

Lost in Phenospace. Questioning the Claims of Popular Neurophilosophy¹

Jan Slaby

Appeared in: Christoph Lumer/Uwe Meyer (Eds.). *Geist und Moral. Analytische Reflexionen für Wolfgang Lenzen*. Paderborn: mentis 2011, pp. 35-53.

Philosophical critique can help clarify issues surrounding the burgeoning cognitive neurosciences and thereby contribute constructively to the field. ›Neurophilosophy‹ has become a fashionable chorus to some of the new experimental approaches to the neural foundations of human traits and capacities. Sometimes, self-declared neurophilosophers – a group that one might call the ›cheerleaders‹ of the new mind sciences – aspire to deliver overarching visions of the human mind and its place in nature.² These visions almost rival older metaphysical approaches in their scope and generality. However, the impression that a broad philosophical conception follows directly from the *experimental results* of neuroscience is untenable. Unfortunately, quite often there is no proper distinction between what seems an obvious lesson derived from experimental findings and what is instead a philosophical assumption invested into the subject matter beforehand. Some of this seems to be the case with a recent, popular book on the subject, Thomas Metzinger's ›The Ego-Tunnel‹ (2009).

In the spirit of Wolfgang Lenzen's sharp and thoughtful critical analyses of neuroscientific and neurophilosophical work, I want to take issue with some of the claims that Metzinger makes in his book, in the hope of helping to clarify the specific relevance of new scientific approaches to the human mind.³ The main focus will be on claims concerning experience and intentionality, but I will also address the ethical aspirations that Metzinger expresses in the final part of his book. In a rather awkward way, Metzinger proposes a ›new approach to ethics‹ as a consequence of recent neuroscientific findings. Hereby, he enrolls in the camp of those commentators of neuroscience who claim a broad social relevance of the new science of the mind. Claims like these need to be viewed with some suspicion, as my arguments below will make clear.

¹ An earlier version of this text grew out of a close collaboration with Jan-Christoph Heilinger. Part 4 below, in particular, contains ideas and arguments developed by Heilinger. I thank him for allowing me to publish this material as part of this contribution.

² Inaugurators of the movement and authors of canonical writings of neurophilosophical texts are of course Patricia and Paul Churchland; see, for instance, P.S. Churchland, 1989 and 2002.

³ See, for instance, Lenzen, 2004; 2005, and especially 2006.

Before I begin discussion of Metzinger's book, I will briefly outline the motivation, aims and scope of a new interdisciplinary initiative that goes by the name of ›Critical Neuroscience‹. This project, which has in part been initiated and developed at the Institute of Cognitive Science at the University of Osnabrück, provides a framework for critical reflections like those developed here. It aspires to bring together competences from different scientific and meta-scientific perspectives in order to help understand, assess and improve practices and theories in the cognitive neurosciences. Philosophical critique is a crucial ingredient of the project, but at the same time the philosophical perspective is significantly broadened and enriched by studies of the social and institutional realities of neuroscientific practice, its historical developments and anthropological contexts. Given today's widespread fascination with brain-related approaches in science and popular culture, a project like this seems to be of increasing relevance.

1. The Project of ›Critical Neuroscience‹

Critical Neuroscience attempts to better understand, explain, contextualise and, if necessary, critique current developments in and around the cognitive – including the social, affective and cultural – neurosciences. A key aim is to help create the competencies needed to deal responsibly with a range of concerns: the impact of new brain-technologies; premature application of potentially dangerous or unreliable devices and chemical agents; intellectual dominance of certain forms of ›neurocentric‹ thinking, including problematic, one-sided or outrightly mistaken philosophical theories; dangers of narrowly brain-oriented approaches to medical diagnosis and treatment etc.⁴ The project addresses scholars in the humanities and, importantly, neuroscientific practitioners, young researchers and students.

Key questions addressed by *Critical Neuroscience* are the following: are the enthusiasts in fact right when they announce that neuroscience will very soon have ›wide-ranging effects upon society‹, or are we collectively overestimating its impact? Maybe the increased focus upon the brain and brain-related scientific approaches distracts attention away from other drivers of social and cultural change, such as economic or political developments. Those who have an interest in over-selling their own work might intentionally overstate the importance of the brain and its new sciences. So how and via what channels is neuroscience *in fact* interacting with contemporary conceptions of subjectivity, identity, and well-being?⁵

⁴ For an initial outline of the project's aims and structure, see Choudhury, Nagel & Slaby (2009) and also Slaby (2010). A collection of essays on Critical Neuroscience is scheduled for publication in early 2011 at Wiley-Blackwell, see Choudhury & Slaby (*forthcoming*). The project is in full swing as an independent research and project group operating from Berlin, Germany. See www.critical-neuroscience.org.

⁵ On this, see Dumit (2004), Rose (2006) and Joyce (2008).

What are the dominant ›styles of thought‹ that have emerged in and around the neurosciences and in the new hyphenated ›neuro‹-disciplines (such as neuro-economics, neuro-law, neuro-education, neuro-theology or neuro-aesthetics)?⁶ Not least: how is neuroscience institutionally and politically entangled with powerful social players such as pharmaceutical companies, funding agencies, policy makers, parts of the popular media?⁷

As a further step, *Critical Neuroscience* strives to make the results of these assessments relevant to the practice of cognitive neuroscience itself. What difference would it make to scientific practice if neuroscientists themselves were involved, from the outset, in the analysis of contextual factors, historical trajectories, conceptual difficulties and potential consequences in connection to their work? The hope is to engage in collaborations focused on specific themes of societal relevance: for example, studies of mental illnesses such as depression, the investigation of social pathologies of various kinds (such as alienation in work and life environments, violence, attention problems), or ideas and popular conceptions of well-being, to name just a few. Our understanding of these phenomena can of course be enriched by neuroscientific approaches, but it seems obvious that neuroscience cannot provide all that there is to know about these matters. First applications of the framework of *Critical Neuroscience* to empirical issues show promising results, for instance in approaches to addiction and adolescence.⁸

In terms of its theoretical foundations, *Critical Neuroscience* draws on a variety of sources, among them Frankfurt School critical theory, Bruno Latour's actor-network theory (see Latour 2005), and Foucauldian approaches that analyse the entanglement of knowledge production and social power (see Rose 1996; 2006). What results is a critical toolbox that helps observers to look beyond the outer appearance and official declarations of scientific practitioners and institutions, and to investigate social influences, historical developments and tacit interests and agendas that shape scientific perspectives. In its explicitly practical orientation, *Critical Neuroscience* strives to close the gap between science students and actual scientific practice. For this reason, the project's activities are centred on teaching activities, workshops, and public events with the aim of bringing diverging perspectives into fruitful tension. The present paper is broadly representative of the aims of the project; however it is

⁶ That it is no longer a style of blatant neuronal reductionism and methodological individualism is persuasively argued by Pickersgill (2009). See also Abi-Rached & Rose (2010).

⁷ More on this in Choudhury, Nagel & Slaby (2009), Slaby (2010) and in Choudhury & Slaby (*forthcoming*).

⁸ For a *Critical Neuroscience* approach to addiction, see Campbell (2010), for work that challenges reductionist approaches to the 'adolescent brain' while at the same time constructively engaging issues surrounding adolescence from various perspectives, see Choudhury (2009).

predominantly focused on philosophical arguments and places less emphasis on the social and political climate that surround current developments in the neurosciences.

2. Metzinger's Neurophilosophy of Mind and Self

In his popular science book *The Ego-Tunnel – The Science of the Mind and the Myth of the Self* (2009), Thomas Metzinger aspires to present an over-arching vision, the big picture towards which, in his opinion, the neurocognitive sciences have been driving in recent years. In the manner of the genre of popular scientific syntheses, he makes the full journey – from laboratory results, to philosophical theory, to social application and ethical reflection: neuroscience, neurophilosophy and neuroethics are all covered. This is quite symptomatic of a recent trend: the field seems not to be content with merely providing theoretical insights into the functioning of the brain, it is also extremely eager to spell out – and popularise – its alleged ›societal and ethical relevance‹.

In the first, longer part of this essay I will take issue with the central theoretical claim Metzinger makes. Afterwards I will have a brief look at Metzinger's case for a new ›consciousness ethics‹ as a version of ›neuro-ethics‹. It is here, in Metzinger's ethical musings, that one can see a powerful new tendency at work: the push to claim social and ethical relevance, moreover a relevance of a ›revolutionary‹ kind, for the recent empirical and theoretical developments in neuroscience. As I hope to show, Metzinger works with a rather awkward conception of ethics, in which we find a focus on *states of consciousness* as the unit of ethical concern. This diverges significantly from common practice and introduces a number of difficulties. As regards the general trend to extend assessments of current neuroscience into the ethical and social realms, I will argue that this is a result of a gross over-estimation of the scientific and philosophical significance not only of the findings of neuroscience made so far, but also, and more importantly, of the potentials it has for future developments. While of potentially significant medical relevance, neuroscience simply does not operate on the level at which ethical guidelines are formulated or the fate of human societies is decided. It is important to correct this wrong impression as it threatens to negatively affect the self-understanding of practitioners as well as the public perception of neuroscience.

Of course, I cannot do justice to the full scope of the rich materials that Metzinger presents in his well-written and resourceful book. In particular, I have to omit the interesting discussion surrounding new research on out-of-body experiences, dreaming, and other new

developments in neuroscience. I move straight to the heart of the matter, to the theory of the Ego-Tunnel, Metzinger's neuroscience-inspired theory of the self and its relation to the world.

The book starts with a rather puzzling statement that immediately presents the reader with one of the key claims to be found in Metzinger's account:

»In this book, I will try to convince you that there is no such thing as a self. Contrary to what most people believe, nobody has ever *been* or *had* a self. [...] to the best of our current knowledge there is no thing, no indivisible entity, that is us, neither in the brain nor in some metaphysical realm beyond this world.« (ET, 1)⁹

Metzinger here denies the existence of ›a self‹, but he somehow fails to distinguish this seemingly radical claim from the trivial one that the self is ›no thing, no indivisible entity‹. Only someone untrained in philosophy – someone unable to see the mistake in nominalizing the personal pronouns so as to literally assume the existence of *an entity* called ›the Self‹ (writ large, as it were) that is somehow distinguishable from the person as a whole – might find this claim really informative. Unfortunately, Metzinger indeed seems to play on this ambiguity, not merely for rhetorical reasons. For instance, in section 8 of his book, he again presents it as a substantial claim and as the result of his reflections, that there »is no self« and that »[w]e must face this fact: We are *self-less* Ego Machines« (p. 208 – emphasis in the original). He thereby creates the impression of a substantial philosophical thesis when in fact all there is to the claim is that what we call the ›self‹ is a dynamic process of interaction of a human organism with its environment. I suspect that it is no accident that Metzinger doesn't mention the names of any authors whom he is arguing against in this part of his theory: apart from the philosophically uninitiated or linguistically careless, there is just no opposition to the claim that what a person ultimately is, is »no little man inside the head« (p. 208).

Metzinger goes on to outline a thorough representationalism pertaining to both world and self. In fact, most of the time, he even uses the word ›simulation‹ instead of ›representation‹, leaving no doubt about the radical nature of his thesis: »First, our brains generate a world-simulation, so perfect that we do not recognise it as an image in our minds. Then, they generate an inner image of ourselves as a whole.« (ET, p. 7). It is not just a claim to the effect that access to the world is mediated by representations, but rather the much more radical claim that what we perceive as the world and as ›ourselves‹ is in fact a *simulation* generated by the brain – a projection into ›phenospace‹, which one probably has to imagine as a kind of inner video screen. With this, Metzinger then elaborates upon the nature of the self-simulation thereby created:

⁹ In all my references to Metzinger's book I will use the abbreviation 'ET' for *The Ego Tunnel*.

»The internal image of the person-as-a-whole is the phenomenal Ego, the »I« or »self« as it appears in conscious experience; therefore, I use the terms ›phenomenal Ego‹ and ›phenomenal self‹ interchangeably. The phenomenal Ego is [...] the content of an inner image — namely, the conscious self-model, or PSM. By placing the self-model within the world-model, a center is created. That center is what we experience as ourselves, the Ego. [...] We are not in direct contact with outside reality or with ourselves, but we do have an inner perspective. We can use the word ›I.‹ We live our conscious lives in the Ego Tunnel.« (ET, p. 7)

The metaphor of the ›Ego Tunnel‹ is adopted from virtual reality engineering where we sometimes find the term ›reality tunnel‹. This means that an artificial or highly selective environment is created, often in the context of computer or video games or simulation devices. So again, the point is this: according to Metzinger, the world as we perceive it is not in fact ›the real world‹, but rather a brain-generated phenospace, a highly selective projection onto a mental screen within each of us. It is quite refreshing to see the explicitness with which Metzinger formulates this position and draws consequences from it:

»The conscious brain is a biological machine — a reality engine — that purports to tell us what exists and what doesn't. It is unsettling to discover that there are no colors out there in front of your eyes. The apricot-pink of the setting sun is not a property of the evening sky; it is a property of the internal model of the evening sky, a model created by your brain. The evening sky is colorless. The world is not inhabited by colored objects at all. It is just as your physics teacher in high school told you: Out there, in front of your eyes, there is just an ocean of electromagnetic radiation, a wild and raging mixture of different wavelengths. Most of them are invisible to you and can never become part of your conscious model of reality. What is really happening is that the visual system in your brain is drilling a tunnel through this inconceivably rich physical environment and in the process is painting the tunnel walls in various shades of color. *Phenomenal color. Appearance.* For your conscious eyes only.« (ET, p. 20)

It is no accident that Metzinger invokes Plato's *Allegory of the Cave* at this point. According to his theory, we are in the same situation as the enchained prisoners in Plato's story, only that the cave is our mind, or rather our ›mental projection screen‹, and the ›shadows‹ dancing at the opposite wall are brain-generated simulations:

»The wall [of Plato's cave] is not a two-dimensional surface but the high-dimensional phenomenal state space of human Technicolor phenomenology. Conscious experiences are full-blown mental models in the representational space opened up by the gigantic neural network in our heads—and

because this space is generated by a person possessing a memory and moving forward in time, it is a tunnel.« (ET, p. 23)

Everything we perceive, everything we know about, including ourselves, is in fact a part of ›virtual reality‹ generated by the brain.

Within the space of this essay, I can problematise only one aspect of this view, but I hope to move close to the centre of the matter with it. I take my hint from one telling phrase in one of the above quotations: »The world is just as your physics teacher in high school told you: an ocean of electromagnetic radiation, a wild and raging mixture of different wavelengths.« I will ask only one question about this: what justification is there for *prioritising* this presumably physical conception of reality over all other conceptions, and especially above the manifest image of the world characteristic of pre-scientific common sense? More precisely, assuming the truth of Metzinger's story about brain-generated virtual reality, how could we ever *know* that this physicalist story is the true story about reality, if in fact all that we are ever in touch with is a reality simulation – *mere appearances*? If we could indeed be »brains in a vat«, as Metzinger at one point happily suggests (ET, p. 21), how could we have ever managed to know with certainty what the world ›out there‹, outside the vat environment, is really like?

Thus, Metzinger's theory is exposed as incoherent. It places us in a position that, if we in fact were in it, we could have no knowledge about. In this way, his story violates a Kantian condition (one that Kant, however, himself violated from time to time): that to delineate the limiting conditions of experience, we cannot move beyond an outer boundary of possible experience and draw the borderline from both sides, as it were.¹⁰ At the heart of the problem lies a kind of scientism that is a piece of dogmatic metaphysics in disguise: it is a symptom of an exaggerated confidence in the epistemic power of the sciences. Somehow, according to Metzinger, the sciences achieve a kind of epistemic success – viz. epistemic access to reality – that by far exceeds the epistemic success that normal human subjects are ever capable of achieving. Given the fact that science obviously is a human endeavour and, as such, at any time tied to human epistemic capacities, it is highly doubtful whether science would be able to operate from such a privileged epistemic standpoint. This might suffice as a short preview of a key argument against Metzinger's account. I will now turn to some of the relevant details.

¹⁰ *Locus classicus* for this kind of characterisation – and subsequently criticism for violating its own principle – of Kant's endeavour is of course Strawson (1966).

The problem just outlined becomes most evident in the awkward theoretical status of the very organ around which Metzinger's theory revolves: the brain. Since the brain anchors the whole account as that which presumably generates all those simulations of a world and of ourselves, it has to be ›real‹ in the classical, non-representational sense of the term. But on the other hand, since it obviously also figures as an object of everyday perception, all Metzinger can claim is that it is also a simulation in phenospace – a factual (that is, transcendent) reality behind its mental appearance remaining unknown to us. Thus, when speaking about the brain as an object of scientific investigation, Metzinger has to assume that science has enabled us to leave the confines of our individual phenospaces and somehow access reality as it is in itself.

One can attack this thought from two sides. First, if indeed the brain as discovered by science is ›real‹ in the transcendent sense of the term, then it is hardly convincing that we stop there, claiming that of all we can see and perceive, *only* one single object, the brain, is ›truly real‹, not just a representation, perceived as it is in itself. This has been nicely and persuasively argued by Wolfgang Lenzen (2006): certainly, a working brain needs to be connected to a functioning body in order to function properly – but can we really say that a *real* brain can be connected to a *virtual* body? And so on – a functioning body needs oxygen and nourishment to survive and operate, so it has to be embedded in the right kind of environment; this environment, in turn, cannot just consist of simulations in phenospace because simulations are not nourishing. Thus, it has to be a *real* environment that is *really* nourishing the *real* body connected to the *real* brain. Conclusion: if the brain – as we (subjectively? collectively?) know it – is indeed real in the classical sense of the term, then pretty much everything else must be real in this sense as well.

Metzinger might block this somewhat trivializing line of argument by claiming, as he certainly does, that science does indeed have the capacity to discover reality as it is in itself. After all, *if* the physics teacher is indeed *right* about reality, then science, notably physics, must have succeeded in breaking free from the veil of appearances, reaching out to transcendent reality. But now a number of pressing questions come up: which reality is the ultimate reality? If fundamental physics is right about wavelengths and electromagnetic radiation, isn't then the brain that current neuroscience investigates again relegated to the status of a ›mere appearance‹? Wouldn't we be forced to stop taking neuroscience findings seriously until they are fully reduced to fundamental physics? A reduction that of course might never happen simply because of the enormous systemic complexity of brain organisation. Again, basing one's account on the reality of the brain as described by current neuroscience would be exposed as unfounded.

But a more fundamental question is this: how has science managed to achieve the remarkable feat of accessing ›true‹ reality in the first place? How have our scientists collectively managed to do what no individual and no non-scientific community ever achieved, namely, to break out of their phenospaces to glance at the world beyond subjective world simulations? It surely doesn't suffice to say, as Metzinger does at one point (ET, p. 8), that science is intersubjective, communicative, and systematically organised, *thereby* transcending individual perspectives. Because it seems evident enough that our pre-scientific lifeworld, our collective human practices, are also intersubjective, communicative, and systematically organised in intricate ways. Metzinger needs some factor that is *unique to science* (in a strong enough sense) that is intelligibly able to carry scientific practitioners beyond their internal representations, out of their Ego-Tunnels into reality. The problem is, every significant feature of science capable of achieving this feat clearly has a corresponding feature in our lifeworld. As Quine famously pointed out: »Science is a continuation of common sense« (Quine 1953, 45). Experiments are a systematic extension of meaningful human action; theories are refined, systematised belief systems; scientific instruments are specific, highly refined tools; scientific observation and measurement is thoroughly dependent upon pre-scientific observation capacities; the scientific community is a well-organised subset of larger human communities that are often equally well-organised; and finally, scientific rationality is no different in principle from rationality at large, which is a human form of practice and a social institution that, to put it carefully, co-evolved with scientific practices in a long historical process. Thus, there is simply no such factor that singles out science, either in Metzinger's account or anywhere else. Moreover, science is so deeply entangled with other human practices and forms of understanding that the instrumental and epistemic achievements of science have immediate ramifications outside scientific practice proper. There is constant exchange between science and society at all sorts of levels, so that it is highly implausible that science could be *utterly different* in such a crucial respect as epistemic access to reality. This view of a broad (and, one might add, rather unspectacular) continuity between science and the rest of human affairs seems to come close to consensus in the philosophy of science and in science studies of the 20th century.¹¹ In the end, Metzinger's faith in the power of science is exposed as no more than an article of faith, unsupported by any of his own theoretical claims and thoroughly out of step with sensible theorising on the nature of scientific world-disclosure. The theory of the Ego-Tunnel is based on an unacknowledged, ungrounded conviction that traditionally goes by the name of ›scientism‹.

¹¹ Nicely articulated, for instance, by Rouse (2002) and Kitcher (2003).

Almost needless to say, the actual claims of his theory are thereby exposed as utterly misguided. If one manages to see beyond the Cartesian heritage of radical mental ›simulationism‹, one may discover that a quite different picture of how brain, organism, and environment interact to produce human experience is currently emerging in the Cognitive Sciences. Crucially, this alternative outlook is also supported by some findings from the neurosciences. The alternative conception goes by various names: ›embodied and embedded cognition‹ – ›enactivism‹ – ›the extended mind‹, and so forth (see, for instance, Varela et.al., 1992; Clark 1997 and 2008; Gallagher 2005; Thompson 2007; Noë 2005; 2009).

In a recent book that parallels Metzinger's ›The Ego-Tunnel‹ in terms of broad accessibility and stylistic elegance, American philosopher Alva Noë has outlined this very different approach (Noë 2009).¹² Noë explains the embodied, embedded theory of the mind – a theory that is centred among other things around the image of ›openness to the world‹. This image invites us to embrace the natural-seeming thought of direct perceptual access to reality for all those cases where one is not manifestly misled by one's senses. Instead of being confined within one's skull and merely enjoying the brain-generated scam of an ›out-of-brain-illusion‹ (Revonsuo 2003), we can reinstate the world as by and large congruent with what it seems to us perceptually. All those not under the spell of the Cartesian picture of mental simulacra on an inner presentation screen as the *only possible* mode of experience should grasp instantly that the image of ›openness to the world‹ is so natural and enthralling that it alone, without further considerations in its favour, could almost provide the basis of a *modus tollens* against Cartesian thinking.¹³

However, I am unable to consider this further as my aim here is to discuss Metzinger's view in its full scope, instead of exploring alternatives to it. Next, I will turn to a set of ideas that mark the turn, within Metzinger's book, from theoretical claims to ethical considerations.

3. Postbiotic Ego-machines and the Ethics of Artificial Consciousness

One of the more spectacular sections of Metzinger's book is entitled ›Artificial Ego-Machines‹ (ET, pp. 187-205). It is the section leading up to the plea for a new approach to ethics, which Metzinger calls a ›consciousness ethics‹. The storyline is simple: first,

¹² Even the title of Noë's book tells much of the story: *Out of our Heads. Why you are not your brain and other lessons from the biology of consciousness*.

¹³ An argument along these lines is developed in great detail by Marcus Willaschek (2003). If philosophical considerations suggest that we are 'not in touch with the real world', then this first and foremost speaks against these philosophical considerations – simply because the common sense assumption of direct access to the real world is so overwhelmingly established in everyday practice.

Metzinger claims that it is possible, given what we currently know about the brain, to construct machines that are conscious and moreover possess a form of self-awareness; thus machines that can be said to possess what Metzinger calls an *ego* (pp. 190-95). Second, Metzinger emphatically warns us that we should do everything within our power to refrain from constructing those machines, because constructing them would likely increase the amount of suffering in the universe. Operating in a broadly utilitarian perspective, Metzinger assumes we should strive to minimise the amount of suffering in the world.

There are two quite remarkable claims here: first, that it is indeed so much as *technically possible* to construct an artificial consciousness that qualifies as a form of subjectivity. Second, that constructing conscious machines would very likely amount to constructing machines that would more or less continuously suffer. Metzinger thinks that at least the first generation of artificial Ego Machines would most probably be designed so imperfectly that, given their assumed level of cognitive and affective self-awareness, they would suffer most of the time. He illustrates the point by drawing the outrageous comparison to mentally retarded human infants (ET, p. 194). In effect, first-generation artificial Ego Machines would be as imperfectly developed as mentally retarded infants and for this reason more or less *constantly in pain* (that mentally impaired infants quite likely won't suffer constantly doesn't occur to Metzinger, who here, in effect, comes close to suggesting that retardation or disability is life not worth living. Shame on him!). An additional reason for the Ego Machine's plight is the fact that they are exploited by their human creators in the name of scientific advances, joining laboratory animals in the ranks of creatures having to suffer for the sake of human knowledge production. Given their alleged level of self-awareness, insight into this predicament adds to the suffering of the artificial Ego Machines (ET, pp. 195-6).

Since everything hinges on the first claim, I will only discuss this claim in the following. The second claim – that first-generation Ego Machines might be suffering more or less constantly – seems highly speculative to say the least. Without knowledge of the technical details of Ego Machine construction it is impossible to assess it properly. Since I will show that the first claim is untenable, discussion of the second one becomes hypothetical anyway. Besides, the more than bizarre comparison of first-generation Ego Machines with mentally retarded children speaks volumes all by itself – I won't dwell on this any longer and let readers draw their own conclusions.

Of course, technoscientific science fiction is a perennial seller in the popular books department. Metzinger happily indulges in the sport of jumping from reports of some state-of-the-art science (bio-robotics, artificial life, and evolutionary robotics among others) to

speculation about ›*what is certainly possible*‹. However, the gap that he thereby seamlessly attempts to close is one that is so wide that it is quite surprising that he makes these remarks almost casually and in passing, more or less hidden away in some back section of his book – instead of making them the core theme headlining his account. Artificial consciousness? What would be an outright scientific and philosophical sensation, by almost anyone’s standards, is to Metzinger little more than a technical problem (›the devil is in the details‹, ET, p. 189). Artificial self-awareness, genuine emotions and interests, insight into one’s finitude and mortality, capacity to suffer – for Metzinger, if not outrightly ›easy‹, then at least only a matter of time before successfully implemented. The trick is done, not via any new theoretical insight, but by assuming that it is not a problem in principle to build (or evolve) an inclusive representation of the world – a world-model – into a machine or into a hybrid ›postbiotic‹ entity. Giving the right kinds of representations to the system – that is basically all we have to do. If a comprehensive representation of the outside world is in place, the decisive additional ingredient, according to Metzinger, is then just this:

›If a system can integrate an equally transparent internal image of itself into this phenomenal reality, then it will appear to itself. It will become an Ego and a naive realist about whatever its self-model says it is. The phenomenal property of selfhood will be exemplified in the artificial system, and it will appear to itself not only as *being someone* but also as *being there*. It will believe in itself.« (ET, p. 191-2)

For all intents and purposes, this is not *solving* the problem of consciousness and subjectivity, but moving into a state of denial. At best, it is a version of stating the task instead of showing how it can be accomplished. If Metzinger were right about this, we would face a rather grand intellectual riddle: why did generations of philosophers and scientists, among them the sharpest minds of our day, keep thinking that it is a substantial philosophical and scientific challenge to figure out how a physically implemented structure can give rise to consciousness of the world and of itself, how a subjective perspective on a world is possible within a naturalistic outlook, and how the experience of ownership of thought and sensation and authorship of action is realised in physical systems? More blatant still becomes the denial of the problems at hand when Metzinger moves from stating the problem in representationalist terms to the capacity for suffering in those self-conscious artificial systems:

›Note that this transition turns the artificial system into an object of moral concern: It is now potentially able to suffer. Pain, negative emotions, and other internal states portraying parts of reality as undesirable can act as causes of suffering only if they are consciously owned. A system that does not appear to itself cannot suffer, because it has no sense of ownership. A system in

which the lights are on but nobody is home would not be an object of ethical considerations; if it has a minimally conscious world model but no self-model, then we can pull the plug at any time. But an Ego Machine can suffer, because it integrates pain signals, states of emotional distress, or negative thoughts into its transparent self-model and they thus appear as someone's pain or negative feelings.« (ET, p. 193)

By ›this transition‹, Metzinger simply refers back to the last step of his exposition, viz. the addition of an ›internal image of itself‹ to the artificial world-model. By this move alone, he now claims, we ensure the full implementation of the capacity for pain, negative emotions and suffering. But that is exactly the riddle: *why* is pain in fact *painful*, why does it *hurt*, indeed, why does it ultimately result in the existential predicament of *suffering*, when all there is to pain is the representation of some state of affairs as ›undesirable‹, that is, negative for the system? What does ›undesirable‹ mean here – what *can* it mean for such systems? Or to put it differently: where is the ›concernedness‹, the capacity to have something matter to one? Without these, the idea of suffering isn't so much as intelligible. The only move Metzinger has to offer is to accuse of obscurantism all those who keep expressing their puzzlement about how representing the world or processing information could give rise to a qualitative dimension of experience.¹⁴ It won't work to close a gap in understanding by denying its existence without actually showing that it isn't there. Ultimately, the section on artificial Ego Machines turns out to be just another expression of Metzinger's convictions, not an argument in their favour.

4. The Call for a New Ethics¹⁵

»We can definitely increase our autonomy by taking control of the conscious mind-brain, exploring it in some of its deeper dimensions. This particular aspect of the new image of humankind is good news. But it is also dangerous news. Either we find a way to deal with these new neurotechnological possibilities in an intelligent and responsible manner, or we will face a series of historically unprecedented risks. That is why we need a new branch of applied ethics — consciousness ethics. We must start thinking about what we want to do with all this new knowledge — and what a good state of consciousness is in the first place.« (ET, p. 218)

¹⁴ This accusation is implicit in the hilarious faked 'interview' with the 'first postbiotic philosopher' in part 7 of *The Ego Tunnel*, see ET, 201-05.

¹⁵ This part of my text is the result of a close collaboration with Jan-Christoph Heilinger, who deserves credit for most of the argumentative substance of the following text.

Metzinger's argument for his suggested ›consciousness ethics‹ goes like this: first, he claims that we are increasingly able, technologically or pharmacologically, to alter our brain states and our conscious experiences. He provides many examples for this kind of ›phenotechnology‹ (ET, p. 222), focussing on the (experimental) use of psychoactive substances. This might have dangerous effects, for instance because »many such experiments include the phenomenology of certainty and automatically lead to the conviction that one is *not* hallucinating« (ET, p. 220). The emerging consciousness technologies are thereby shown to have both an ethical and a political dimension (ET, p. 222). That is why, Metzinger concludes, we need a consciousness ethics to regulate which brain states and which conscious experiences we want to bring about. The key question of this novel approach to ethics, then, is this: »Which brain states should be legal? Which regions of the phenomenal-state space (if any) should be declared off-limits?« (ET, p. 229) Metzinger calls this the assessment of »the ethical value of various kinds of subjective experience *as such* ... the rational search for a normative psychology or a normative neurophenomenology« (ET, p. 233).

Usually, ethics deals with actions, not with states.¹⁶ Metzinger focuses on conscious states as the unit of ethical concern. In so doing, he suggests a significant shift in moral theory. He gives several examples to substantiate the ethical and political dimension of phenotechnology. The first one goes like this: »Should it be legal, for instance, to let children experience their parents in a drunken state?« (ET, p. 222). This example is telling in many ways.

If Metzinger sees a potential moral problem in parents being drunk – and potentially aggressive towards their children – and seen by their children in this state, then he should in the first place ask whether parents should be allowed to *act* in a way that brings about these situations. As having children has obviously nowhere brought about a *legal* ban on drinking (which is in itself usually not considered a moral problem), the *moral* question to ask could be: should parents be allowed to booze when they risk being seen by their children? It is hard to say whether this really qualifies as an interesting moral question in a somewhat non-trivial sense. It seems rather to be a question of prudence. Boozing per se is not necessarily a morally bad thing and even if a parent would be tipsy or even completely drunk on an occasion, as long as he or she is not *behaving* aggressively or *immorally*, the fact of being tipsy or drunk per se does not seem to constitute a moral problem. What matters are the

¹⁶ Even the Aristotelian virtue ethics approach is no exception: ethics in this understanding is not about providing rules for individual actions, but about assessing the overall quality of a human life. In this approach, the ultimate aim consists in living a ‘worthwhile’ life, which can be predicated on the form of life – and its inherent dispositions to act in certain ways – which one chooses (e.g. *bios theoreticos* – the life of a philosopher – or *bios politicos* – the life of an engaged citizen – and so forth).

possible *actions* resulting from a state of drunkenness: It would probably – by most moral standards – be morally bad to drink on an occasion where one parent has some particular obligation to care for the child. Or it would be bad to behave aggressively towards others, whether you were drunk or not. So, by conventional standards, the morally salient elements of the situation described by Metzinger are 1. on the *parents'* side, who are the *agents* in the situation, and they consist 2. in the *actions* these agents perform.

The specific question of the suggested consciousness ethics in the given case should instead be this: should it be legal that children *experience* their parents in a drunken state? This would amount to a massive change in ethical thinking, because it looks at the morally problematic situation from a completely new angle. The problem is, Metzinger here runs the risk of getting the ethical dilemma completely wrong. It would be like asking: should it be legal that a victim of rape *experiences* her rapist in a happy state of mind? The ethical question is not just whether one should *experience* the – good or bad – actions of someone else, but whether this someone should *perform* these actions or not. There should be – from a moral point of view – no rapists or aggressive, drunk parents whatsoever, whether they are consciously experienced or not. Hence the focus on the experience of a witness of other people's actions is at best secondary for ethical judgement.

4.1. Desirable states of consciousness

To substantiate his ethical re-orientation towards conscious experience, Metzinger provides three criteria for a »desirable state of consciousness« (ET, p. 233). According to him, these states

- minimise suffering
- possess an epistemic potential (expanding insight and expanding knowledge)
- have behavioural consequences that increase the occurrence of future valuable types of experience (ET, p. 233)

Here we can see how Metzinger gets lost in phenospace. The first condition is a purely phenomenal criterion, as is the third which combines a maximising utilitarian thought with the idea of minimal suffering as the best form of experience. But also the second condition – seemingly intended to provide some kind of connection to the ›outer world‹ – remains completely self-related and fully confined within phenospace: as we know by now, the ›bridge to the world‹ he is looking for has already been destroyed beforehand. On the one hand, Metzinger has raised doubt about the fact that we can ever make any reasonable claims about a ›real world‹ and a ›real self‹. Thus, any kind of ›true‹ insight and knowledge about the

world would be impossible. On the other hand, he has claimed that phenomenal states can intrinsically be connected with a *feeling of certainty*, so that any distinction between »real« insight and knowledge and its faked counterparts becomes indistinguishable within the realm of phenomenal experience. As Metzinger puts it »Self-deception may feel like insight« (ET, p. 220). But if we cannot tell self-deception from insight, the three criteria of valuable or desirable states of consciousness confine us to the domain of subjective experience. We are lost in phenospace.

With this, Metzinger faces the challenge of Robert Nozick's famous experience machine (Nozick 1974, 42-45). Nozick ponders the possibility of a machine that produces in us whatever desirable or pleasurable experience we could possibly wish for. Future neuropsychologists, so the imagined scenario goes, have found a way to stimulate a person's brain to induce all kinds of pleasurable experiences. Nozick then asks us the decisive question: given the choice, would we choose the experience machine over our real life?

While Nozick argues that we have several reasons to reject the temptation to hook up with the machine, Metzinger's answer remains ambiguous. On the one hand, he claims: »If it makes any sense at all to speak about the value of human existence, we must concede that it depends on more than the conscious experience of happiness.« (ET, p. 201). He suggests that engagement with truth and creative activity might be »at least as valuable as happiness« (ET, p. 201). On the other hand, his account of ›valuable conscious states‹ fails to provide any way of distinguishing between the value of states achieved in using such a machine and those brought about without it. His theory of valuable conscious states would speak in favour of using the machine – at least it evaluates both states equally – even if Metzinger seems somewhat reluctant to admit it.

In light of this it is not surprising that Metzinger reports in great detail the results of a study that shows how important and valuable an artificially induced ›spiritual experience‹ under the influence of psilocybin was to a group of subjects (ET, pp. 225-27). He speaks in favour of the controlled use of such drugs as ›phenotechnologies‹ and praises its enriching outcomes for human beings.

The question whether one wants to hook up with a phenotechnological machine or lead a real life is ultimately an existential choice to be made. One has to choose between isolating oneself into an – admittedly blissful – engagement with machines or just to live a life. The latter crucially consists in a real-world striving for ends that one deems worthy of pursuit. Knowledge to the effect that our lives have no real-world consequences – that our striving and experiencing is without effective contact with reality – would inevitably devalue

one's experiences, whatever their hedonic qualities might be. Can one really want to be constantly fooled about one's existential situation?

Here we see once again how Metzinger's consciousness ethics depends upon his theoretical position concerning the Ego-Tunnel: if what appears to be the world we live in is in fact no more than a world simulation within phenospace, then indeed it makes no difference whether we stick with our ›natural‹ world-simulations or whether we artificially produce different ones. If there is no real world in the first place, taking the world as we know it away is not such a big deal. Life, philosophy and ethics would all only be matters of conscious experiences – idle plays of projections in phenospace, or on Plato's cave walls. Seen in this light, Metzinger's tacit inclination towards choosing the experience machine over living a real life is understandable. But those who do not conceive of themselves as Ego Tunnels might prefer not to willingly enter the confines of an artificially enhanced phenospace brought about by altered brain states. Instead, they would be much better off continuing to engage as full organisms in exchange with the world and with others and keep living their lives in the world.

Consciousness ethics, in the end, would be no more than mundane reflections about the dangers and possibilities of new neuro-technologies, asking how these should be integrated in individual human lives, who should be responsible for administering them, and how abuses can be avoided. Of course, Metzinger has a point when he says that so far there has been too little reflection upon the values and dangers of certain extreme states of consciousness. But does this warrant a complete re-orientation of ethics? If indeed scientific advances make spectacular and wide-ranging alterations to human experience and human capacities possible, one might ultimately also reflect upon how these new capacities alter our conception of human beings and of worthwhile human lives. But in the end this is a task no different in nature from conventional ethics and philosophical reflection about the nature of man and the good human life. No revolution – in ethics or elsewhere – is called-for.

Bibliography

Abi-Rached, Joelle M.; N. Rose (2010): The Birth of the Neuromolecular Gaze. In: *History of the Human Sciences*, 23, 1 (2010). 1-26.

Campbell, Nancy (2010): Towards a Critical Neuroscience of ›Addiction‹. In: *BioSocieties*, 5, 1 (2010). 89-104.

- Choudhury, Suparna; S. K. Nagel; J. Slaby (2009): Critical Neuroscience: Linking Neuroscience and Society through Critical Practice. In: *BioSocieties* 4, 1 (2009). 61-77.
- Choudhury, Suparna (2009): Culturing the Adolescent Brain: What Can Neuroscience Learn from Anthropology? In: *Scan*, Epub ahead of print, doi: 10.1093/scan/nsp030.
- Choudhury, Suparna; J. Slaby (Eds.) (2011): *Critical Neuroscience. Handbook of the Social and Cultural Contexts of Neuroscience*. Chichester: Wiley-Blackwell.
- Churchland, Patricia Smith (1989): *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, MA.: The MIT Press 1989.
- Churchland, Patricia Smith (2002): *Brain-Wise: Studies in Neurophilosophy*. Cambridge, MA.: The MIT Press 2002.
- Clark, Andy (1997): *Being There. Putting Brain, Body and World Together Again*. Cambridge, MA.: MIT Press 1997.
- Clark, Andy (2008): *Supersizing the Mind. Embodiment, Action, and Cognitive Extension*. New York/Oxford: Oxford University Press 2008.
- Dumit, Joseph (2004): *Picturing Personhood. Brain Scans and Biomedical Identity*. Princeton: Princeton University Press 2004.
- Joyce, Kelly A. (2008): *Magnetic Appeal. MRI and the Myth of Transparency*. Ithaca: Cornell University Press 2008.
- Kitcher, Philip (2001): *Science, Truth, and Democracy*. New York/Oxford: Oxford University Press 2001.
- Latour, Bruno (2005): *Reassembling the Social. An Introduction to Actor-Network-Theory*. New York/Oxford: Oxford University Press 2005.
- Lenzen, Wolfgang (2006): Auf der Suche nach dem verlorenen »Selbst« – Thomas Metzinger und die »letzte Kränkung« der Menschheit. In: *Facta Philosophica*, 8 (2006). 161-192.
- Lenzen, Wolfgang (2005): Alles nur Illusionen? Philosophische (In-)Konsequenzen der Neurobiologie. In: *Facta Philosophica*, 7 (2005). 189-229.
- Lenzen, Wolfgang (2004): Damasio Theorie der Emotionen. In: *Facta Philosophica*, 6 (2004). 269-309.
- Metzinger, Thomas (2003): *Being No-One. The Self-Model Theory of Subjectivity*.

- Cambridge, MA.: The MIT Press 2003.
- Metzinger, Thomas (2009): *The Ego-Tunnel. The Science of the Mind and the Myth of the Self*. New York: Basic Books 2009.
- Noë, Alva (2009): *Out of our Heads. Why you are not your Brain and Other Lessons from the Biology of Consciousness*. New York: Hill and Wang 2009.
- Noë, Alva (2005): *Action in Perception*. Cambridge, MA.: The MIT Press 2005.
- Nozick, Robert (1974): *Anarchy, State, and Utopia*. New York: Basic Books 1974.
- Pickersgill, Martin (2009): *Between Soma and Society. Neuroscience and the Ontology of Psychiatry*. In: *BioSocieties*, 4 (2009). 45-60.
- Quine, Willard Van Orman (1953): *Two Dogmas of Empiricism*. In: Quine, W.V.O.: *From a Logical Point of View*. Cambridge, MA.: Harvard University Press 1953. S. 20-46.
- Revonsuo, Antti (2003): *The Contents of Phenomenal Consciousness*. In: *Psyche*, 9, 8 (2003).
- Rose, Nikolas (2006): *The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century*. Princeton: Princeton University Press 2006.
- Rose, Nikolas (1996): *Inventing Our Selves: Psychology, Power, and Personhood*. Cambridge: Cambridge University Press 1996.
- Rouse, Joseph (2002): *How Scientific Practices Matter. Reclaiming Philosophical Naturalism*. Chicago: Chicago University Press 2002.
- Slaby, Jan (2010): *Steps Towards a Critical Neuroscience*. In: *Phenomenology and the Cognitive Sciences*, 9 (2010). 397-416.
- Strawson, Peter F. (1966): *The Bounds of Sense. An Essay on Kant's Critique of Pure Reason*. London: Routledge 1996.
- Thompson, Evan (2007): *Mind in Life. Biology, Phenomenology, and the Sciences of the Mind*. Cambridge, MA.: Harvard University Press 2007.
- Varela, Francisco, E. Thompson, E. Rosch (Eds.) (1991): *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA.: The MIT Press 1991.
- Willaschek, Marcus (2003): *Der mentale Zugang zur Welt*. Frankfurt/M.: Vittorio Klostermann 2003.