Journal for Politics, Gender and Culture Vol. 8 / No. 2 / Summer 2011

Ray
Brassier

The View from Nowhere

> "True" is a sign that something is to be done,
> *for inferring is a doing.*
>
> (Sellars 1991b, 206)

Philosophy, said Wilfrid Sellars, is the attempt "to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term." (Sellars 1991a, 1) Despite its apparent vagueness, this is as good a way of encapsulating the concerns of philosophy as anyone has ever given, since we can specify what the "broadest possible sense" of the terms "things" and "hang together" is here. For Sellars, "things in the broadest possible sense" covers everything from theorems to fermions. By the same token, the philosophical sense of "hanging together" should furnish an insight into the link between things as disparate as logical norms and elementary particles. The philosophical vision ought not only to encompass but also to *explain* the intrication of conceptual ideality and physical reality. Is this to reiterate an antiquated dualism? No. A dualism is a distinction that fails to explain the connection between the terms it distinguishes. Philosophy discriminates, it distinguishes and separates, but always with a view to ultimate integration. In this regard, philosophy discriminates precisely in order to avoid dualism. The animus towards dualism should not excuse insensitivity towards distinction. To distinguish between the normative and the factual is not to promulgate dualism once it is understood that this distinction furnishes the precondition for understanding the intrication of the conceptual and the physical; an intrication that is constitutive of what we call "reality." Philosophy is synoptic in that it strives to reconcile a basic disjunction in our conception of reality. This disjunction is a consequence of the fundamental conceptual discrepancy bequeathed to us by philosophical modernity. If Sellars' work (unlike that of many of his analytic contemporaries) retain its contemporaneity for us today, fifty years after the bulk of it was written, this is because, over and above its sometimes forbidding difficulty, it represents one of the most sustained attempts to think through the implications of a fundamental diremption which extends into our very conception of what we are. This is the diremption between our self-understanding as rational subjects and our scientific understanding of ourselves as physical objects. Throughout his work, Sellars sought to arbitrate the conflict between these two increasingly divergent *images* of man-in-the-the-world:

the *manifest* image of man as a self-conscious rational agent and the *scientific* image of man as a "complex physical system." Yet Sellars was careful not to portray this divergence as a conflict between naïve pre-theoretical common-sense and sophisticated theoretical reason. Rather, he insisted it be understood in terms of the tension between the disciplined and critical *refinement* of common-sense through which a perennial tradition of philosophical reflection has taught us to conceive of ourselves as rational beings bound by conceptual norms; and the methodical *extrapolation* from ordinary perception through which modern science has taught us to explain manifest phenomena by postulating increasingly complex systems of imperceptible entities (e.g., molecules, electro-magnetic radiation, gravitational fields, etc.). In this regard, the fundamental contrast at issue is one between man's manifest self-image as a rule-bound rational agent participating in but not governed by the realm of physical law, and man conceived through the optic of natural science as a "complex physical system" whose capacity for agency can ultimately be accounted for in terms of concatenations of spatio-temporal causation.

Yet there is a persistent ambiguity in Sellars' account of the relation between manifest and scientific images. On one hand, he seems to insist that the philosophical task is to recognize the *parity* of the two images. The acknowledgement of parity follows from the realization that the images are not in fact competing over the same territory. Philosophy can adjudicate between the competing claims of the manifest and scientific images by distinguishing the *normative* privileges of the former from the *ontological* rights of the latter. Thus, apparently undermining his commitment to parity, Sellars upholds the *priority* of the scientific image by famously insisting that "in the dimension of describing and explaining the world, science is the measure of all things, of what is, that it is, and of what is not, that it is not." (Sellars 1991, 173) This apparent inconsistency can be defused once we recognize that the commitment to parity and the commitment to priority operate at two distinct levels: that of conceptual interpretation (giving and asking for reasons) and that of ontological description and explanation. Parity at the level of conceptual interpretation is compatible with priority at the level of ontological description and explanation. The claim for parity follows from the recognition that the manifest image furnishes us with the fundamental framework in terms of which we understand ourselves as "concept mongers,"[1] creatures continually engaged in giving and asking for reasons. But we are able to do things with concepts precisely insofar as concepts are able to do things to us. It is this capacity to be gripped by concepts that makes us answerable to conceptual norms. And it is this susceptibility to norms that makes us subjects. The manifest image is indispensable insofar as it provides the structure within which we exercise our capacity for rational thought. Hence the parity between images: both are governed by the norm of truth, understood as maximally warranted assertion, despite the conceptual incommensurability between manifest and scientific truth claims. Yet the manifest image remains indispensable as the originary medium for the normative. To the extent that this normative framework does not survive, Sellars warned, "man himself would not survive." (Sellars 1991a, 18) But it is man qua rational agent, not anthropological object, which Sellars wishes to safeguard here. The manifest image remains indispensable because it provides us with the necessary conceptual resources we require in order to make sense of ourselves as

persons, that is to say, concept-governed creatures continually engaged in giving and asking for reasons. It is not privileged because of what it describes and explains, but because it renders us susceptible to the force of reasons. It is the medium for the normative commitments that underwrite our ability to change our minds about things, to revise our beliefs in the face of new evidence and correct our understanding when confronted with a superior argument. In this regard, science itself grows out of the manifest image precisely insofar as it constitutes a *self-correcting* enterprise. Indeed, for Sellars, a proto-scientific theory lies at the heart of the normative structure of the manifest image. We had to learn to postulate thoughts as unobservable inner episodes in order to explain publicly observable speech. Only in doing so did we acquire the ability to understand ourselves as rational agents operating in the concept-governed space of reasons. Once ushered into this normative dimension, we developed ever more sophisticated resources for describing and explaining what we observe in terms of what we do not observe. Thus Sellars is a resolutely modern philosopher in his insistence that normativity is not found but *made*. The rational compunction enshrined in the manifest image is the source of our ability to continually revise our beliefs, and this revisability has proven crucial in facilitating the ongoing expansion of the scientific image. Once this is acknowledged, it seems we are bound to conclude that science cannot lead us to abandon our manifest self-conception as rationally responsible agents, since to do so would be to abandon the source of the imperative to revise. It is our manifest self-understanding as persons that furnishes us, qua community of rational agents, with the ultimate horizon of rational purposiveness with regard to which we are motivated to try to understand the world. Shorn of this horizon, all cognitive activity, and with it science's investigation of reality, would become pointless. Is this to say that the manifest image subordinates the ends of enquiry to human interests? Does the manifest image predetermine our understanding of what a person is? I think the answer to both questions is no.

Sellars aligns himself with a rationalist lineage that postulates an intimate link between rationality and subjective agency. It is encapsulated in this Sellarsian dictum: "'True' is a sign that something is to be done, *for inferring is a doing*." The capacity to draw inferences requires the ability to be bound by a rule. This binding is spontaneously undertaken by a subject, not passively submitted to by an object. The agent is a subject precisely insofar as she is able to submit to a rule. Our capacity to do things with concepts presupposes that concepts can do things to us. Our grasp of a concept requires that we be gripped by the concept. But if rationality is indissociable from subjectivity, and subjectivity is synonymous with selfhood, does this mean that the capacity for rationality requires the existence of selves? Does the institution of rationality necessitate the canonization of selfhood? Not if we learn to distinguish the normative realm of subjective rationality from the phenomenological domain of conscious experience. To acknowledge a constitutive link between subjectivity and rationality is not to preclude the possibility of rationally investigating the biological roots of subjectivity. Indeed, maintaining the integrity of rationality arguably obliges us to examine its material basis. Philosophers seeking to uphold the privileges of rationality cannot but acknowledge the cognitive authority of the empirical science that is perhaps its most impressive offspring. Among its most promising manifestations is

cognitive neurobiology, which, as its name implies, investigates the neurobiological mechanisms responsible for generating subjective experience. Does this threaten the integrity of conceptual rationality? It does not, so long as we distinguish the phenomenon of selfhood from the function of the subject. We must learn to dissociate subjectivity from selfhood and realize that if, as Sellars put it, inferring is an *act* - the distillation of the subjectivity of reason - then reason itself enjoins the destitution of selfhood.

*

It is instructive to contrast Sellars' account of conceptual parity and explanatory priority between the manifest and scientific images with Jürgen Habermas' recent attempt to adjudicate the relation between the factual and normative in a controversy over the implications of cognitive neurobiology. In a 2008 paper entitled "The Language-Game of Responsible Agency and the Problem of Free-Will," Habermas invokes the Sellarsian schema in order to refute what he sees as the attempt by contemporary neuroscientists to undermine the norm of rational agency which plays such a fundamental role not only in ethical and political theorizing, but also in legal and psychiatric discourse. (Habermas 2008, 13-50) Habermas' text is largely concerned with responding to a manifesto in which eleven distinguished German neuroscientists claim that our ordinary concept of "free-will" is on the verge of being overthrown by recent advances in cognitive neurobiology. As Habermas himself notes, "neurologists expect the results of their research to lead to a profound revision in our self-understanding." (ibid., 14) According to these neuroscientists themselves: "We stand at the threshold of seeing our image of ourselves considerably shaken in the foreseeable future" (Elger et al 2004, 37). The Sellarsian resonances of both formulations are striking. But Habermas accuses the neuroscientists who would deploy the methods of natural scientific investigation to explain some of the fundamental features of our manifest self-conception - specifically, our understanding of ourselves as agents - of illegitimately extending the resources of objectification beyond their proper remit. For Habermas, the attempt to study first-person subjective experience from the third-person, objectifying viewpoint, involves the theorist in a performative contradiction, since objectification presupposes participation in an intersubjectively instituted system of linguistic practices whose normative valence conditions the scientist's cognitive activity. Attempts to interrogate the normative status of agency within the manifest image unwittingly undermine the very concept in whose name every rational investigation is ultimately undertaken, since it is the collectively instantiated norm of agency that provides the rationale for producing "truer," more accurate descriptions of reality in the first place. Thus, according to Habermas, attempts to explain agency naturalistically fail because "the social constitution of the human mind which unfolds within interpersonal relationships can be made accessible only from the perspective of participants and cannot be captured from the perspective of an observer who objectivates everything into an event in the world." (Habermas 2008, 34) Habermas characterizes this intersubjective domain of rational validity as the dimension of "objective mind," which cannot be understood in terms of the phenomenological profiles of the community of conscious selves comprised in it. Accordingly, it is the intrinsically intersubjective status of the normative realm

that precludes any attempt to account for its operation or genesis in terms of entities or processes simpler than the system itself. Neither the phenomenological nor neurobiological profiling of participants can be cited as a constituting condition for this socially "objective mind" since it is the source of the capacity for intentional objectivation presupposed by both:

> It is not the subjectivity of our conscious life that distinguishes humans from other creatures but the intentional stance and the interlocking of the intersubjective relations between persons with an objectivating attitude to something in the world. The linguistic socialization of consciousness and the intentional relation to the world are mutually constitutive in the circular sense that each presupposes the other conceptually. (ibid., 35)

The objectivity of social mind is grounded in the relation of reciprocal presupposition between an inherently *linguistic* (and hence constitutively social) consciousness and the cognitive relation to the world. For Habermas, the interdependence between language and intentionality implies not only that neither can be studied independently of the other, but more strongly, that neither can be intelligibly distinguished from the other. Here Habermas certainly echoes Sellars, whose attack on "the myth of the given" challenges the idealist attempt to ground "originary" intentionality in transcendental consciousness. Consciousness construed as originary condition of givenness becomes an unexplained explainer. This brand of transcendental idealism is inimical to naturalism, since if consciousness is the originary condition of objectivation, of which science is one instance, it follows that science cannot investigate consciousness. Upending this idealist order of explanation, Sellars roots the intentionality of the mental in socially instantiated linguistic practice. While the normative order retains a quasi-transcendental status, its linguistic embodiment allows us to understand how it is embedded in the empirical order. Thus, while Sellars maintains the irreducibly normative status of intentionality, the fact that it is always linguistically embodied allows us to investigate when or how this normative dimension might have arisen in the course of our evolutionary and social history.

Habermas, for his part, rightly emphasizes the necessity of distinguishing the normative from the natural, or reasons from causes, and accurately diagnoses the contradictions and confusions attendant upon any pre-emptive collapse of the former into the latter. But because his account is so largely reactive, unlike Sellars, he is unable to propose any positive account of the *intrication* between concepts and causes. Conflating naturalism with empiricism, Habermas upholds Sellars's distinction at the cost of eliding its scientific realist corollary, viz., that mind, and hence the normative order, possesses a neurobiological as well as socio-historical conditions of emergence. As a result, Habermas pre-emptively disqualifies by conceptual fiat every scientific attempt to describe and explain the transition from pre-linguistic to linguistic consciousness, from the sub-personal to the personal, and from neurobiology to culture. For Habermas, the explanatory resources required in order to provide such an account threaten to cost too much: they would incur a self-objectification which would irrevocably estrange us from ourselves. As he puts it: "The limits of naturalistic self-objectification are trespassed when persons describe themselves in such a way that they cannot recognize themselves as persons anymore" (ibid., 25). Such an

objectification of the human, Habermas maintains, would bring about a "fictionalization" of selfhood which would conjure "the image of a consciousness that hangs like a marionette from an inscrutable criss-cross of strings" (ibid., 24). Yet such depersonalization remains impossible, Habermas contends, because it could only come about through the attainment of a hypothetical "view from nowhere" which science cannot realize:

> The resistance to a naturalistic self-description stemming from our self-understanding as persons is explained by the fact that there is no getting round a dualism of epistemic perspectives that must interlock in order to make it possible for the mind, situated as it is within the world, to get an orienting overview of its own situation. Even the gaze of a purportedly absolute observer cannot sever the ties to one standpoint in particular, namely that of a counterfactually extended argumentation community. (ibid., 35)

This dualism of epistemic perspectives invoked by Habermas is the dualism of observer and participant. And in fact, Habermas recodes the Sellarsian distinction between manifest and scientific images in terms of a dualism of theory and practice wherein the former indexes the objectifying stance of scientific naturalism while the latter expresses subjective participation in intersubjective discourse (the "argumentation community"). Yet even as Habermas insists on the complementarity of scientific theory and discursive practice, he inscribes the former within a horizon of conceptual possibility entirely delimited by the latter. Thus, he insists, "the conceptual constitution of domains of enquiry, the construction of designs and measurements, and the experimental production of data are all rooted in pre-scientific practice" (ibid., 38). Yet as Habermas knows, there is a crucial difference between methodological priority and nomological dependence, and the fact that pre-scientific practice enjoys *chronological* precedence over scientific theorizing in no way entails that the latter is *logically* dependent upon or reducible to the former. In his determination to ward off the naturalistic dissolution of the normative, Habermas resorts to an instrumentalization of science - of the sort Sellars repeatedly warned against - which inadvertently suggests that nothing we learn about ourselves from the perspective of scientific theory could force us to revise the content of our subjective or "participatory" self-understanding. Habermas' epistemological dualism of objectifying theory and discursive practice is in many ways an exacerbation of the more familiar dualism of first and third-person perspectives in Anglo-American philosophy of mind. Ultimately, the dualism of epistemic perspectives seems to point toward the conceptual impossibility of arriving at a synoptic vision that would finally bridge the gap between the conceptual and the natural, or the subjective and the objective. What Anglo-American philosophy characterizes as the "explanatory gap" between mind and brain, or first and third person perspectives, Habermas rashly inflates into a "pragmatic contradiction" between the neuroscientist's practico-discursive reliance on intersubjectively instituted semantic norms and her conceptual disavowal of those conditions in her theoretical propositions.

*

Is it possible to describe and explain the correlation between first-person experience and neurobiological processes without lapsing into the sort of conceptual incoherence denounced by Habermas? In *Being No One: The*

Self-Model Theory of Subjectivity (Metzinger 2004; originally published in 2003, four years prior to Habermas' article), Thomas Metzinger describes and explains *in principle* how normatively regulated social interaction between conscious selves supervenes upon un-conscious, sub-symbolic neurobiological processes. Moreover, Metzinger does so by explaining how the phenomenon of selfhood, and hence the first-person subjective perspective, can be understood as arising out of sub-personal representational mechanisms. First however, it is necessary to stave off a potential misunderstanding. Although unequivocally naturalistic in its methodology and uncompromisingly "materialist" in tenor, Metzinger does not adopt the kind of straightforwardly "reductionist" strategy espoused by traditional mind-brain identity theories, whether in their strong versions, where identity is construed as obtaining between mental and physical *types*, or in their weaker formulations, where the identity in question is merely between mental and physical *tokens*.[2] Rather than postulating direct token or type identities between psychological and neurological states, Metzinger proceeds by elaborating a naturalized theory of representation wherein the latter is construed as a dynamic process involving three distinct types of state - internal representations, which are always unconscious; mental representations, which are only sometimes conscious; and phenomenal representations, which are always conscious. Furthermore, every representational state comprises a relation between a *representing* - i.e., the concrete internal state of the system - and a *represented* - the particular feature of the world or of the system itself about which the representational state carries information. In many ways, Metzinger's distinction between representing and represented corresponds to the familiar distinction between the "vehicule" and the "content" of representation. However, for Metzinger, the representing or "vehicule" does not have its boundaries at the skin of the organism but can extend out into the environment from which it extracts a represented "content." Consequently, in Metzinger's account, the representing may be defined as "internal" to the representational system even when it is constituted by spatially external events. Moreover, where much philosophy of mind tends to hypostatize the vehicule/content distinction, with the result that vehicule and content are construed as distinct entities which can then all too easily be interpreted as instances of mental or physical events respectively, Metzinger insists that representing and represented be conceived as conjoined aspects of a single informational process whose deep-structure needs to be mapped according to five distinct levels of analysis: phenomenological, representational *sensu stricto*, information-computational, functional, and neurobiological. Although each level of representational structure remains conceptually distinct, its autonomy is constrained by the minimal requirement that any "slice" of the representational process remains correlated with events at the neurobiological level. Thus, rather than trying to directly identify the mental with the physical, Metzinger maintains the relative irreducibility of these distinct levels of description, carefully distinguishing the structural properties and features specific to each, while insisting that every representational state invariably supervenes upon the neurobiological level - the guiding hypothesis being that there must always be minimally sufficient neural correlates for every representational state, even in those cases where we are not yet in a position to identify them.

On the basis of this characterization of conscious states as a variety of representational states, Metzinger is able

to propose a novel account of the nature of conscious experience as a special case of phenomenal representation in which an individual information processing system generates a *reality-model*. At its simplest level then, consciousness can be defined as obtaining whenever a representational system generates a *phenomenal world model*: "Conscious experience then consists in the activation of a coherent and transparent world model within a window of presence." (ibid., 213) Metzinger goes on to specify three minimal constraints for the experience of phenomenal consciousness:

> 1. *Presentationality*, or the generation of a window of temporal presence through which the system represents the world.
>
> 2. *Globality*, or the availability of information for guided attention, cognitive reference, and control of action.
>
> 3. *Transparency*, defined as "inversely proportional to the introspective degree of attentional availability of earlier processing stages." (2004, 165)

Transparency, the third constraint, is arguably the most significant for Metzinger's entire account. Here again, it is important to distinguish it from more familiar philosophical definitions of "transparency" in terms of the inaccessibility of vehicule as opposed to content properties (or of the properties of the representing as opposed to those of the represented). Metzinger refuses this orthodox construal of transparency because, once again, it encourages the temptation to reify the distinction between content and vehicule in terms of traditional distinctions between the mental and the physical. Thus, the mental would be defined as transparent in contradistinction to the opacity of the physical. But on Metzinger's account, it is simply not the case that the representational vehicule is a physical entity while its represented content is mental: both vehicule and content, representing and represented, are indissociable aspects of an informational continuum wherein each can switch role and serve as content or vehicule for another, higher order representation. Consequently, transparency is fundamentally a phenomenological rather than epistemological notion: phenomenal content is not epistemic content: "The transparency of phenomenal representations is cognitively impenetrable; phenomenal knowledge is not identical to conceptual or propositional knowledge." (ibid., 174) Accordingly, the fact that something is phenomenologically transparent does not entail that it is cognitively accessible to the system itself; as we shall see, the reverse is far more often liable to be the case. In fact, phenomenal transparency implies the unavailability of the representational character of the contents of conscious experience:

> Truly transparent phenomenal representations force a conscious system to functionally become a naïve realist with regard to their contents: whatever is transparently represented is experienced as real and as undoubtedly existing by this system. (ibid., 167)

Thus, in a move strikingly redolent of Kant, Metzinger characterizes what U.T. Place originally identified as "the phenomenological fallacy" - "the mistaken idea that descriptions of the appearances of things are descriptions of actual state of affairs in a mysterious inner environment" (Place 1970, 42) - in terms of the abstraction of the represented from the process of representation. Transparency understood as the occlusion of the process

of representation to the benefit of its phenomenal contents encourages the system to remain a "naïve realist" about what it experiences. It generates the subjective impression of phenomenological immediacy. As a result, phenomenal transparency, which is among the defining features of the subjective experience of conscious immediacy, is in fact "a special form of darkness." (Metzinger 2004, 169)

Once consciousness is minimally defined as the activation of an integrated world-model within a window of presence, then self-consciousness can be defined as the activation of a phenomenal self-model (PSM) nested within this world-model: "A self-model is a model of the very representational system that is currently activating it within itself." (ibid., 302) Metzinger identifies three regards in which the system may benefit from the ability to consciously represent its own states to itself:

> 1. The possession of phenomenal states clearly increases the flexibility of the system's behavioural profile by amplifying its sensitivity to context and its capacity for discrimination.
>
> 2. The PSM "not only allows a system to make choices about itself but adds an internal context to the overall conscious model of reality under which the system operates." (ibid., 308)
>
> 3. Lastly, the PSM exerts an important causal influence, not only by differentiating but also by integrating the system's behavioural profile. Thus, "as one's bodily movements for the first time became globally available as one's *own* movements, the foundations for agency and autonomy are laid. A specific subset of events perceived in the world can now for the first time be treated as systematically correlated *self-generated* events." (ibid., 309)

Through the PSM, a system becomes able to treat itself as a second-order intentional system - one capable of entertaining beliefs about its own beliefs[3] - and is thereby transformed from something merely exhibiting behaviour into an entity capable of exerting the sort of self-regulation characteristic of what we call "agency." Accordingly, given any system for which the constraints of presentationality, globality, and transparency obtain, the acquirement of a PSM will necessarily entail the emergence of a *phenomenal self*. Yet the latter is not an autonomous or independent entity but merely the represented of a phenomenal representation. Moreover, it is precisely the system's lack of access to the process through which it generates its own self-model that engenders the condition of "autoepistemic closure" whereby the represented of the system's self-representation occludes the representing that gave rise to it:

> Phenomenal selfhood results from autoepistemic closure in a self-representing system; it is a lack of information … The phenomenal property of selfhood is constituted by transparent, non-epistemic self-representation - and it is on this level of representationalist analysis that the refutation of the corresponding phenomenological fallacy becomes truly radical, because it has a straightforward ontological interpretation: no such things as selves exist in the world … What exists are information processing systems engaged in the transparent process of phenomenal self-modelling. All that can be explained by the phenomenological notion of a "self" can also be explained using the representationalist notion of a transparent *self-model*. (ibid., 337)

Ultimately, the PSM is simply the shadow cast by the occlusion of global, attentionally available information about the workings of the system. But why should this transparency have come about? Metzinger's answer is that autoepistemic closure is imposed by the need to minimize the amount of computational resources required in order to make system-related information consciously available. Transparent self-modelling provides systemic information without generating a potentially debilitating regress of recursive self-modelling, for if the system had to include every representing involved in generating its self-represented within the latter, then it would also have to incorporate within it the representing required in order to generate this new, second-order self-represented, and so on ad infinitum. Phenomenal transparency is a cheap way of minimizing the neurocomputationally exorbitant cost of representational opacity.

Metzinger concludes by summarizing his principal claim in terms of three heuristic metaphors: the neurophenomenological cave; the phenomenal map; and total simulational immersion. The first is a reworking of Plato's allegory of the cave. Recall that according to the latter, the human mind's relationship to reality is akin to that of a prisoner held captive in a cave - the prisoner has never seen anything but the shadows cast onto the wall facing her by puppet-simulacra of objects which are paraded in front of the fire that is burning behind her. In Metzinger's version of this Platonic allegory, the cave is the physical organism or information processing system as a whole; the fire its neurocomputational dynamics; the puppet-simulacra of objects its mental representings; and the shadows cast on the cave wall its phenomenal representeds. But according to Metzinger, there is no prisoner in the cave; indeed there is no-one there at all. The conscious self is not an entity but a *shadow*; not an individual object, but rather the ongoing process of *shading* through which a multidimensional neurocomputational representation is projected as a much lower dimensional phenomenal model onto the surface provide by the system's world-model. Thus the PSM is not the shadow of a captive individual, nor the avatar of a supposedly authentic or even "transcendental" subject beneath or behind the conscious individual, but rather a shadow cast by the cave *as a whole*: "It is the physical organism as a whole, including all of its brain, its cognitive activity, and its social relationships, that is projecting inward from all directions at the same time … The cave shadow is there, the cave itself is empty." (ibid., 550)

In Metzinger's second metaphor, phenomenal experience constitutes a dynamic, multidimensional map of the world. And like the maps in subway stations, the phenomenal world model features a little red arrow in it that allows the user to locate herself within the map. The PSM is analogous to this little red arrow saying "You are here:" "Mental self-models are the little red arrows that help a phenomenal geographer to navigate her own complex mental map of reality by once again depicting a subset of her own properties *for* herself." (ibid., 552) But whereas the red arrow in the subway map is *opaque* to the map user, and hence explicitly apprehended by her as a representation, the PSM is transparent: its status as a representation is occluded for the system because of the introspective unavailability of all those earlier processing stages through which it has been produced. Yet this is not to say that we are mistakenly identifying ourselves with our own PSM - there can be no question of misidentification

here since the PSM is all we are. There is no transcendental or noumenal self who could mistakenly identify itself with the phenomenal self since, as Metzinger insists, the cave is empty. But its multidimensional neural self-image generates a condition of "full immersion." Thus, in the third and last of Metzinger's heuristic metaphors, the PSM operates like a *total simulation*: "A total flight simulator is a self-modelling aeroplane that has always flown without a pilot and has generated a complex internal image of itself within its *own* internal flight simulator." (ibid., 557) The PSM is this internal image which functions as an invisible interface for the interaction between system and world. And just as the total flight simulator generates its own virtual pilot, the human brain activates its PSM when it requires a representational instrument to integrate, monitor, predict, and remember the activities of the system as a whole:

> As long as the pilot is needed to navigate the world, the puppet-shadow dances on the wall of the neurophenomenological caveman's phenomenal state-space. As soon as the system does not need a globally available self-model, it simply turns it off. Together with the model, the conscious experience of selfhood disappears. Sleep is the little brother of death. (ibid., 558)

Ultimately then, Metzinger explains the phenomenological experience of selfhood as a specific type of representational content: the self is the represented of a phenomenally transparent self-model. But it is not necessary to postulate the existence of entities called "selves" over and above the dynamic web of relations between the complex physical system known as the human organism, its internal representational economy, and its physical environment. All the salient cognitive and phenomenal data can be accounted for in terms of the PSM. Is this then to say that the notion of "the self" as an autonomous reality can be dispensed with and relegated to the dustbin of intellectual history? Before we address this question and some of the objections voiced against Metzinger's thesis, let us consider some further implications of the latter.

According to Metzinger, even if it is the case that we cannot help experiencing ourselves as "selves" and find it impossible to phenomenologically imagine selfless experience, the latter remains an *epistemic* possibility. Clearly, organisms can satisfy the minimal constraints for phenomenal consciousness (presentationality, globality, transparency) without being in possession of a PSM. Undoubtedly, many forms of animal life provide instances of selfless consciousness in this sense. But they remain incapable of generating sophisticated conceptual representations of themselves and their world. Thus, for Metzinger, the philosophically interesting question is whether it is possible to envisage systems capable of generating sophisticated conceptual representations of themselves and their world without the benefit of a PSM. Metzinger suggests that such systems are indeed envisageable, but would have to be characterized as systems whose representational models have been rendered *fully opaque*. Recall that phenomenal transparency is a function of epistemic *darkness*: for any representation, its degree of transparency is inversely proportional to the degree of available epistemic information about the representational processes that preceded its instantiation. But it is possible to imagine systems endowed with the same cognitive capacities as humans, but for whom the transparency constraint, specifically as pertaining to the PSM, would not obtain. Thus, "earlier processing stages would be attentionally available

for all partitions of its conscious self-representation; it would continuously recognize it as a representational construct, as an internally generated internal structure." (ibid., 565) Such a system would possess a system-model without instantiating selfhood. It would retain the functional advantages of possessing a coherent self-model (integration, monitoring, prediction, memory) but without experiencing itself *as* a self. It would be burdened with an additional computational load, which it would have to find some way of discharging without getting trapped into infinite loops of self-representation, but if it could find some means of solving this problem without resorting to the transparency solution, then this would indeed constitute an example of a cognitive system operating with a non-phenomenologically centred model of reality. Such a system would be *nemocentric*: it would satisfy a sufficiently rich set of constraints for conscious experience without exemplifying phenomenal selfhood. It would quite likely remain *functionally* egocentric, in order to satisfy the requirements of biological adaptation, but it would remain phenomenologically selfless. Moreover, such a system's reality-model would be richer in informational content than our own, because at every stage of processing, more information about earlier processing stages would be globally available for the system as a whole. Thus such a system would instantiate what Metzinger calls a "first-object" perspective because it would experience its own phenomenal self-model not only as a *represented* but also and simultaneously as a *representing*. It would be aware of the representational vehicule as well as of the represented content.

There is an interesting comparison to be made between this hypothetical nemocentric perspective and the transcendental perspective of pure phenomenological consciousness as effected by what Husserl called the "transcendental reduction." The goal of the latter is to "bracket off" or suspend the assumption of the autonomous reality of objects in order to isolate the ideal objectifying acts through which intentional consciousness generates its objective correlates. Obviously, in Husserl's idealist schema, this reduction is carried out by and for a transcendental subject, the better to separate the world-less realm of intentional consciousness as originary source and locus for the possibility of scientific objectification. By way of contrast, the hypothesis of the nemocentric perspective suggested by Metzinger is one in which the representational process's reincorporation into the represented object serves to foreground the sub-personal dimension of neurocomputational processing that underlies objectifying representation, and hence the objective processes through which objectivity is partly produced. Over and above its status as a phenomenological anomaly, the hypothesis of nemocentric consciousness provides a possible model for the new type of experience that could be engendered were scientists to succeed in objectifying their own neurobiological processes of objectification. The nemocentric subject of a hypothetically completed neuroscience in which all the possible neural correlates of representational states have been identified would provide an empirically situated and biologically embodied locus for the exhaustively objective "view from nowhere," which Habermas and others have denounced as a conceptual impossibility. Yet here, as Metzinger's work suggests, empirical possibility outstrips a priori stipulations of conceivability. In railing against the possibility of the mind's complete theoretical self-objectification, Habermas inadvertently reiterates the conflation of personhood as conceptual norm with

selfhood as phenomenological reality - the very confusion he initially sought to denounce. Here we have an example of what could be called "the philosopher's fallacy:" a failure of imagination paraded as an insight into necessity.[4] Habermas refuses to envisage the possibility of a convergence between self-objectification and self-knowledge because he continues to assume that self-knowledge must be knowledge *of* the self:

> [N]euroscientific enlightenment about the illusion of free will crosses the conceptual border into self-objectification … For this shift in the naturalization of the mind dissolves the perspective from which alone an increase in knowledge could be experienced as emancipation from constraints. (Habermas 2008, 24)

But what Habermas fails to see is how the genitive in the proposition "self-knowledge is not knowledge of the self" is at once subjective and objective: if the subject is not a self, then the subject who knows herself to be selfless is neither the proprietor of this knowledge (since it is not *hers*) nor its object (since there is *no-one* to know). Ultimately, Habermas' inability to articulate the distinction between theoretical objectification and discursive practice ends up promulgating a dualism of theory and practice, objective and subjective, which results from the refusal to acknowledge their interpenetration. For as Sellars so clearly saw, it is precisely the norm-governed domain of subjective practice that demands the conceptual integration of the subjective and the objective, reasons and causes, in the obligation to attain a maximally integrated understanding of the world and our position within it as creatures who are at once conceptually motivated and cause-governed. Unlike Sellars, Habermas pushes the irreducibility of the normative to the point where it generates a schism within the conceptual order in the form of a dualism of the normative and the natural. Lacking any understanding of the interplay between subjective practice and objective explanation, Habermas' account of rationality becomes internally contradictory: it seeks to defend rationality by excluding a key part of it, viz., the naturalistic explanation of empirical subjectivity, which can only increase, not compromise, our understanding of the conceptual, both in its distinction and emergence from the empirical. Disregarding the imperative to understand the latter, Habermas posits a distinction that he reifies into a substantive dualism of reasons and causes.

*

Critics have objected that the notion of "self" which Metzinger claims to have eliminated is a straw man: Hume, Kant and Nietzsche had already demolished this (supposedly) Cartesian conception of the self as an autonomous metaphysical substance. Others have responded to his work by insisting that phenomenology in the Husserlian tradition abjures precisely this metaphysical reification of the self: phenomenology construes the subjectivity of conscious experience in terms of a pre-reflective dimension of *ipseity* according to which phenomenal experience is necessarily "owned." One of Metzinger's phenomenological critics, Dan Zahavi, insists that it is in terms of the unobjectifiable "mineness" of conscious experience - which Heidegger called *Jemeinigkeit* - that selfhood ought to be understood once liberated from its metaphysical reification as *res cogitans*:

> Whether a certain experience is experienced as mine or not does not depend on something apart from the experience,

> but on the givenness of the experience. If the experience is given to me in a first-personal mode of presentation, it is experienced as my experience, otherwise not. To be conscious of oneself, is consequently not to capture a pure self that exists in separation from the stream of consciousness, rather it just entails being conscious of an experience in its first-personal mode of givenness. In short, the self referred to is not something standing beyond or opposed to the stream of experiences, rather it is a feature or function of their givenness. It is the invariant dimension of first-personal givenness in the multitude of changing experiences. (Zahavi 2005, 9)

It is this focus on the allegedly transcendental dimension of "givenness" (which is "ontological," as opposed to the merely "ontic" given) that distinguishes phenomenology from psychology, and phenomenological experience *stricto sensu* from any merely empirical cataloguing of introspectively accessible psychic states or processes. Indeed, Zahavi cites Husserl approvingly to the effect that the phenomenological domain is "neither psychic nor physical:"

> Rather, phenomenology is interested in the very dimension of givenness or appearance and seeks to explore its essential structures and conditions of possibility. Such an investigation is beyond any divide between psychical interiority and physical exteriority, since it is an investigation of the dimension in which any object - be it external or internal - manifests itself. (ibid., 14)"

Thus Zahavi insists that for phenomenology, the self is not something given - it is precisely never something given at the level of content of experience - but rather the *form* of givenness or of experience as such. This form is precisely what Heidegger called *eigentlichkeit* or "mineness:" the *owning* of experience. Consequently, Zahavi contests Metzinger's use of the PSM theory of subjectivity to explain the fracturing of selfhood and the anomalous phenomenologies involved in pathologies such as anosognosia, schizophrenia, and Cotard's syndrome. He objects that even in cases of thought insertion, where the subject experiences thoughts that she ascribes to another, she continues to own the experience, since her very estrangement from the thought reveals how, even in disavowing that the thought is hers, she continues to own the experience in which this estrangement is registered and this disavowal occurs. Thus, Zahavi insists, selfhood remains an ineluctable phenomenological feature of the form of the given, rather than of its content. The schizophrenic continues to experience alien thought episodes as occurring to her, rather than to someone else: "Rather than involving a lack of a sense of ownership, passivity phenomena like thought insertions involve a lack of a sense of authorship (or self-agency) and a misattribution of agency to someone or something else." (ibid., 6) Zahavi demotes subjective agency to the level of empirical content, the better to elevate selfhood into a formal condition of experience. Accordingly, he concludes, even schizophrenic depersonalization presupposes this irreducible proprietary relation to experience, which phenomenology identifies as this dimension of "ownness."

But *who* owns experience? What remains of the self once it has been de-substantialized and transposed to the level of form? If phenomenological selfhood pertains to the form rather than the content of experience, then what formal property (or set of properties) can we invoke to *identify* an experience as our own, or discriminate one self from another? What characteristics distinguish my

experience from yours at the level of phenomenological form? The problem is that everything that distinguishes my self from yours subsists at the level of experienced content, not the form of experiencing. Phenomenology inflates selfhood into a structurally necessary property of experience, the invariant form for the givenness of the given, when precisely what distinguishes my self from yours is something given, rather than its givenness. To insist that it is given *to* me, rather than *to* you, is simply to beg the question as to the identity of the dative, by reiterating a distinction experienced at the level of given content and projecting it back onto the form of its givenness. So what is the explanatory worth of the phenomenological postulate according to which selfhood is a formally necessary property of experience? In descriptive terms, all that distinguishes the phenomenological postulate of "mineness" as originary form from the self-model theory of subjectivity is the fact that the former stipulates as a necessary condition of experience a phenomenon that the latter derives as a conditioned experience. Instead of providing some property or set of properties, whether conceptual, qualitative, or experiential, that would mark the difference between the phenomenological structures governing the possibility of appearance and those of its phenomenal counterparts, which can be accounted for in terms of the sub-personal mechanisms mapped by Metzinger, Zahavi invokes a dimension of givenness which, although defined using all those features of phenomenal consciousness accounted for by the PSM, is nevertheless "neither physical nor psychical."[5] Moreover, the claim that this givenness provides the dimension wherein any object "whether internal or external" must manifest itself remains unpersuasive: in what sense does a saccadic eye movement or a lesion of the occipital lobe appear as phenomenologically "given" in the same way as a pub conversation or a religious experience? The fact that saccades and lesions can be turned into intentional correlates of consciousness does not make them "phenomena" in the same sense in which conversations and sensations are said to be. Just as unconscious phenomena can be viewed as intentional correlates, conscious phenomena can be turned into objects and investigated from the third person perspective. The former is no more a vindication of phenomenology than the latter is of naturalism. Playing on the inherent ambiguity of the word "phenomena," Zahavi elides the distinction between intentional and conscious phenomena and reduces the former to the level of the latter. But he adduces no argument for the claim that phenomenological "givenness" remains irreducible to psychological and/or cognitive experience; he simply stipulates it.

Ultimately, the claim that givenness itself must be accepted as an undeniable datum is merely the most radical version of the myth attacked by Sellars.[6] On the one hand, subjectivity understood as "mineness" is precisely an aspect of experience that Metzinger is at pains to describe and explain via his PSM theory. Having relinquished the metaphysical postulate of a noumenal self subsisting behind or beyond appearances, the phenomenologist cannot then maintain that the reality proper to the experiencing self is *more* than just an experience. To understand the subject as a structurally necessary condition of experience in the Kantian sense is precisely not to construe it as a self exercising a proprietary grip over its experiences, since the Kantian subject is an impersonal function, not a titled individual proprietor endowed with deeds of ownership. The relation between subjective condition and conditioned object does not map onto the relation between

proprietary self and owned experience. Questions as to the reality of experience are undoubtedly metaphysical. Zahavi denounces Metzinger's denial of the existence of selves as a dubious piece of scientistic metaphysics. But Zahavi cannot then proclaim the indubitable reality of selfhood simply because it is given as an experienced content. For as both Metzinger and Sellars point out, phenomenal transparency is not *epistemic* transparency. To insist on the epistemic authority of conscious experiences is to reiterate the dogmatic pre-Kantian postulate according to which experiences are cognitively self-authenticating. It is one thing to insist, as Descartes did, that where phenomenal seeming is concerned, doubt is inappropriate, since there can be no appearance-reality distinction of the sort subject to epistemological adjudication. But where doubt is inappropriate, so is certainty. The corollary of the admission that we cannot doubt how things seem is the recognition that we cannot be certain of it either, since certainty is doubt's epistemic obverse. It is as inadmissible to proclaim the indubitable epistemic authority of phenomenal experience as to denounce it as illusory.

Thus, just as Metzinger exposes phenomenal transparency as a kind of epistemic blindness, Sellars (like Kant before him) insists that self-knowledge is mediated by knowledge of objects. The phenomenon that Metzinger describes and explains subtends the epistemic assumption that Sellars diagnoses and analyses in his critique of the given. Zahavi reiterates this assumption when he insists that "At its most primitive, self-consciousness is simply a question of having first-personal access to one's own consciousness; it is a question of the first-personal givenness or manifestation of experiential life." (ibid., 7) Self-knowledge certainly comprises a dimension of non-inferential immediacy that endows us with a privileged epistemic access to our own internal states, but only within certain limits, since the immediacy of self-knowledge is itself the result of conceptual mediation and cannot be evoked to ratify the appeal to an allegedly intuitive, pre-conceptual self-acquaintance. The prejudice that immediacy is not the result of a mediating self-relation seduces us into absolutizing phenomenal experience. Phenomenology's absolutization of givenness as such is the most extreme variant of the myth dismantled by Sellars.

Consequently, Zahavi is no more entitled to infer the reality of selfhood from its experience than Metzinger is to deny it. Here it is important to bear in mind the distinction between different levels of analysis: *concepts are not phenomena*. The concept of the subject, understood as a rational agent responsible for its utterances and actions, is a constraint acquired via enculturation. The moral to be drawn from Metzinger's work here is that subjectivity is not a natural phenomenon in the way in which selfhood is. But Metzinger need not even deny the reality of the self (we might say that self-models are "real" in some suitably qualified sense - though justifying this would require working out a full blown metaphysics), only the phenomenological postulate of its absolute explanatory priority. He draws a metaphysical conclusion where a methodological one would be more apt: the self-model theory of subjectivity describes and explains the phenomenon of selfhood in a way that allows it to be reintegrated into the domain investigated by the natural sciences. It forces us to revise our concept of what a self is. But this does not warrant the elimination of the category of agent,

since an agent is not a self. An agent is a physical entity gripped by concepts: a bridge between two reasons, a function implemented by causal processes but distinct from them. And the proper metaphysical framework for explaining the neurobiological bases of subjective experience is that of a scientific realism rooted in an account of conceptual normativity that supervenes on, but cannot be identified with, socially instantiated and historically mediated linguistic practices.

Notes:

1. The phrase is Robert Brandom's.
2. For canonical statements of the position, see the first four papers by Herbert Feigl, U.T. Place, J.J.C. Smart and David Armstrong in Borst 1970, 33-79. See also Armstrong 1968, Feigl 1967, and Smart 1963. Donald Davidson's "Mental Events" is the classic statement of the case for token identity (Davidson 2011).
3. "Let us define a *second-order intentional system* as one to which we ascribe not only simple beliefs, desires, and other intentions, but beliefs, desires, and other intentions *about* beliefs, desires, and other intentions" (Dennett 1978, 273).
4. Daniel Dennett was of course the first to identify this fallacy.
5. The claim that for phenomenology consciousness is neither psychical nor physical is of course made by Husserl in the second volume of his *Logical Investigations*. Zahavi (2005) cites it approvingly on p. 13.
6. "Many things have been said to be 'given': sense contents, material objects, universals, propositions, real connections, first principles, *even givenness itself*" (Sellars 1991, 127; my emphasis).

References:

Armstrong, D. M. 1968. *A materialist theory of mind*. London: Routledge.

Borst, C.V., ed. 1970. *The mind/brain identity theory*. London: Palgrave MacMillan.

Davidson, Donald. 2001. Mental events. In *Essays on action and events,* 2nd edition, 207-228. Oxford: Clarendon Press.

Dennett, Daniel. 1978. *Brainstorms: philosophical essays on mind and psychology*. Harmondsworth: Penguin.

Elger, C. E., A. D. Friederici, C. Koch, H. Luhmann, C. von der Malsburg, R. Menzel, H. Monyer, F. Rösler, G. Roth, H. Scheich, and W. Singer. 2004. Das Manifest: Elf führende Neurowissenschaftler über Gegenwart und Zukunft der Hirnforschung [The Manifesto: Eleven prominent neuroscientists on the present state and future of brain Research]. *Gehirn und Geist,* 6, October 13, http://www.gehirn-und-geist.de/artikel/839085.

Feigl, Herbert. 1967. *The "mental" and the "physical."* Minneapolis: University of Minnesota Press.

Habermas, Jürgen. 2008. The language game of responsible agency and the problem of free will: How can epistemic dualism be reconciled with ontological monism? *Philosophical Explorations* 10, 1: 13-50.

Metzinger, Thomas. 2004. *Being no one: The self-model theory of subjectivity*. Cambridge: MIT Press.

Place, U.T. 1970. Is consciousness a brain process? In *The mind/brain identity theory*, ed. C.V. Borst. London: MacMillan.

Sellars, Wilfrid. 1991. Empiricism and the philosophy of mind. In *Science, perception, and reality*, 127-196. Atascadero: Ridgeview Publishing Co.

___. 1991a. Philosophy and the scientific image of man. In *Science, perception, and reality*, 1-40. Atascadero: Ridgeview Publishing Co.

___. 1991b. Truth and "correspondence." In *Science, Perception, and Reality*, 197-224. Atascadero: Ridgeview Publishing Co.

Smart, J.J.C. 1963. *Philosophy and scientific realism*. London: Routledge and Kegan Paul.

Zahavi, Dan. 2005. Being someone. In *Psyche: An Interdisciplinary Journal of Research on Consciousness*, 11.5: 1-20. http://the-assc.org/files/assc/2611.pdf (accessed October 1, 2011).