the* proto-object as such. If anything grounds our naive-realistic world-view that reality is composed out of individual substances possessing intrinsic, context-invariant properties and standing in certain relations to each other, it is exactly the phenomenology of selfhood. Cognitively, the conscious experience of selfhood leads directly to the metaphysical prototype of 'objecthood' and to the idea of an *individual* substance. This observation implies the interesting conclusion that many of our irresistible theoretical intuitions about substancehood are ultimately anchored in the conscious experience of selfhood. With this wider context in mind, let us now return to our brief sketch of theoretical options.

Epistemic anti-realism about 'the self'

For those who cannot resist the intuition that individual substances like selves exist, the obvious move will be to posit the existence of unknowable individuals possessing an unknowable intrinsic nature:

(ARS_E) The self is part of an unknowable realm of individuals, possessing an unknowable intrinsic nature.

Just as with the unknowable nature of the Kantian *Ding an sich* (thing in itself), we could posit an unknowable self pulling the strings behind the observable behavior of self-conscious agents and underlying the introspective phenomenology of selfhood. Consequently, all we could ever know would be the *structure* of the self—for example, its form of interaction with other selves and the laws and regularities guiding its cognitive and bodily behavior. Its *nature* (the sort of entity it really *is*) would remain epistemically inaccessible to us. Epistemological anti-realism treats the self as an unobservable entity, but derives no specific metaphysical position from its central claim. We can view it as a form of agnosticism.

However, there is at least one specific problem, which arises in the context of 'selves' as objects of knowledge. If they really are unknowable individual substances or have an unknowable intrinsic nature, then self-consciousness can no longer be seen as a process that provides us with a direct and epistemologically relevant form of *acquaintance* with ourselves. On this view, substantive forms of self-knowledge are no longer possible: in introspection and in phenomenal self-awareness, we never grasp our own true nature—it may well be that we have an essence, but this essence will forever remain inaccessible to us. Therefore, the distinct, characteristic, and Cartesian *phenomenology of certainty* which accompanies self-consciousness is an illusion. Furthermore, as the phenomenology of certainty is not about the existence of some merely objective, historical person, but about the indubitable ontological status of oneself *as self*,[1] epistemic anti-realism renders the self not only epistemologically irrelevant, but also leaves us with no further metaphysical issue to be resolved.

[1] This type of phenomenological first-person certainty is intimately related to the capacity of thinking Cartesian I*-thoughts of the form [I am certain that I* myself exist]. In other words, Descartes claimed that he was certain that he* (he *himself*) existed, not that he was certain that Descartes existed. Lynne Baker has made this point very clear in Baker 1998, 2007.

Methodological anti-realism about 'the self'

It is perhaps on this level that we find the most straightforward and convincing argument for the elimination of the concept of 'a' self:

(ARS_M) Nothing in the scientific investigation of self-consciousness commits us to assume the existence of individual selves.

As already noted in my discussion of ontological anti-realism above, there are no strong empirical data whatsoever supporting the existence of selves. There are first-person reports, and as such they may function as data-points, but of course there are no first-person *data* in any more rigorous sense.[2] More importantly, the process of generating and testing new hypotheses in empirical research programs investigating self-consciousness, agency, social cognition, etc. simply does not *require* the assumption of a theoretical entity by the name of 'the self'. Science can achieve its predictive success, describe and explain the available data, and integrate them into a larger evolutionary or neuroscientific framework without assuming that there is a mysterious thing called 'the self' which is represented in self-consciousness, initiates actions, or engages in social cognition related to other mysterious individuals called 'selves'. Prediction, testing, and explanation can take place in a much more parsimonious conceptual framework, for instance by introducing the concept of a 'transparent self-model' (Metzinger 2003*a*, 2006, 2008, 2009).

Semantic anti-realism about 'the self'

We refer to ourselves using the word 'I'. But what, exactly, is the meaning of the linguistic expression 'I'? If—as seems obvious—it doesn't *refer* to a specific part of reality in an object-mode of reference, what exactly is its logical or semantic

[2] Seriously assuming the existence of 'first-person data' rests on an extended usage of a concept that is well defined in another (namely, scientific) context. First, the whole concept of a 'first-person perspective' is just a visuo-grammatical metaphor, without a theory to back it up—currently, we simply don't know what that could be, 'a' first-person perspective (for a first conceptual differentiation, see Blanke and Metzinger 2009). Second, 'data' are extracted from the physical world by *technical* measuring devices, in a *public procedure*, which is well-defined and well-understood, replicable, and improvable; and which is necessarily *intersubjective*. Therefore, speaking of 'first-person data' would rest on an extended usage of a concept which is only well-defined in another context of application. 'Data' are typically (though not always) gathered with the help of technical measuring devices (and not individual brains) and by groups of people who mutually control and criticize each other's methods of data gathering (namely, by large scientific communities). In particular, data are gathered in the context of rational theories aiming at ever better predictions, theories that—as opposed to phenomenological reports—are capable of falsification. Autophenomenological reports themselves can be treated as data, the experience itself cannot (for a dissenting view, cf. Thompson 2007: 474 and 338 n. 10). All of this is not to deny that what are sometimes called 'first-person methods' could have an enormous impact on our way towards a rigorous, empirically based theory of self-consciousness.

function? In order to be semantic realists about 'I', we would have to assume that it reliably connects us to an irreducible, elementary aspect of reality. Semantic anti-realism denies this:

(ARS$_S$) The indexical expression 'I' does not refer to any entity that is ontologically fundamental.

A lot of excellent philosophical work has been done with regard to the semantics of the indexical term 'I' (see e.g. Boër and Lycan 1980; Castañeda 1966, 1967; Perry 1979, 1993; Recanati 2007). Does 'I' refer to some sort of invisible object, like some other linguistic expressions do? Does 'I' refer to a Cartesian Ego? Does it perhaps refer to some sort of 'objective self' (see Nagel 1986: ch. 4), because sentences like 'I am Thomas Metzinger' always possess a second reading, a second set of truth-conditions superseding the trivial, purely objective self-identification with a particular, historical person? Is the object of reference for 'I' always a *person*, that is, an entity which is ontologically basic and which simultaneously possesses mental and physical properties (Strawson 1959)? But how, then, do we explain the particular structure of self-consciousness—is it perhaps best characterized by a specific *mode of presentation*, perhaps an 'EGO-mode' of internally presenting information (Newen 1997; Perry 1979; Recanati 1993)? Or does 'I' simply not refer at all—at least not to a 'self', but simply to an organism that has gained knowledge about itself (Anscombe 1975)?

Clearly there is a distinct, fourth strategy the anti-realist about 'the self' can adopt: she can investigate the semantics and the logical deep structure of *linguistic* acts of self-reference. Employing her prescribed theory of meaning, she can then deny the referential nature of our uses of 'I'. Yet again, every single act of linguistic self-reference is accompanied by a distinct phenomenology, which, although it locally supervenes on functional properties of our central nervous systems, transports exactly those intuitions that turn us into naive realists and make us doubt any anti-realism about 'the self'. In uttering an 'I*'-sentence, we often actually *do* have the feeling of directly referring to something deep and real, to an invariant and substantial core of our own being. If anything could be a paradigm case for direct acquaintance or immediate epistemic access, then it is the phenomenology of non-conceptual self-representation, in which such linguistic (or even purely cognitive) acts of seemingly 'direct' self-reference are anchored: we feel infinitely close to ourselves. And any philosophical anti-realism about 'the self' can only succeed if it complements its theoretical strategy with a convincing account of the phenomenology that gives rise to our Cartesian intuitions, to our enduring feeling that some sort of self simply *must* exist.

Therefore, having briefly sketched the four main variants of the no-self alternative and some of the arguments supporting them, we must now take a closer look at the role of intuitions and their relationship to the metaphysics of selfhood. This will be done in the second, and last, part of this chapter.

Why is Anti-Realism about 'the Self' Counterintuitive?

Intuitiveness

Intuitiveness is a property of theoretical claims or arguments, relative to a class of representational systems exhibiting a specific functional architecture. Conscious human beings are one example of such a class. The brains of human beings are naturally evolved information-processing systems, and when engaging in explicit, high-level cognition they use specific representational formats and employ characteristic styles of processing. Whenever we try to comprehend a certain theory, an argument or a specific philosophical claim, our brains construct a *mental model* of this theory, argument, or claim (Johnson-Laird 1983, 2008; Knauff 2009). This mostly automatic process of constructing mental models of theories possesses a phenomenology of its own: some theories just 'feel right', because they elicit subtle visceral and emotional responses, some claims 'come easily' and some arguments (including implicit assumptions they make in their premises) seem 'just plain natural'.

There are two overarching reasons for this well-known fact. First, such theories exhibit a high degree of 'goodness of fit', in regard to our network of explicit prior convictions, and further, in optimally satisfying the constraints provided by our conscious and unconscious models of reality as a whole. These latter represent both the totality of the knowledge we have acquired during our lifetime, as well as certain assumptions about the causal deep structure of the world that proved functionally adequate for our biological ancestors. Theories that immediately feel good because they are characterized by a high degree of intuitiveness maximize internal harmony. What we introspectively detect is a high degree of consistency, but in a non-linguistic, subsymbolic medium. Therefore we could also replace the term 'intuitiveness' by a notion like 'intuitive soundness' or 'introspectively detected consistency'. In principle it ought to be possible to spell this point out on a mathematical level, by describing the underlying neural computations and their properties in a connectionist framework, or utilizing the conceptual tools provided by dynamical systems theory.

The second major causal factor underlying the conscious experience of 'intuitive soundness' simply is the amount of energy it takes to activate and sustain a mental model of a given theory, plus the amount of energy it would take to permanently *integrate* it into our pre-existing model of reality. Our mental space of intuitive plausibility can be described as an energy landscape: claims that 'come easily' do so because they allow us to reach a stable state quickly and easily; theories that 'feel good' are theories that can be appropriated without a high demand of energy. Theories that *don't* feel good have the opposite characteristics: they 'don't add up', they 'just don't compute', because they endanger our internal harmony and functional coherence, and it would take a lot of energy to permanently integrate them into our overall mental model of reality. They are costly.

Obviously, the no-self hypothesis simply is *the* paradigm example of a theory that just 'doesn't compute' for beings like us. However, what does or does not compute is a contingent fact determined by the functional architecture of our brain, shaped by millions of years of biological evolution on this planet, and—to a much lesser degree—by our individual cognitive history and a given cultural/linguistic context. The phenomenology of intuitive soundness—the fact that some arguments seem 'just natural'—ultimately is a biological phenomenon, with a short cultural history supporting it as well. However, the inner landscape of our space of intuitive plausibility is not only contingent on our evolutionary history and on certain physical and functional properties of our brains—it was optimized for *functional adequacy* only. It serves to sustain an organism's coherence and physical existence, but this does not mean that the *content* of intuitions is epistemically justified in any way. The no-self hypothesis therefore becomes the paradigm example of an almost insurmountable obstacle, a major challenge to our intellectual honesty: It demands that we investigate a claim even if it contradicts our deepest intuitions, something that somehow is 'just too radical', way too costly, likely to be self-damaging, and cries out for a more moderate, weaker version because it just 'doesn't compute'. The no-self alternative is a theoretical construct, which nicely exemplifies the point made at the very beginning of this section: 'intuitiveness' and counterintuitiveness are *functional properties* that depend on the way in which a theoretical construct is encoded in a given class of systems with a given internal structure. We also see how any philosophical methodology that just tries to make our 'deepest intuitions' explicit in a conceptually coherent way immediately turns into a rather trivial enterprise. At best, it just charts our intuition space; at worst, it confuses failures of imagination with insights into conceptual necessity ('philosopher's syndrome', according to Dennett 1991: 401).

Maybe there is another, perhaps even better, way to put the point. Philosophers sometimes like to speak of sets of 'logically possible worlds', of 'nomologically possible worlds', or even of 'metaphysically possible worlds', and then proceed to investigate relations of epistemic access between such sets of worlds. We could refine our definition of intuitive soundness by introducing the notion of a 'phenomenologically possible world': every world that can be mentally simulated on the level of conscious experience by a given class of systems realizing a specific set of functional constraints is a phenomenologically possible world. For example, some worlds are close neighbors to our current perceptual world-model, others are more distant from it (but still are possible conscious experiences, representing something that could, phenomenologically, *be the case*), and still others are strictly phenomenologically impossible. For certain conscious cognitive agents, some theories will be false—will feel false—in all possible phenomenal worlds. Here is the central point: the no-self hypothesis describes an impossible scenario, because it cannot be *actively* simulated and embedded—the property of cognitive agency itself would have to be dissolved (see the next section). For beings like us, it is therefore not a phenomenologically possible world. But of course 'phenomenal possibility' is relative to the functional architecture that a certain class of systems contingently happens to have.

There are not many interesting relations of epistemic access pointing from the space of intuitive possibility to other sets of possible worlds. Epistemologically speaking, 'phenomenal possibility' is a rather irrelevant notion: logically inconsistent scenarios could always *appear* as possible phenomenal experiences to certain cognitive systems; nomological necessities could be encoded as contingent (because there is at least one possible phenomenal world in which, for this kind of system, they are false); a claim that turns out to be false in the actual world could be simulated as true in all possible phenomenal worlds, as metaphysically necessary, and so on.

The counterintuitive nature of the no-self alternative

We should now ask, 'Why is anti-realism about 'the self' counterintuitive?' It would be good to have an empirically grounded answer to the following question: what, in systems like us, determines the *phenomenology of actual existence* and what determines the *phenomenology of metaphysical necessity*? Let me explain. If we want to know why anti-realism about the self is counterintuitive, it would be helpful to understand how realism about the self comes to be intuitive in the first place. It would be good to understand how something must be experienced as 'definitely real', and sometimes also as necessarily existing.

First, we must clearly separate the phenomenology of 'realness' and the phenomenology of 'metaphysical necessity'. One phenomenological constraint that clearly has been neglected too much in most current work on consciousness is the 'realness constraint'. One and the same phenomenal content—a visually perceived object, or, say, the bodily self—can appear as more or less real, while the content itself remains stable: the content of subjective experience can vary along a dimension of realness. There are well-documented psychiatric conditions like derealization and depersonalization in which the core of the problem is that either the environment or the content of the patient's self-model appear as less and less real (Hunter *et al.* 2004; Simeon and Abugel 2006). Different parts of phantom limbs in congenital aplasia appear as more or less real depending on the respective subregion of the body-model, and all of them are less real than representations of actually existing body-parts, and so on. For example, a recent case study by Brugger and colleagues introduced a vividness rating on a seven-point scale that showed highly consistent judgments across sessions for their subject AZ, a 44-year-old university-educated woman born without forearms and legs. For as long as she remembers, she has experienced mental images of forearms (including fingers) and legs (with feet and first and fifth toes)—but these were not *as* realistic as the content of her non-hallucinatory self-model (cf. Brugger *et al.* 2000: 6168; Metzinger 2008).

Then we have conditions during intense religious experiences or under the influence of psychoactive substances where the specific content of phenomenal experience is held constant while everything becomes *more* real than normal. On the representational level of analysis, the phenomenology of realness can be captured by the intensity constraint for presentational content (Metzinger 2003*a*: section 3.2.9) and by the transparency constraint (ibid., section 3.2.7). On the functional and neuroscientific levels of description, plausible correlates are sheer stimulus strength, the general level of arousal, and the temporal coherence of neural responses.

Then there is the phenomenology of 'metaphysical necessity'. It means that we experience something—for example, the self—not only as real, but as something *of which it is not possible that it could not exist.* To consciously represent something as 'metaphysically necessary' (in this purely phenomenological sense) means that it is not possible that the thing in question could *not* be real. Phenomenologically, there is not a single possible world in which this could actually be the case, actual non-existence is inconceivable, only centered worlds exist. 'Inconceivability' here means 'phenomenal impossibility' as explained above. We confuse the apparent phenomenal necessity of the self with metaphysical necessity.

Of course, we can all construct conscious mental models of worlds in which we ourselves do not exist—for instance, because we have not yet been born or because we have already died. But what the no-self hypothesis demands is something else: In order to appear as intuitively sound, we should be able to conceive of the *actual* world—i.e. *given* the ongoing phenomenology of substantial selfhood, of ontological self-subsistence, and transtemporal identity—minus our own existence. We should be able to imagine a selfless world, but *from a first-person perspective.* For beings like ourselves, however, this is at least a functional impossibility: we cannot mentally simulate the conscious experience of a world in which we are not present as selves, but which we nevertheless experience from a first-person perspective. This fact is a contingent fact about the causal structure of our brains, about the functional architecture underlying our conscious minds. Here, I will make no stronger claim to interpret this fact in terms of conceptual necessities or possibilities—perhaps selfless first-person perspectives can be coherently described. The intriguing question, lying outside the scope of this entry, is whether there could be *epistemic* first-person perspectives without *phenomenal* selves as constituting factors.

So why does anti-realism about 'the' self seem counterintuitive? After having clearly separated the phenomenology of 'realness' and the phenomenology of 'metaphysical necessity' we can now proceed to the core of the problem: conscious mental agency. A lot of empirical research demonstrates how the experience of global control is intimately linked to the experience of selfhood. However, global control connotes not only bodily agency but also the focus of attention and (in some cases) the focus of high-level, conceptually mediated cognition. A completely selfless mental model of reality cannot be *actively* simulated, because the property of cognitive agency itself would then have to be dissolved. Reality-models in which the no-self alternative is true are of course possible, but what is not possible is a first-person phenomenal simulation of such a world. In particular, the 'first-personness' of the simulation itself, the ongoing activity of the cognitive self, could not be mapped onto such a world.

We can illustrate the point by briefly looking at the fallacy underlying Thomas Nagel's beautiful but failed argument for the existence of an 'objective self' (Nagel 1986: ch. 4). What Thomas Nagel terms the *objective self* is a conceptual reification of an ongoing misrepresentational process. This process takes place within a perspectivally structured model of reality, in the conscious mind of his readers experimenting with the *View from Nowhere*. This is what happens to you when reading Nagel: propositional input activates a chain of phenomenal mental models in your brain. As a cognitive agent, you try to understand his argument by constructing a mental model. In particular, you now simulate a non-centered reality *within* a centered model of reality. Under Nagel's description, this non-centered 'conception' of the world also contains all experiences and the perspective of Thomas Nagel as well:

> Essentially I have no particular point of view at all, but apprehend the world as centerless. As it happens, I ordinarily view the world from a certain vantage point, using the eyes, the person, and the daily life of TN as a kind of window. But the experiences and the perspective of TN with which I am directly presented are not the point of view of the true self, for the true self has no point of view and includes in its conception of the centerless world TN and his perspective among the contents of that world. (Nagel 1986: 61)

But this description is false: *this* inner experience, the current *View from Nowhere* as initiated and executed by the cognitive agent TN is *not* contained in the 'centerless conception of the world'. The *last* phenomenal event—namely, the intended shift in perspective—is not contained in the centerless conception, because this would lead to an infinite regress. However, it must obviously be contained in Nagel's autobiographical self-model—otherwise it would not be reportable. The current perspective is not part of reality as non-perspectivally seen by the true self Nagel postulates. Logically speaking, the threat of infinite regress is blocked by an object-formation, by introducing a metaphysical entity—the objective self. Technically, the fallacy is an act—object equivocation: What we have is not a thing, but a process.

Here is what *really* happens. A conscious, self-modeling system internally simulates a non-centered reality. This simulation is opaque (i.e. phenomenally experienced *as* a form of mental content), and it is embedded in the currently active phenomenal self-model: at any time you know that this is only a thought-experiment, and you know that *you* are carrying it through. Anything else would either be a manifest daydream or a full-blown mystical experience—and this certainly is not the phenomenology Nagel describes. In this phenomenally simulated reality there is a model of a person, TN (or, respectively, yourself), enriched by all the properties until then only known under the phenomenal self-model, as your *own* properties. This person-model forms the object-component of your consciously experienced first-person perspective; it is part of a comprehensive simulational process. Following Nagel's implicit instruction, you generate the simulation of an 'inner third-person perspective', namely by forming a model of yourself, which is *not* a self-model, but the model of yourself as if you were only given to yourself through indirect, external sources of knowledge. It is a model of a person alone in oceans of space and time, '*a momentary blip on the cosmic TV-screen*' (Nagel 1986: 61).

This process is fully reversible: In a second step you can now *reintegrate* the simulated person with the transparent partition of your phenomenal self-model, which, of course, has been there all along. Like a monkey or a dolphin recognizing himself in a mirror, as it were, you discover yourself in the *internal* mirror of your ongoing phenomenal simulation of a centerless world, by discovering a strong structural isomorphism to one of the persons contained in this world. To this representational event you can linguistically refer by exclaiming sentences of the type 'I am TN!' in their second, 'philosophical' reading. But there is no second set of truth-conditions, there is no additional 'perspectival fact' to be expressed,[3] and

[3] There are a number of problems here, if one looks more closely. First, the logical structure of the alleged perspectival fact is never clearly stated (see Lycan 1987: 78; 1996: 50; Metzinger 1993: 233; 2003). Second, the *objective self*—which is more similar to Husserl's notion of a 'transcendental ego' in his later philosophy than to the Wittgensteinian subject as forming the border of the world—in being used in Nagel's ubiquitous visual metaphor of the 'taking' of perspectives immediately creates distal objects as its counterparts. In its conceptual interpretation this then leads to persisting act–object equivocations, to the freezing of phenomenal *events* into irreducible phenomenal individuals. Again, see Lycan 1987: 79; 1996: 51). Thirdly, at a closer look, Nagel's concept of an 'objective self' is inconsistent. It is not a mental object any more, because the concept of mentality was *introduced* via the notion of a perspective, as referring to subjective points of view and their modifications (see Nagel 1986: 37). 'Self', however, is a mentalistic term *par excellence.* Norman Malcolm has pointed out how an aperspectival objective self would be a 'mindless thing' because in its striving for objectivity it would have distanced itself so radically from the point of view of the *psychological* subject that now it could itself not be grasped by any mental concept any more. See Malcolm 1988: 158. The most important mistake, however, consists in using 'I' as a designator and not as an indicator in the 'philosophical' reading of the relevant identity-statements. There are no criteria of identity offered for the individual in question. As Malcolm (1988: 154 and 159) puts it: 'Does this make any sense? *It would if there were criteria of identity for an I.* [emphasis TM] . . . When we are uncertain about the identity of a person, sometimes we succeed in determining his identity, sometimes we make mistakes. But in regard to the identity of an *I* that supposedly occupies the point of view of a person, we could be neither right nor wrong. After a bout of severe amnesia Nagel might be able to identify himself as TN—but not as *I.* "I am TN" could announce a discovery—but not "I am I". . . . An important source

there is no homunculus that was briefly united with the transcendental ego (Nagel's *objective self*) and is now hurled back into the empirical subject.

The *View from Nowhere* is a phenomenal simulation, but one in which the phenomenal properties of selfhood and cognitive agency never disappear—it is only a thought experiment carried out by a cognitive agent, always accompanied by an intact phenomenal self comfortably seated in the proverbial armchair. Therefore, it generates only a 'small' version of the *View from Nowhere.* The 'big' version would be a full-blown mystical experience, a phenomenal model of reality in which the property of selfhood is instantiated in no form whatsoever. We might interestingly view this type of conscious reality-model as the counterpart of the no-self hypothesis on the level of the brain's internal ontology. In conceptually interpreting Nagel's 'small' armchair version of the *View from Nowhere*, realism about the purported objective self creeps in at the very moment one forgets about the processuality (i.e. the event-character and the sustained agency-component of the ongoing phenomenal simulations described), and the phenomenal opacity (i.e. the subjectively experienced *representational* nature characterizing the overall process).[4]

Conclusion

The no-self alternative comes in a variety of different flavours and strengths, each being supported by a number of ontological, epistemological, methodological, and semantic arguments. One may well view it as the default assumption for all rational, data-driven approaches to self-consciousness and subjectivity. Its central problem is its radically counterintuitive nature.

Intuitiveness is a property of theoretical claims or arguments, relative to a class of representational systems exhibiting a specific functional architecture. Our intuition that 'the self' *exists* and that it *must* exist therefore is grounded in three different factors. First, it is a functional feature of our own mental architecture that any attempt to consciously simulate a world in which the no-self hypothesis turns

of confusion in Nagel's thinking is his assumption that the word 'I' is used by each speaker, to *refer* to, to *designate—something.* But that is not how "I" is used. If it were, then "I am I" might be false, because "I" in these two occurrences had been used to refer to different things. Nagel's statement, "I am TN", could also be false, not because the speaker was not TN, but because Nagel had mistakenly used "I" to refer to the wrong thing. If Nagel had not assumed that "I" is used, like a name, to designate something, he would not have had the notion that in each person there dwells an *I* or *Self* or *Subject*—which uses that person as its point of viewing.'

[4] For a concise definition of the terms 'phenomenal opacity' and 'phenomenal transparency', plus their role in cognitive self-reference, cf. Metzinger 2●●3*b*.

out to be true generates the phenomenology of mental agency, cognitive control, and therefore selfhood. Second, in describing this phenomenology, we are prone to act–object equivocations, because we turn a process into a thing, hypostatizing a phenomenal individual where there is only an intermittent chain of events, using 'I' as a designator where it only is an indicator. Third, evolution has shaped our brains in a specific way. We have a strong inbuilt tendency to interpret a functionally grounded phenomenal necessity as a metaphysical necessity. But of course the fact that we simply cannot consciously imagine a certain state of affairs has no implications for the deeper structure of reality. Metaphysics is underdetermined by phenomenology.

At the outset, I pointed out that we have no empirical evidence and no truly convincing conceptual argument that supports the actual existence of 'a' self. Consequently, debates about 'the self' can easily degenerate into merely ideological disputes, turning into projection screens for metaphysical desires and hopes. There are two ways this can happen. One is the endorsement of the no-self hypothesis for purely ideological reasons, because it seems to lend support to a metaphysical world-view, which is grounded in tradition, organized religion, a specific creed or faith, and so on. The obvious example here is what I would like to term 'ideological Buddhism'. The other standard case is the rejection of the no-self hypothesis for equally ideological reasons, and here the obvious examples are all principled and purely ideological forms of anti-reductionism, as we find them in some forms of philosophical phenomenology or any substantialist metaphysics of the 'soul', for example, in Western religions. I believe that for anyone with a more serious interest in epistemic progress on self-consciousness in all its aspects, the first task will consist in effectively protecting more rigorous philosophical discussions and scientific research programs from degenerated debates of this kind.

References

Albahari, M. (2007). *Analytical Buddhism: The Two-Tiered Illusion of Self* (New York: Palgrave Macmillan).

Anscombe, G. E. M. (1975). 'The First Person', in Samuel Guttenplan (ed.), *Mind and Language* (Oxford: Clarendon Press).

Baker, L. (1998). 'The First-Person Perspective: A Test for Naturalism', *American Philosophical Quarterly*, 35: 327–46.

—— (2007). 'Naturalism and the First-Person Perspective', in G. Gasser (ed.), *How Successful is Naturalism? Publications of the Austrian Ludwig Wittgenstein Society* (Frankfurt am Main: Ontos-Verlag), 203–26.

Blanke, O., and Metzinger, T. (2009). 'Full-Body Illusions and Minimal Phenomenal Selfhood', *Trends in Cognitive Sciences*, 13/1: 7–13.

Boër, S., and Lycan, W. (1980). 'Who, Me?', *Philosophical Review*, 89: 427–66.

Brugger, P., Kollias, S. K., Müri, R. M., Crelier, G., Hepp-Reymond, M.-C., and Regard, M. (2000). 'Beyond Re-membering: Phantom Sensations of Congenitally Absent Limbs', *Proceedings of the National Academy of Science USA*, 97: 6167–72.

Castañeda, H.-N. (1966). "He': A Study in the Logic of Self-Consciousness', *Ratio*, 8: 130–57.

——(1967). 'Indicators and Quasi-Indicators', *American Philosophical Quarterly*, 4: 85–100.

Dennett, D. C. (1991). *Consciousness Explained* (Boston: Little, Brown & Co.).

Ghin, M. (2005). 'What a Self could be', *Psyche*, 11(5): www.theassc.org/files/assc/2617.pdf.

Hume, D. (1740). *A Treatise of Human Nature* (Oxford: Oxford University Press, 1967 edn).

Hunter, E. C. M., Sierra, M., and David, A. S. (2004). 'The Epidemiology of Depersonalization and Derealization: A Systematic Review', *Society for Psychiatry and Psychiatric Epidemiology*, 39: 9–18.

Johnson-Laird, P. N. (1983). *Mental Models* (Cambridge, Mass.: MIT Press).

——(2008). 'Mental Models and Deductive Reasoning', in L. Rips and J. Adler (eds), *Reasoning: Studies in Human Inference and its Foundations* (Cambridge: Cambridge University Press).

Knauff, M. (2009). 'A Neurocognitive Theory of Deductive Relational Reasoning with Mental Models and Visual Images', *Spatial Cognition and Computation*, 9: 109–37.

Ladyman, J., and Ross, D. (2009). *Every Thing Must Go: Metaphysics Naturalized* (Oxford: Oxford University Press).

Legrand, D. (2005). 'Transparently Oneself', *Psyche*, 11(5): www.theassc.org/files/assc/2616.pdf.

Lycan, W. G. (1987). *Consciousness* (Cambridge, Mass., and London: MIT Press).

——(1996). *Consciousness and Experience* (Cambridge, Mass.: MIT Press).

Malcolm, N. (1988). 'Subjectivity', *Philosophy*, 63: 147–60.

Metzinger, T. (1993). *Subjekt und Selbstmodell* (Paderborn: mentis; 2nd edn. 1999).

——(1995). 'Faster than Thought: Holism, Homogeneity and Temporal Coding', in T. Metzinger (ed.), *Conscious Experience* (Thorverton: Imprint Academic; Paderborn: mentis), 425–61.

——(2003*a*). *Being No One: The Self-Model Theory of Subjectivity* (Cambridge, Mass.: MIT Press).

——(2003*b*). 'Phenomenal Transparency and Cognitive Self-Reference', *Phenomenology and the Cognitive Sciences*, 2: 353–93.

——(2006). 'Conscious Volition and Mental Representation: Towards a More Fine-Grained Analysis', in N. Sebanz and W. Prinz (eds.), *Disorders of Volition* (Cambridge, Mass.: MIT Press).

——(2008). 'Empirical Perspectives from the Self-Model Theory of Subjectivity: A Brief Summary with Examples', in Rahul Banerjee and Bikas K. Chakrabarti (eds), *Progress in Brain Research*, 168 (Amsterdam: Elsevier), 215–46.

——(2009). *The Ego Tunnel: The Science of the Mind and the Myth of the Self* (New York: Basic Books).

——and Gallese, V. (2003). 'The Emergence of a Shared Action Ontology: Building Blocks for a Theory', in G. Knoblich, B. Elsner, G. von Aschersleben, and T. Metzinger (eds), 'Self and Action'. Special issue of *Consciousness and Cognition*, 12/4 (Dec.): 549–71.

Nagel, T. (1986). *The View from Nowhere* (New York and Oxford: Oxford University Press).

NEWEN, A. (1997). 'The Logic of Indexical Thoughts and the Metaphysics of the '"Self"', in W. Künne, A. Newen, and M. Anduschus (eds), *Direct Reference, Indexicality and Propositional Attitudes* (Stanford, Calif.: CSLI).

PARFIT, D. (1984). *Reasons and Persons* (Oxford: Oxford University Press).

PERRY, J. (1979). 'The Problem of the Essential Indexical', *Noûs*, 13: 3–22.

——(1993). *The Problem of the Essential Indexical and Other Essays* (New York: Oxford University Press).

RECANATI, F. (2007). 'Content, Mode, and Self-Reference', in S. L. Tsohatzidis (ed.), *John Searle's Philosophy of Language: Force, Meaning and Mind* (New York: Cambridge University Press), 49–64.

SIDERITS, M. (2003). *Personal Identity and Buddhist Philosophy: Empty Persons* (Aldershot: Ashgate).

——(2007). *Buddhism as Philosophy: Empty Persons* (Aldershot: Ashgate).

SIMEON, D., and ABUGEL, J. (2006). *Feeling Unreal: Depersonalization Disorder and the Loss of the Self* (New York: Oxford University Press).

SINGER, W., and GRAY, C. (1995). 'Visual Feature Integration and the Temporal Correlation Hypothesis', *Annual Review of Neuroscience*, 18: 555–86.

STRAWSON, P. (1959). *Individuals: An Essay in Descriptive Metaphysics* (London: Routledge).

THOMPSON, E. (2007). *Mind in Life* (Cambridge, Mass.: MIT Press).

WESTERHOFF, J. C. (2009). *Nagarjuna's Madhyamaka. A Philosophical Investigation* (Oxford: Oxford University Press).