Naturalism and the First-Person Perspective

Lynne Rudder Baker
University of Massachusetts Amherst

The first-person perspective is a challenge to naturalism. Naturalistic theories are relentlessly third-personal. The first-person perspective is, well, first-personal; it is the perspective from which one thinks of oneself *as* oneself* without the aid of any third-person name, description, demonstrative or other referential device. The exercise of the capacity to think of oneself in this first-personal way is the necessary condition of all our self-knowledge, indeed of all our self-consciousness. As important as the first-person perspective is, many philosophers have not appreciated the force of the data from the first-person perspective, and suppose that the first-person perspective presents no particular problems for the naturalizing philosopher. For example, Ned Block commented, "It is of course [phenomenal] consciousness rather than...self-consciousness that has seemed such a scientific mystery" (Block 1995, 230). And David Chalmers says that self-consciousness is one of those psychological states that "pose no deep metaphysical enigmas" (Chalmers 1996, 24).

In contrast to those with such casual attitudes toward self-consciousness stands Thomas Metzinger, a naturalistic philosopher who sees the complexity of self-consciousness and treats it in detail. Some years ago, I wrote an article, "The First-Person Perspective: A Test for Naturalism" (Baker 1998), in which I presented the first-person perspective as a challenge to naturalism—at least for the robust form of reductive naturalism that aims to provide accounts of all phenomena in terms accepted by the natural sciences. Metzinger has taken up this challenge, both in his article "Phenomenal Transparency and Cognitive Self-Reference" (Metzinger 2003a) and in his book *Being No One: The Self-Model Theory of Subjectivity* (Metzinger 2003b). These works offer by far the most comprehensive naturalistic theory of the first-person perspective that I know of.

I want to use Metzinger's view of the first-person perspective as a case study for naturalism. First, I'll present my own view of the first-person perspective (and point out its naturalistic and nonnaturalistic aspects), then I'll present Metzinger's reductive naturalistic account. After challenging some aspects of Metzinger's account, I want to consider some of the consequences of his account. Finally, I'll discuss naturalism more broadly and ask: Could there be a well-confirmed naturalistic theory that is rationally untenable and/or self-defeating?

1. THE FIRST-PERSON PERSPECTIVE: BAKER'S VIEW

All conscious beings—dogs, as well as human beings—have a perspective. They have points of view from which they perceive and act in the world. They solve problems by employing perspectival attitudes. Although a dog has a certain perspective on its surroundings with itself as "origin", the dog does not conceive of itself as a subject of experience. Metzinger puts it well:

> As Baker points out, it is not only necessary to have thoughts that can be expressed using 'I'. What is necessary is the possession of a concept of oneself as the *thinker* of these thoughts, as the *owner* of a subjective point of view. In short, what is needed is not only reference from the first-person point of view, but the capacity to mentally 'ascribe' this act of reference *to* oneself while it is taking place. (Metzinger 2003b, 396)

We not only *make* first-person references—e.g., 'I am registered to vote', but we also *attribute* first-person references to ourselves—e.g., 'I believe that I am registered to vote'. A first-person perspective[1] is a *conceptual* capacity to attribute first-person references to ourselves. This is a capacity to form complex first-person thoughts that have first-person references embedded in clauses following linguistic or psychological verbs. Call such thoughts and the sentences expressing them 'I*-thoughts' and 'I*-sentences', respectively. For example, 'I think (hope, fear, said) that I* am

---

[1] Throughout this paper, 'first-person perspective' should be understood as what I have lately called a 'robust first-person perspective' to distinguish it from a 'rudimentary first-person perspective' (Baker 2005).

tall' is an I*-sentence.[2] Note that I* thoughts include but are not limited to "Cartesian" thoughts (like 'I am certain that I* exist'). Mundane thoughts (like 'I hope that I* won't be late' or 'I wish that I* could buy a car') are I*-thoughts as well. I*-thoughts are first-person attributions of first-person reference, whereby one thinks of oneself as oneself*, without identifying oneself by means of any third-person referential device, such as a name, description, or demonstrative. Ability to express one's thoughts by means of I*-sentences is conclusive evidence of a first-person perspective.

From a first-person perspective, I have the ability to think of myself in a unique way, but there is no funny object that is myself-as-myself; there is no entity other than the person who I am. The referent of 'I' and of 'I*' is the person: not a body, not a disembodied ego. When I say, "I wonder whether I'll be happy in five years", I refer twice to myself—to the person, Lynne Baker, in her embodied concreteness. When I attribute first-person reference to myself by means of 'I*', what I <u>refer</u> to is no different from what you refer to by means of 'Lynne Baker'. What is special about my use of 'I*' is that I can conceive of that person in a way that you cannot, from "the inside," so to speak. This ability opens up a whole new realm of inwardness, of self-consciousness, of subjectivity.

On my view, having a first-person perspective is the defining characteristic of persons. What distinguishes us persons from other beings is our capacity to think of ourselves in a certain, first-personal way. A first-person perspective concerns *how* we think about ourselves, regardless of *what* we think about ourselves. If I am right, it is essential to your existence, to your being an entity in the world, that you have a first-person perspective. If you irretrievably lost your ability to think of yourself as yourself*, you would go out of existence—even if your brainstem still maintained the organic functions of your body. Since you the person and your body have different persistence conditions (yours depends on a first-person perspective; your body's depends on organic functioning), you are not identical to your body. On the other hand, you (the person) are not your

---

[2] Hector-Neri Castañeda introduced 'he*, and Gareth B. Matthews extended the he* from sentences with a third-person subject to 'I*' for sentences with a first-person subject. Castañeda studied phenomena expressed by sentences like 'The editor believes that he* is F.' See Castañeda 1966, and Castañeda 1967. Matthews discussed phenomena expressed by 'I think that I* am F'. See Matthews 1992.

body plus something else, just as a statue is not a piece of marble plus something else. The relation between you and your body (and the relation between the statue and the piece of marble) is what I've called 'constitution', a relation of unity that is not identity. (I worked out this view in *Persons and Bodies*, and in subsequent publications (Baker 2000, Baker 2002a, Baker 2002b).

The aim of my view of persons is to combine a fully Darwinian account of human organisms with a traditional concern of philosophers—namely, a concern with understanding our inwardness made possible by the first-person perspective. My account of the first-person perspective has some naturalistic and some non-naturalistic aspects. It is naturalistic in that it does not appeal to immaterial souls. The first-person perspective may well have evolved by means of natural selection; we human persons, with our first-person perspectives, are as much a part of the natural world as were dinosaurs.

I have no doubt that there's something going on in my brain that makes it possible for me to have I* thoughts, and I have no doubt that our capacity to have I* thoughts is a product of natural selection. The sub-personal sciences (e.g., neuroscience and parts of psychology) are sources of knowledge about mechanisms necessary for a first-person perspective in beings like us. But while I agree that the sciences may enhance our understanding of the *mechanisms* underlying the first-person perspective, I strongly disagree that knowledge of mechanisms can supplant or replace knowledge of phenomena that the mechanisms make possible.

Indeed, in many cases, knowledge of underlying mechanisms—though interesting in their own right—would not explain the phenomena that we set out to explain. For example, if we are interested in how winning the lottery changes the lives of lottery winners, a non-intentional explanation in terms of the natural sciences would be beside the point. And even where knowledge of underlying mechanisms is useful (as in, say, knowing the molecular events that trigger Alzheimer's disease), such knowledge does not unseat or replace knowledge of the disease as the destroyer of a person's life. In any event, I do not think that the natural sciences can

explain everything that there is to understand.[3] So, in this respect, I am not a naturalist.

Moreover, my Constitution view of persons may be considered to be non-naturalistic in other respects: One is that I deny that the biological origin of the first-person perspective tells us ontological significance of the first-person perspective. Ontology does not recapitulate biology. Now let us turn to Metzinger's view.

2. COGNITIVE CONSCIOUS SELF-REFERENCE: METZINGER'S VIEW

Metzinger writes sympathetically about my account of the first-person perspective. He writes that the conceptual distinction between merely having a perspective and conceiving of oneself as having a perspective—a distinction at the heart of my account of the first-person perspective—"is important for cognitive science in general, and also for the philosophical notion of a true cognitive subject" (Metzinger 2003b, 396). However, when I say, "[A]ttribution of first-person reference to one's self seems to be ineliminable" (Baker 1998, 331), Metzinger disagrees. He offers an alternative view that eliminates reference to any self or genuine subject of experience. On his view, "all that exists are conscious systems operating under transparent self-models" (Metzinger 2003b, 397). On my view, I (me, the person, a first-personal being, a genuine subject of experience, a "self") am an entity in the world. So, the issue between Metzinger and me is joined in a profound and intriguing way: When I affirm that there are persons with irreducible first-person perspectives in the world, I am affirming that there are genuine subjects of experience (essentially first-personal beings) in the world. When Metzinger denies that there are

---

[3] Some naturalists (e.g., Quine 1960) confine science to the so-called natural sciences; intentional descriptions are simply a dramatic idiom. I'll call this version Reductive Natualism. Other naturalists (e.g., Kornblith 1993) who are antireductionists may countenance irreducible social and psychological sciences that advert to intentional phenomena. Metzinger clearly aims for an account in terms of sub-personal mechanisms and is a reductive naturalist.

"selves," he is denying that there are genuine subjects of experience in the world.[4]

Let me make two terminological points:

(1) I follow Metzinger's use of the word 'phenomenal' to apply to the qualitative contents of conscious experience; phenomenal experience is characterized by how it feels or "what it's like" to have it. This leaves it open whether or not a phenomenal content represents anything real, or is, as Metzinger puts it, "epistemically justified" (Metzinger 2003b, 401). Phenomenal content may or may not depict anything in reality.

(2) Metzinger denies that there are any entities in the world that are "selves" or genuine subjects of experience. By the term 'genuine subject of experience,' I mean an entity that must be included as such in ontology—a first-personal entity that exists in the world and not just as an artifact of an information-processing system. Although I do not believe that there exist "selves" as distinct from persons, I do believe that there are persons, who are essentially first-personal, and are genuine subjects of experience (call them 'selves' if you'd like). I prefer the word 'persons' or 'genuine subjects of experience' to the word 'self', but I'll use all of these locutions to mean the same thing.

Although Metzinger emphasizes the importance of the first-person perspective in the very terms in which I describe it, he argues that we can account for the first-person perspective without supposing that there are "selves" or genuine subjects of experience. The question, then, comes down to this: Can there be an adequate ontology—an inventory of what really exists—that includes no first-personal subjects of experience, but only information-processing systems and self-models that are understandable in wholly third-personal terms?

The portion of Metzinger's argument that concerns me here has three parts:

---

[4] If all Metzinger means by a self or a subject of experience is "an internal and nonphysical object," Metzinger, 2003b, 271, then almost everyone agrees with him that there are none; and there would be no argument. I do not suppose him to be taking on a "straw man."

(i)   a sub-personal, naturalistic account of subjective experience,

(ii) an account of how it can *seem* to us that we are genuine subjects of experience, and

(iii) an account of the (putative) fact that there really are no genuine subjects of experience in the world. Metzinger offers a theory both that denies that I am a genuine subject of experience and that shows what is really going on when it *seems* to me that I am a genuine subject of experience.

The first part of Metzinger's argument is to give an account of subjective experience. Our brains activate mental models that contain mental representations. Mental representations have both phenomenal content (smells, colors, etc.) that supervenes on brain states, and intentional content (wishing you were here, believing that global warming is a serious threat) that depend in part on relations to an environment. Our representations are part of mental models, some of which represent the world (world-models) and some of which represent the system generating the models (self-models).

> [A] self-model is a model *of* the very representational system that is currently activating it within itself. (Metzinger 2003b, 302) The content of a phenomenal self-model (PSM) is the conscious self: your bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processes. (Metzinger 2003b, 299)

Some properties of a self-model are transparent—that is, we don't see them, we look through them; they are not introspectively accessible. Transparency here is a phenomenological, not an epistemological, notion. Other properties are opaque—that is, we are aware of them; they are introspectively accessible. E.g., as G.E. Moore pointed out, when we try to introspect the sensation of blue, the *sensation* (what the sensation of blue has in common with the sensation of green) is transparent: "we look through it and see nothing but the blue" (Moore 1903, 446). But the blue is opaque; it is what we see. Metzinger says:

> A transparent representation is characterized by the fact that the only properties accessible to introspective attention are their content properties. (Metzinger 2003b, 387)

Our subjective experience, in the first instance, is activation of representations in transparent models—i.e., only the representational contents are experienced, not the models themselves. In other words, subjective experience is phenomenal experience. It consists of activation of models of representations. We cannot experience the models. We experience only the content properties of representations, whether the contents depict anything outside the model or not.

The second part of Metzinger's argument is to show how it can seem to us that we are subjects of experience. Metzinger distinguishes between a phenomenal first-person perspective and a cognitive first person perspective (Metzinger 2003b, 405). A phenomenal first-person perspective allows an information-processing system to have phenomenal (i.e., subjective) experience; a cognitive first-person perspective allows an information-processing system to have I* thoughts that make it seem that it is a genuine subject of experience in the world.[5]

I*-thoughts require integrating part of an opaque self-model into a preexisting transparent self-model[6] (Metzinger 2003b, 402). The opaque self-model is a phenomenal model of the intentionality relation (PMIR) that "represents itself in an ongoing, episodic *subject-object relation*." (Metzinger 2003b, 411) What we think about when we consciously think about ourselves is really just the content of a self-model. In having I* thoughts, we are unable to consciously experience that "we are referring to

---

[5] To show how it can seem to us that we are subjects of experience, Metzinger begins with a transparent phenomenal self-model that can be generated by an animal or pre-linguistic being; then, a conscious cognitive subject emerges when the system generates opaque representations of itself and integrates them into the transparent phenomenal self-model. In Metzinger's words: "My claim is that, all other constraints for perspectival phenomenality satisfied, a conscious cognitive subject is generated as soon as a globally available representation of the system as currently generating and operating with the help of quasi-linguistic, opaque mental representations is integrated into the already existing transparent self-model" (Metzinger 2003a, 367f.; Metzinger 2003b, 395).

[6] Metzinger defines a minimal notion of self-consciousness as having three properties: "the content of the self-model has to be embedded into a currently active world-model; it has to be activated within a virtual window of presence; and it has to be transparent" (Metzinger 2003a, 373).

the content of a *representation* that is 'in ourselves' (in terms of locally supervening on brain properties)" (Metzinger 2003b, 402). Metzinger continues:

> Cognitive self-reference always is reference to the phenomenal content of a transparent self-model. More precisely, it is a *second-order* variant of phenomenal self-modeling, which, however, is mediated by *one and the same* integrated vehicle of representation. The capacity to conceive of oneself as oneself* consists in being able to activate a dynamic, 'hybrid' self-model: Phenomenally opaque, quasi-symbolic, and second-order representations of a preexisting phenomenally transparent self-model are being activated and continuously reembedded in it. This process is the process of [conscious cognitive self-reference]....Reflexive self-consciousness consists in establishing a subject-object relation within the [phenomenal self-model]. [7] (Metzinger 2003b, 403)

Let me try to put this in my own words. If someone thinks, "I am hungry," she is activating a transparent phenomenal self-model. She sees through the 'I' (so to speak) to the feeling of hunger. The 'I' is invisible to her. But if she thinks, "I believe that I* am hungry," the first occurrence of 'I' is part of an opaque self-model that is integrated into the preexisting transparent self-model. The second occurrence of 'I' in 'I believe that I* am hungry' (the 'I*') is phenomenologically transparent. The first occurrence of 'I' is opaque since she is thinking of herself as the subject of her thought. What remains invisible to her is precisely what she is referring to. A conscious information-processing system seems to be a subject of experience when it generates subjective experiences that include the experience of being a subject of experience. Thus, we seem to be subjects of experience in the world. But the experience of being a subject of experience remains phenomenal.

The third part of Metzinger's argument is to show that the experience of being a substantial subject is *merely* phenomenal. The conscious cognitive subject is not part of reality, but only part of a self-model. Metzinger holds that a cognitive first-person perspective (that is, the ability to have I*

---

[7] Metzinger 2003b, 403. I inserted 'consciously experienced cognitive self-reference' for 'introspection$_4$'. Metzinger characterizes introspection$_4$ as "a conceptual (or quasi-conceptual) kind of metarepresentation, operating on a pre-existing , coherent self-model" (Metzinger 2000a, 367).

thoughts) is a special case of a phenomenal first-person perspective: "Cognitive self-reference is a process of phenomenally modeling certain aspects of the content of a preexisting transparent self-model, which in turn can be interpreted as the capacity of conceiving of oneself as oneself*" (Metzinger 2003b, 405). In cognitive self-reference, what is referred to is the phenomenal content of a transparent self-model. So, the reference will be to an element of the self-model, not to a self existing in the world. In short, the conscious cognitive subject is just an element of the self-model. Metzinger says:

> Any conscious system operating under a phenomenally transparent self-model will by necessity instantiate the phenomenal property of selfhood in a way that is untranscendable for this system itself. (Metzinger 2003a, 363)

I believe that the word 'untranscendable' in this passage means that the system lacks resources to uncover the fact that the phenomenal property of selfhood is *merely* the content of a self-model. But according to Metzinger, what we refer to in cognitive self-reference is a mental representation: "[I*]," he says, "is the content of the transparent self-model. (Metzinger 2003b, 400).

Metzinger's claim that the *cognitive* first-person perspective can be reduced to a complex *phenomenal* first-person perspective has a strong consequence about subjects of experience: No belief about the worldly existence of *what* is being mentally represented is "epistemically justified." That is, we cannot conclude that what is represented exists in reality. Metzinger says that the belief that a self carries out the act of cognitive self-reference is not epistemically justified, and hence is apt for rejection (Metzinger 2003b, 403). Thus, we can see how the Cartesian claim of epistemic transparency (my certainty that I am a genuine subject of experience that exists in reality,) is intelligible, even if it is false (Metzinger 2003a, 363).

In sum, Metzinger denies that conscious experience really has a subject in the world (a self or person who does the experiencing). Our experience of being subjects of experience is only phenomenal. We are mistaken if we think that, because we experience being a subject of experience, there actually *is* (in reality) a subject of experience who we are. We lack "epistemic justification" for "all corresponding belief states about what is

actually being represented" (Metzinger 2003b, 404; Metzinger 2003a, 375). The subjective experience of being someone in the world is an illusion. Just as dreams and hallucinations tell us nothing veridical about what's really going on in the environment, so too does subjective experience tell us nothing veridical about what we are. There are no selves, just self-models. "For ontological purposes," he says, "'self' can therefore be substituted by 'PSM' [phenomenal self-model]" (Metzinger 2003b, 626).

Metzinger says that the main thesis of his book *Being No One*

> is that no such things as selves exist in the world: Nobody ever was or had a self. All that ever existed were conscious self-models that could not be recognized as models. (Metzinger 2003b, 1)

The experience of oneself is only a phenomenological consequence of a system operating under a phenomenal self-model (Metzinger 2003b, 387). This is compatible with saying either that I (a subject of experience) do not exist, or that I exist but that what I am is only a part of the content of a self-model.

However, I believe that the most charitable way to read Metzinger is not as an eliminativist about subjects of experience, but as a reductionist. Despite the misleading title of his book, *Being No One*, and despite what I just quoted him as saying, perhaps he is not saying that I do not exist, or that I am no one. Perhaps he is saying that *what I am* is an information-processing system that has generated a phenomenal self-model (PSM), and that *what I think about* when I think about myself is only the content of a mental representation in my self-model.

In any case, whether Metzinger is an eliminativist about selves (as his quotations suggest) or a reductionist (as I think is the more charitable interpretation), he denies that there exist what I have called 'genuine subjects of experience'—first-personal entities that must be included as such in ontology. If Metzinger is correct, then the fact that you and I seem to be subjects of experience has no ontological significance. Persons (selves, subjects of I* thoughts) belong to appearance, not to reality.

3. TWO ISSUES INTERNAL TO METZINGER'S THEORY

Let me express my admiration for the cleverness of Metzinger's theory. Indeed, there are a number of points of broad agreement between Metzinger and me. Here are some examples:

(1) Self-consciousness is importantly different from mere sentience, or the kind of consciousness that nonhuman animals have (Metzinger 2003b, 396).

(2) Self-conscious beings possess the distinction between the first and third person "on a *conceptual* level, and actually use it" (Metzinger 2003b, 396).

(3) Philosophers cannot "decide on the truth or falsity of empirical statements by logical argument alone" (Metzinger 2003b, 3).

(4) The phenomenology of conscious experience should be taken seriously (Metzinger 2003b, 301 n2).

(5) A human being can "conceive of itself *as a whole*" (Metzinger 2003b, 1).

Despite these areas of agreement, I would like to critically discuss two issues internal to Metzinger's view, and then turn to the main difference between my view and Metzinger's: The main difference between us is the ontological difference, stemming from his commitment to reductive naturalism. Whereas I think that a complete ontology must include persons ("selves" or genuine subjects of experience), Metzinger does not. That is, although I think that there are selves in reality (again, I really prefer the word 'person'), Metzinger thinks that selves are only matters of appearance, not reality. On his view, as we have seen, reality includes no selves, only self-models.

    The two issues internal to Metzinger's view that I want to discuss are, first, Metzinger's "analysis" of cognitive first-person reference from a third-person point of view, and second, his notion of phenomenal content and the use that he makes of it.

    First, consider Metzinger's argument against my claim that attribution of first-person reference to oneself is ineliminable. In the article of mine

that Metzinger discusses, I used the example of Descartes' I*-thought, [I am certain that I* exist][8], and I pointed out that the certainty that Descartes claimed was first-personal: Descartes claimed that he was certain that he* (he himself) existed, not that he was certain that Descartes existed. Although Metzinger agrees that Descartes was not making a third-person reference to Descartes (Metzinger 2003b, 398), he also holds that the mental content of Descartes' thought [I am certain that I* exist] and the linguistic content of the sentence 'I am certain that I* exist' can be understood in third-person terms.

All the mental content of the thought [I am certain that I* exist] is merely phenomenal and, as Metzinger says, "not epistemically justified" (Metzinger 2003a, 373).[9] In short, my certainty that I* exist is understood as a complex relation of parts of the content of a self-model. In general, I*-thoughts are to be understood without supposing that a subject of experience exists in reality.

Metzinger also treats *linguistic* self-reference by the sentence <I am certain that I* exist>. The linguistic content of <I am certain that I* exist> may be "analyzed", he says, from a third person perspective as follows:

(A) <The speaker of this sentence currently activates a PSM (a phenomenal self-model) in which second-order, opaque self-representations have been embedded. These representations are characterized by three properties:

First, they possess a quasi-conceptual format (e.g., through a connectionist emulation of constituent-structure, etc.);

second, their content is exclusively formed by operations on the transparent partitions of the currently active PSM;

third, the resulting relation between the system as a whole and content is phenomenally modeled as a relation of certainty.> (Metzinger 2003b, 402)

---

[8] Metzinger uses square brackets ([....]) to denote thoughts, and pointed brackets (<...>) to denote linguistic expressions.

[9] Such phenomenal certainty has two defining characteristics. The first is that "the object-component of the phenomenal first-person perspective is transparent and the respective person is therefore, on the level of phenomenal experience, forced into an (epistemically unjustified) existence assumption with respect to the object-component." The second defining characteristic is "transparency of the self-model yielding a phenomenal self depicted as being certain" (Metzinger 2003a, 374).

Let us label this account (A). Can (A) be a correct analysis of a first-person assertion <I am certain that I* exist>? My assertion <I am certain that I* exist> is necessarily about me, Lynne Baker. But the analysis is not. The analysis is about anybody who asserts that she* is certain that she* exists. Neither my assertion <I am certain that I* exist> nor (A) entails the other. So, the proposed analysis (A) is not an analysis in a traditional sense. Nor can (A) replace anyone's assertion of 'I am certain that I* exist.' The target sentence and (A) simply do not convey the same information.[10]

What is at issue is not the specific Cartesian example <I am certain that I* exist>, however, but rather my broader claim that the attribution of first-person reference to one's self seems to be ineliminable," (Baker 1998, 331). It is this broader claim—one that applies to all I*-thoughts and I*-sentences that is at stake.

So perhaps (A)—even if it is not an analysis—should be regarded as an application of part of an empirical theory. Metzinger predicts that the phenomenal self-model (PSM) is a real entity that will be empirically discovered—"for instance, as a specific stage of the global neural dynamics in the human brain, characterized by a discrete and unitary functional role" (Metzinger 2003b, 411). The only thing to say here is that we will have to wait and see whether neural correlates of phenomenal self-models are actually discovered in the brain.

Even if they are discovered, however, the most that a third-person empirical theory of I*-sentences or I* thoughts can hope to do is to provide necessary and sufficient conditions for the production of I* sentences or I* thoughts. But this would be a far cry from eliminating or replacing I* sentences or I* thoughts by third-person sentences or thoughts. Even if (A) is part of an empirical theory that is eventually confirmed, it still cannot *replace* the I*-sentence, which remains ineliminable.

The second question that I want to raise that is internal to Metzinger's theory is whether the notion of phenomenal content can bear the load that Metzinger puts upon it. Phenomenal content is qualitative content and

---

[10] Maybe (A) is what makes an assertion of <I am certain that I* exist> true. Maybe (A) is the truth-masker for such assertions. But the notion of truth-makers is part of a controversial metaphysical theory outside the purview of any empirical science known to me. So, as a naturalist, Metzinger should be reluctant to appeal to truth-makers. (And, as far as I know, he does not appeal to truth-makers.)

(supposedly) supervenes on the brain; representational content is intentional content (Metzinger 2003b, 71). Metzinger says:

> The central characteristic feature in individuating mental states is their phenomenal content: the way in which they *feel* from a first-person perspective. (Metzinger 2003b, 71)

In my opinion, this is not the way that mental states should (or even could) be individuated—at least those mental states that have truth-conditions, as all I*-thoughts do. We have no criterion for sameness of feeling: I wake up at night and on some occasion my subjective experience is hope that I'll get a certain paper finished on time; on another occasion, my subjective experience is hope that it won't rain tomorrow. My subjective experience is certainly not the same on both occasions of hope, but not because of any difference in feeling. The difference—even the difference in what it's like to be in the states—depends on the *intentional content* of the hopes, not on any feeling associated with them. So, I do not think that purely phenomenal content can individuate mental states.

According to Metzinger, "conceptual forms of self-knowledge" (I* thoughts) are generated

> by directing cognitive processes towards certain aspects of internal system states, the intentional content of which is being constituted by a part of the world depicted as *internal*. (Metzinger 2003a, 367; his emphasis.)

He says that the phenomenology associated with this type of representational activity

> includes all situations in which we consciously think about ourselves *as* ourselves (i.e., when we think what some philosophers call I* thoughts; for an example see Baker 1998). (Metzinger 2003a, 367)

It seems to me to be phenomenologically mistaken to suppose that the intentional contents of I* thoughts depict part of the world as internal. When I think: "I believe that I* can get money from this ATM", the intentional content of my I*-thought is not constituted by a part of the world depicted as *internal*. Still less is internality "phenomenally experienced." When I consciously think, "I believe that I* can get money

from this ATM," the intentional content of my thought depicts a relation between a machine and myself—a relation that is not internal to me.

Metzinger endorses a principle of local supervenience for phenomenal content: "phenomenal content supervenes on spatially and temporally internal system properties" (Metzinger 2003b, 112). He goes on:

> If all properties of my central nervous system are fixed, the contents of my subjective experience are fixed as well. What in many cases, of course, is not fixed is the *intentional* content of those subjective states. (Metzinger 2003b, 112)

But almost all subjective experience (mine, anyway) has intentional content. Any mental state that can be true or false, or that can be fulfilled or unfufilled, has intentional content, no matter what it feels like.[11]

For example, it suddenly occurs to me that I locked my keys in my office, and I experience a feeling of panic. The subjective experience has intentional, not just phenomenal, content; it includes a thought that has a truth value. And I'm greatly relieved if I discover that the truth value of my thought is false: Here the keys are in my pocket. The subjective experiences were not just the panic and the relief; they included the sudden thought with its specific intentional content and the happy discovery that the thought was false. Not only are we embodied, but also we are embedded—embedded in a real world, not just in representations of a world. And the contents of our subjective experience are typically infected by relations with the environment.[12]

Since, according to Metzinger, phenomenal content supervenes on brains, and most of our subjective experience has intentional content, which does not supervene on the brain, phenomenal content cannot account for our subjective experience. Our brains, and what supervene on them, are only one determinant of subjective experience. I may wake up in

---

[11] Although I cannot argue for it here, I believe that none (or almost none?) of our intentional mental states supervene on our brain states. See, Baker, forthcoming.

[12] Metzinger notes that one "of the most important theoretical problems today consists in putting the concepts of 'phenomenal content' and 'intentional content' into the right kind of logical relation" (Metzinger 2003b, 112). That seems to me a problem easily solved: Do not insist that phenomenal content (content that is experienced) supervene on brain states. With the exception of qualia (if there are any), all content depends on interaction with the environment.

the night, thinking that a search committee meeting the next day may be unpleasant. That particular subjective experience would be metaphysically impossible (and not just causally impossible) in a world without search committees and all the intentional apparatus surrounding hiring new people. So my subjective experience of thinking that tomorrow's meeting may be unpleasant does not supervene on my brain. Hence, phenomenal content, which does supervene on my brain, does not suffice for ordinary subjective experience. Metzinger asserts:

> Phenomenal content can be dissociated from intentional content: a brain in a vat could possess states subjectively representing object colors as immediately and directly given. (Metzinger 2003a, 359)

This claim brings to the fore the dilemma that phenomenal content faces: If phenomenal content is dissociated from intentional content, it does not account for much of our subjective experience, as the above examples show. But if phenomenal content is not dissociated from intentional content, then phenomenal content does not supervene locally on brain states and it loses the neuroscientific legitimacy that Metzinger claims for it. Either way, phenomenal content cannot play the role that Metzinger assigns it.

To recapitulate, my two objections internal to Metzinger's view concern his attempt to eliminate I*-thoughts and I*-sentences (or to reduce them to the third-person), and his use of phenomenal content to carry the weight of subjective experience. Now let us turn to some consequences of Metzinger's theory.

4. CONSEQUENCES OF METZINGER'S THEORY

In this section, I want to consider three kinds of consequences of Metzinger's view that I find untenable—semantic, epistemic, and moral consequences.

First, I believe that Metzinger's view requires an ineliminable equivocation on the word 'I'. Sometimes 'I' refers to the whole information-processing system, and sometimes 'I' refers to the content of a part of a self-model. This becomes apparent if we consider I* sentences.

Consider an ordinary I* thought—e.g., 'I believe that I am in Austria'. Metzinger says:

> I experience myself as the thinker of the I*-thoughts. (Metzinger 2003a, 373)

The reality that the first occurrence of 'I' in this thought refers to is the whole information-processing system.

> The content of [I] is the thinker, currently representing herself as operating with mental representations. (Metzinger 2003b, 401)

It is the whole system that thinks of itself as the thinker of thoughts. On the other hand, the second occurrence of 'I' (the 'I*' in 'I believe that I am in Austria') "is the content of the transparent self-model." As Metzinger explains:

> Any conscious system operating under a transparent self-model will by necessity instantiate a phenomenal self to which, linguistically, it *must* refer using <I*>. (Metzinger 2003b, 400, emphasis his.)

So, the referent of 'I' is sometimes the whole information-processing system and sometimes the content of a self-model. It is utterly implausible that 'I' could be equivocal in a single thought of a single thinker. This would make us all hopelessly schizophrenic: Which am I—the whole information-processing system or part of the transparent content of its currently active self-model?

We can see this tension in another way when we consider Metzinger's metaphor that

> you constantly confuse yourself with the content of the self-model currently activated by your brain. (Metzinger 2003b, 1)

Who is doing the confusing? On the last page of his book, Metzinger says that we should not take this metaphor too literally:

> There is no one *whose* illusion the conscious self could be, no one *who* is confusing herself with anything. (Metzinger 2003b, 634).

What, exactly, then is the confusion that has no bearer? It is difficult to see how there is a confusion to be made (with or without someone to make it).

When I think, "I believe that I* in Austria", my belief is that I (all of me) am in Austria. Perhaps, Metzinger is saying that, unbeknownst to me, the information-processing system that I am has a transparent self-model representing being in Austria, and the system integrates part of an opaque self-model representing itself into the transparent self-model, and thus generates a representation of a representation of being in Austria within the self-model. This would completely misrepresent the content of my thought "I believe that I* am in Austria." If you and I agree that I believe that I* am in Austria, then we are agreeing about me, about where I believe I am (even if I am an information-processing system operating with a self-model); we are not agreeing about my self-model. So, I think that it is not coherent to construe the subjects of I* thoughts to be parts of self-models.

Second, consider an epistemic consequence of Metzinger's view. The theory cannot make sense of what is going on when people reflect on what they are doing while they are doing it. Suppose that a scientist using an electron microscope for the first time thinks to herself, "I can hardly believe that I'm looking at electrons." If the scientist is not a subject of experience that exists in the world, how is she to make sense of her own thought, on Metzinger's view? Well, maybe this: The scientist has the experience of being the subject of the thought expressed by "I can hardly believe that I'm looking at electrons," but she is not "epistemically justified" in supposing that she really is a genuine subject of experience in the world.

From Metzinger's point of view, the scientist is an information-processing system that is integrating

> its own operations with opaque mental representations, i.e., with mental simulations of propositional structures that could be true or false, into its already existing transparent self-model while simultaneously attributing the causal role of generating these representational states to itself. (Metzinger 2003a, 369)

But, on Metzinger's view, *the scientist herself* cannot see her own thoughts and activity in this light; indeed, she is deceived about what is going on. Of course, Metzinger has an account of why she cannot see her own thoughts and activity in this light; but that's beside the point. The point is that the scientist cannot comprehend what is really going on while she is engaging in scientific activity. Metzinger's theory would seem to make it impossible

for anyone to think clearly about what she is doing while she is doing it. A view of subjectivity that makes it impossible for scientists (and everyone else) to think clearly about what they are doing as they are doing it is dubious.

Third, Metzinger's view has consequences that are morally questionable. Consider a soldier long ago who experienced excruciating pain while undergoing a battlefield amputation. Metzinger says that we should minimize "the overall amount of suffering in all beings capable of conscious suffering" (Metzinger 2003b, 570). I do not see what epistemic grounds we can have for this "simple principle of solidarity," as he calls it. If Metzinger's view is correct, then we are epistemically unjustified in supposing that there is any substantial entity in the world that actually undergoes excruciating pain; rather, there is an information-processing system with a self-model that made it appear that there was such a subject of pain. There was a subjective experience of pain, but the bearer of the pain was just a phenomenal self, who was "epistemically unjustified." If we are unjustified in supposing that there was a substantial entity (the soldier) who was a subject of pain, then we would be under no obligation to alleviate the pain. I think that this consequence would make our moral experience unintelligible.[13]

I am prepared to accept theories with counterintuitive consequences (e.g., I find it counterintuitive that there's no absolute ongoing now; but I accept this as a result of well-confirmed theories of physics). But Metzinger's view of the first-person perspective and its I* thoughts is not just counterintuitive. It has consequences that seem to me to be semantically, epistemically and morally untenable. So, what should we do?

5. WHITHER NATURALISM?

Metzinger's theory is a naturalistic one. Naturalism is often characterized by two themes—an ontological one that is committed to an exclusively

---

[13] In an email to me, Metzinger said that he was very interested in ethical consequences of his view. He said that he believes that there can be selfless suffering subjects, and that phenomenal suffering is real and should be minimized. I hope that he pursues these issues at length. It is not obvious to me how to work out a morally acceptable position within the confines of his view.

scientific conception of nature, and a methodological one that conceives of philosophical inquiry as continuous with science (De Caro & Macarthur 2004, 3). Reductive naturalism recognizes as real only third-personal entities and properties.[14]

Metzinger's third-person sub-personal account of the first-person perspective fits this characterization of reductive naturalism nicely. So, I shall continue to use Metzinger as a case study. On being presented with a theory, each of us decides: Do I accept this theory? I invite you to join me in thinking of Metzinger's theory from the point of view of a prospective adherent of it. Would it be *rational* for me to accept it? Would it even be *possible* for me to accept it? Let's consider each of these questions in turn.

(i) If Metzinger's view is correct, then there are no selves and no genuine subjects of experience in the world. I just argued that without subjects of experience in reality, I cannot make sense of my own experience while I'm having it. A view with this consequence renders my experience unintelligible to me. Is it rational for me to endorse a theory that renders my experience unintelligible to me? My experience of being a conscious subject is evidence that I am a subject, and this evidence overwhelms any possible evidence that I may have for any scientific theory to the contrary. Hence, rationally, I should reject the view that would have me repudiate myself as a genuine subject of experience.

(ii) It seems that Metzinger's theory cannot coherently be endorsed or accepted. I may have the subjective experience that I* am accepting Metzinger's theory. I think to myself, "I am having the experience that I* am accepting Metzinger's theory." But the "I*" doing the accepting is not an entity in the world; it is just part of the content of a transparent self-model (Metzinger 2003a, 372; Metzinger 2003b, 400). When I refer to myself by means of 'I*', I am referring to the content of a mental representation. It is incoherent to suppose that a mental representation can actually accept a theory. On Metzinger's view, all there can be is a subjectless subjective experience of accepting his theory; but for me to accept a theory is not just for there to be a subjectless subjective experience of accepting. So, it seems doubtful that Metzinger's theory can be endorsed

---

[14] Whether non-reductive naturalism can allow irreducibly first-person phenomena remains to be seen.

or accepted. If a theory cannot coherently be endorsed or accepted, it is self-defeating. It is paradoxical, if not self-contradictory, to suppose that I should accept a theory that I cannot coherently accept.

Here is my recommendation: Give up reductive naturalism. Do not confine ontological conclusions to those that can be gleaned by scientific methods. As we have seen in the best attempt to naturalize the first-person perspective, science (at least as it stands today) cannot intelligibly be the final word on what there is. Even if philosophers gave up naturalism as a *global* commitment to the methods and ontology of natural sciences, however, we may still keep those naturalistic theories that explain what we want explained. The way to accomplish this is to attend to what the naturalistic philosophical theories are (or should be) theories *of*.

We should distinguish between phenomena that interest philosophers and the underlying mechanisms that subserve those phenomena. For example, we may hope for a naturalistic theory of the mechanisms that underwrite a first-person perspective (Metzinger 2003b, 395). But on my view, the "I" who is the genuine subject of experience is a person: an object in the world whose first-person perspective is irreducible and ineliminable.

Why is my view to be preferred to Metzinger's? First, his theory (with a phenomenal self that is not a genuine object in reality) is paradoxical; mine is not. Second, his theory relies on an inadequate view of subjective experiences as supervening on the brain; mine does not. Third, his theory would leave the work undone that the first-person perspective does—e.g., in understanding moral agency; mine does not (Baker 2000). Fourth, his interpretation has unfortunate semantic, epistemic, and moral consequences; mine does not.

Reductive naturalism often seems like a change of subject that lacks respect for the peculiar projects and puzzles that traditionally preoccupy philosophers. In particular, nonnaturalists resist the tendency to assimilate the phenomena that piqued our philosophical interest to the mechanisms that support those phenomena. No one doubts that there are underlying mechanisms and that they are worthy of understanding. The nonnaturalist resistance is to *supplanting* philosophical questions by empirical questions about the underlying mechanisms that make the philosophically-interesting phenomena possible—as if questions about the 1985 world-championship

chess match between Kasparov and Karpov could be replaced by questions about the physics involved in the motions of little bits of wood.

Taking Metzinger's view as the best case, I now suspect that the challenge that the first-person perspective poses for reductive naturalism cannot be met.[*]

6. REFERENCES

Baker, L. R. 1998: "The First-Person Perspective: A Test for Naturalism." *American Philosophical Quarterly* 35, 327-348.

Baker, L. R. 2000: *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Baker, L. R. 2002a: "On Making Things Up: Constitution and its Critics." *Philosophical Topics* 30, 31-51.

Baker, L. R 2002b: "Precis" and "Reply to Critics." *Philosophy and Phenomenological Research* 64, 592-598 and 623-635.

Baker, L. R. forthcoming: "First-Person Externalism". *The Modern Schoolman.*

Block, N. 1995: "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences* 18, 227-247.

Castañeda, H.-N., 1966: "He: A Study in the Logic of Self-Consciousness." *Ratio* 8, 130-157.

Castañada, H.-N.. 1967: "Indicators and Quasi-Indicators." *American Philosophical Quarterly* 4, 85-100.

Chalmers, D. J. 1996: *The Conscious Mind: Toward a Fundamental Theory*. Oxford: Oxford University Press.

De Caro, M. and Macarthur, D. 2004: *Naturalism in Question*. Cambridge, MA: Harvard University Press.

Kornblith, H. 1993: *Inductive Inference and its Natural Ground*: *An Essay in Naturalistic Epistemology*. Cambridge, MA: MIT Press.

Matthews, G. B. 1992: *Thought's Ego in Augustine and Descartes*. Ithaca. NY: Cornell University Press.

Metzinger, T. 2003a: "Phenomenal Transparency and Cognitive Self-Reference." *Phenomenology and the Cognitive Sciences* 2, 353-393.

Metzinger, T. 2003b: *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Moore, G. E. 1903: "The Refutation of Idealism." *Mind* 12, 433-53.

Quine, W. V. O. 1960: *Word and Object*. Cambridge, MA: MIT Press.