The Return of Consciousness

A new science on old questions

★

Editors
Kurt Almqvist & Anders Haag

Axel and Margaret Ax:son Johnson Foundation

Axel and Margaret Ax:son Johnson Foundation

Stureplan 3
SE-103 75 Stockholm
SWEDEN

Tel: int+46 (0)8-788 50 00
Fax: int+46(0)8-788 50 60

Editing: Kurt Almqvist & Anders Haag
Cover image from *Utriusque cosmi maioris scilicet et minoris metaphysica, physica atqve technica historia: in duo volumina secundum cosmi differentiam diuisa*, 1617, by Robert Fludd
Graphic design and production: Johan Laserna
Typeface: Indigo
Printed and bound in Riga
by Studio RBB 2016

ISBN 978-91-89672-90-1

Table of Contents

THOMAS METZINGER

Suffering[1]

★ ★
★

The cognitive scotoma

In this essay, I will present a first, rudimentary, working concept of suffering and then derive six logical possibilities for its minimisation. Rigorous philosophical and scientific research programmes on consciousness, which have seen a great renaissance in the last three decades,[2] are reaching a level of maturity which suggests the careful introduction of additional relevance constraints: we can now begin to go beyond foundational research and ask what, given a wider and perhaps even normative context, the really *important* aspects or forms of conscious experience actually are.

This introduction sets one such possible context, by drawing attention to some interesting phenomenological characteristics of conscious suffering and to the equally interesting fact that suffering *as such* has been largely ignored by modern philosophy of mind and cognitive science. There clearly seems to exist a "cognitive scotoma", a systematic blind spot in our thinking about consciousness. While in medicine and psychiatry, for example, a large body of empirical data is already in existence, the search for a more abstract, general and unified theory of suffering clearly seems to be an unattractive epistemic goal for most researchers. Why is this so?

The "Eternal-Playlist" thought experiment

Consider a thought experiment. Let us assume there is life after death. This afterlife is temporally unbounded, it lasts eternally and within it conscious experience continues to exist. There is, however, an important difference with regard to the life you are living right now, as an active subject of experience: all conscious experiences after death are experiences you were permitted to choose from the set of subjective experiences you had in your current life – because after death there are no new experiences. Before death, by contrast, you lived through a large number of inner experiences and conscious states, and some of them were *actively created* by you: by going to the movies, by taking a hike, by reading books, by taking certain drugs or by participating in a meditation retreat. For most of our lives we are busy seeking, in one way or another, conscious states we will experience as pleasant or valuable. Of course, we also actively avoid or try to end unpleasant states, and often our search for pleasant states may simply be instrumental in pursuing this second goal.

Let us take it that the smallest unit of conscious experience is one single subjective moment – because, if we look carefully, we find that we are always living our life through conscious moments. In so doing, most of us are always looking for the "meaningful now", for those small "perfect" moments of satisfaction and happiness or even an enduring inner experience of wholeness and meaning. Often, but not always, the two subjective qualities of positive affect and of successfully "making sense" are deeply intertwined. Our life before death then is constituted by a finite chain of conscious moments, whereas the succession of consciously experienced psychological moments after the disappearance of our physical body is infinite – it will never end.

Our thought experiment now consists of an idea and a question. The idea is that you are allowed to select exactly which conscious moments from your finite life will be transferred to a "playlist for eternity": after your physical death all subjective experiences on this list will be replayed again and again, in random order. This process then creates your very own personal conscious eternity, and it is based on your very own selection of conscious experiences. During your lifetime you are like a phenomenological Cinderella attempting to pick out just the right ones: "The good into the pot, the bad into the crop!" And here is the question: if you were permitted to make this irrevocable selection all by yourself,

without asking the tame pigeons, the turtle-doves and all the birds beneath the sky for help and if it really was only yourself who could pick the good grains from the ashes of transience, into which the Evil Step-Mother has thrown them, which moments would you choose? Most importantly, how *many* moments, according to your own criteria, would you actually rank as truly worth living – in the sense of worth being *re*lived?

At Johannes Gutenberg University in Mainz, we began a first series of small pilot studies with a group of advanced philosophy students. We chose a signal-contingent, externally cued form of experience sampling.[3] One tech-savvy student programmed an SMS server in such a way that, for seven days, it sent ten signals a day at random points in time to the participants, whose cell phones would then briefly vibrate. The participants' task was to decide whether the *last* moment before the conscious experience of the vibration was a moment they would take with them into life after death. For many, the result was surprising: the number of positive conscious moments per week varied between 0 and 36, with an average of 11.8 or almost 31 per cent of the phenomenological samples, while at 69 per cent a little more than two thirds of the moments were spontaneously ranked as not worth reliving. If you are cued externally, it seems, less than a third of such experiential samples would have a chance of entering your very own "eternal playlist".

Clearly, one first has to distinguish conceptually the *subjective* and *objective* value of conscious moments. It would be conceivable that an objectively valuable, subjective conscious experience – for instance, a painful learning experience in the external world or a deep inner insight into a permanently recurring self-deception – would be subjectively perceived as unattractive and worthless. Conversely, there could be states that would be subjectively perceived as extremely meaningful, but which at the same time would appear as worthless from a critical, third-person perspective – for example, certain states induced by psychoactive substances or deeper delusional states caused by ideological indoctrination, religious belief systems and so on.

It may also be the case that, ultimately, it makes no sense to assign an "objective value" to a specific sample of conscious experience in the first place. But it is clear that the bulk of our phenomenal states instantiate a *subjective* quality of "valence", a sense or inner representation of value. Of course, subjective value is also determined by the policy we pursue –

by an expected utility in the more distant future. Nevertheless, it typically expresses itself on an emotional level, in the *hedonic* valence or affective quality of the present conscious moment.

Obviously, our "data" cannot count as significant, nor is a group of advanced philosophy students a representative sample. These were only exploratory pilot studies to give us initial ideas about feasibility and effect size. We were primarily interested in gaining a better understanding of the mechanism through which we subjectively experience conscious moments as pleasant or valuable. In so doing, we searched for a fine-grained form of assessment, as simple as possible, that would always register the current moment alone, as independently as possible from philosophical theories, ideologies and conceptual presumptions. For example, to test its influence, in a second study we dropped the afterlife assumption and the "eternity condition", replacing them with the following question: "Would you like to relive the very last conscious moment in *this* life?" Interestingly, under this condition only a little over 28 per cent of life moments were ranked as positive, while just below 72 per cent were considered not worth reliving. It seems like the original afterlife condition adds a small positive bias to one's phenomenological self-assessment.

Of course, there are a host of justified, and important, theoretical questions about this thought experiment. Could I also take only the one very best moment of my life and repeat this single moment indefinitely? *When* do I have to make the final decision? Many positive experiences include an aspect of "novelty" or "surprise", would this aspect be indefinitely present in the afterlife as well? What are the temporal boundaries of "a" conscious moment? How does one introspectively individuate such allegedly "single" moments, and how theory-contaminated are the corresponding verbal reports? How *reliable* is introspective knowledge in the first place? And should we trust our own intuitive value judgments; can they be epistemically justified? And so on.

The results were however striking: first, the initial surprise the low scores caused in all participants and, secondly, what could be interpreted as a consequent attempt to explain the phenomenon away. Indeed, "explaining away" a prediction error, in this case, may simply consist in an updating of one's conscious self-model, preferably in a way that allows one to reduce introspective uncertainty by dampening and suppressing future "unexpected news" of this type. Many participants immediately

started to develop more or less intricate "cognitive confabulations", trying to make their own self-assessment appear not so bad after all: "Happiness certainly is not the most important thing in life, I am writing a doctoral dissertation that will make a contribution to the knowledge of mankind and epistemic progress certainly adds more value and meaning to my life than momentary hedonic valences or even their life-time average!"; "Most of my moments are neither good nor bad anyway, they are just neutral!"; "Life-time average value is philosophically irrelevant, it is only the peak experiences and our memory of them which make a life a *good* life!"; "What *really* counts in life are the larger time-windows, not decontextualised phenomenal snapshots!" And so on.

Narrative self-deception

If, on the finest introspective level of phenomenological granularity that is functionally available to it, a self-conscious system would discover too many negatively valenced moments, then this discovery might paralyse it and prevent it from procreating. If the human organism would not repeat most individual conscious moments if it had any choice, then the logic of psychological evolution mandates concealment of the fact from the self-modelling system caught on the hedonic treadmill.[4] It would be an advantage if insights into the deep structure of its own mind, insights of the type just sketched, were not reflected in its conscious self-model too strongly, and if it suffered from a robust version of optimism bias. Perhaps it is exactly the main function of the human self-model's higher levels to drive the organism continuously forward, to generate a functionally adequate form of self-deception glossing over everyday life's ugly details by developing a grandiose and unrealistically optimistic inner story – a "narrative self-model" with which we can identify?

Any successful agent must be able to motivate itself. One solution to this problem of autonomous self-motivation could have consisted in what I would like to call "autobiographical Gestalt formation", an automatic escape into larger timescales. Maybe evolution has created self-models that automatically expand their predictive horizons as soon as the present is boring or simply too unpleasant?[5] Such a strategy of flexible, dynamic self-representation across a hierarchy of timescales could have a causal effect in continuously remotivating the self-conscious organism, systematically distracting it from the potential insight that the

life of an anti-entropic system is one big uphill battle, a strenuous affair with minimal prospect of enduring success. Let us call this speculative hypothesis "narrative self-deception". If something like this is true, one would also expect it to have an observable effect in academic philosophy and science as well.

Perhaps the foremost theoretical "blind spot" of current philosophy of mind is conscious suffering. Thousands of pages have been written about colour qualia and zombies, but almost no theoretical work is devoted to ubiquitous phenomenal states like boredom, the subclinical depression folk-psychologically known as "everyday sadness" or the suffering caused by physical pain.[6] Pain qualia are frequent examples, but suffering is rarely mentioned, because in philosophical debates pain qualia often are barely more than a stand-in, easily replaceable by other allegedly primitive forms of sensory consciousness. As Sascha Fink has pointed out, however, the sensation of pain and the emotional *affect* of unpleasantness are as distinct as hue and saturation in colour experience, and pain and suffering are clearly metaphysically independent mental phenomena, although they are similar in structure and formal object, and tied together by evolutionary ancestry.[7]

What exactly is it that suffering *represents*? We have no rigorous phenomenology of *Weltschmerz* (world-weariness), loss of one's confidence in humanity, loss of one's ethical integrity or the introspective profile following moral failure – or any more precise analysis of, say, the suffering that goes along with information overload, continuous partial attention and the consequent, frequently recurring loss of cognitive control and mental autonomy.[8] The same is true of panic, despair, shame, the suffering going along with some types of empathy, the phenomenology of losing one's dignity or the conscious experience of mortality. Why are these forms of conscious content generally ignored by the best of today's philosophers of mind? Is it simple careerism ("Nobody wants to read too much about suffering, no matter how insightful and important the arguments are!") or are there deeper, evolutionary reasons for what I have termed the "cognitive scotoma" at the outset, our blind spot in investigating conscious experience?

Take the experiential insight into one's own mortality. In its unprecedentedly high degree of explicitness it can probably count as a unique feature of human self-consciousness. It is interesting to note how a considerable part of current work in philosophy which also achieves

higher academic impact and actually gets attention from a wider audience is at the same time characterised by a vague potential for supporting mortality denial: just think of recent metaphysical discussions relating to property or even substance dualism, of dynamically extended minds "beyond skin and skull", panpsychism and quantum models of consciousness, or (to name an offence committed by the present author) philosophical attempts at supporting interdisciplinary research programmes into out-of-body experiences or virtual re-embodiment in robots and avatars.[9]

I think there may be a deeper problem here: we are systems that have been optimised to procreate as effectively as possible and to sustain their existence for millions of years. In this process, a large set of cognitive biases have been installed in our self-model. Our deepest cognitive bias is "existence bias", which means that we will simply do almost anything to prolong our own existence. Sustaining one's existence is the default goal in almost every case of uncertainty, even if it may violate rationality constraints, simply because it is a biological imperative that has been burned into our nervous systems over millennia. From a neuro-computational perspective, we appear as systems constantly and actively trying to maximise the evidence for their own existence, by minimising sensory surprisal (not to be confused with the personal-level state of being surprised).[10]

My point goes far beyond what is discussed as variants of "status-quo bias" in social or personality psychology.[11] Rather, it refers to the fact that we will almost always opt for continuing and protecting the life process as such, always preserving our own existence. But our brand-new cognitive self-model tells us that this organism's predictive horizon will inevitably shrink to zero, that our individual "maximum credible accident" is bound to happen. This creates a new situation in our inner environment to which we must adapt, a challenge creating a potentially permanent inner conflict and involving a specific, novel kind of suffering. This point therefore raises a deeper question, which we will encounter again when discussing the possibility of future suffering in self-conscious artificial systems: can there be conscious intelligence without existence bias?

Mortality-salient information and death-related cognition pose a constant threat to our self-esteem, because we need constantly to "buffer" the resulting existential anxiety – which in turn consumes an enormous

amount of resources. For example, to stabilise the layer of our self-model dynamically representing an overall emotional evaluation of our own worth, we now have to qualify for either literal or symbolic immortality by creating a new "meta-context". We need to invent some sort of ideology or other, and then successfully live up to the standards of value that are part of it. On the other hand, any uncertainty regarding the validity of one's worldview constantly undermines the efficacy of our attempt to protect the conscious self-model from being flooded by the terror going along with knowledge about the inevitability of death.[12] Creating too much awareness of mortality may therefore not only be bad for an academic career: it actually threatens the integrity of your and other people's conscious self-model – because it implies a deep existential loss of control. Almost nobody wants to gaze into an abyss for too long, because, as Nietzsche famously remarked, the abyss might eventually gaze back into you as well.[13]

Our constant inner battle for mortality denial is, however, only one introductory example. When one examines the ongoing phenomenology of biological systems on our planet, one finds that the varieties of conscious suffering are at least as dominant as, say, the phenomenology of colour vision or the capacity for conscious thought. The ability to see colour consciously appeared only very recently in evolutionary terms, and the ability consciously to think abstract thoughts, of a complex and ordered character, arose only with the advent of human beings. Pain, panic, jealousy, despair and the fear of dying, however, appeared millions of years earlier and in a much greater number of species.[14] Now, as the environment human beings have created for themselves gets increasingly complex, the sheer quantity of preferences and thereby of potentially frustrated preferences continues to rise.

Suffering

We lack a comprehensive theory of conscious suffering. One of the key desiderata is a conceptually convincing and empirically plausible model of this very specific class of phenomenal states: those that we do *not* want to experience if we have any choice, those states of consciousness which folk-psychology describes as "suffering". We need a general framework for philosophy of mind and empirical research, which allows of continuous refinement and updating. On an abstract conceptual level, I would

like to begin by proposing two phenomenological characteristics which may serve as markers of this class: *loss of control* and *disintegration of the self* (either on a "mental" or a "bodily" level of representational content). This makes sense, because, first, the phenomenal self-model (PSM)[15] is exactly an instrument for global self-control and, secondly, it constantly signals the current status of organismic integrity to the organism itself. If the self-model unexpectedly disintegrates, this typically is a sign that the biological organism itself is in great danger of losing its coherence as well.

In addition, many forms of suffering can be described as a loss of autonomy: bodily diseases and impairments typically result in a reduced potential for global self-control on the level of bodily action; experienced pain can be described as a shrinking of the space of attentional agency accompanied by loss of attentional self-control, because functionally it tends to fixate attention on the painful, negatively valenced bodily state itself. And there are many instances where psychological suffering is expressed as a loss of cognitive control, for example in depressive rumination, neurotic threat sensitivity and mind wandering;[16] similarly, in insomnia people are plagued by intrusive thoughts, feelings of regret, shame and guilt while suffering from dysfunctional forms of cognitive control.[17]

Obviously, an empirically informed, philosophically coherent theory of suffering would be of high relevance for applied ethics, policymaking and legal theory. I am not going to present such a theory here, but in what follows I will sketch four necessary conditions for the concept of "suffering" while making no claims about sufficiency or offering a technical definition. The hope is that for practical purposes these short remarks constitute a good starting point, perhaps already a "minimal model of conscious suffering", but at least a working concept that we can use and gradually refine.

The C-condition: conscious experience

"Suffering" is a *phenomenological* concept. Only beings with conscious experience can suffer. Zombies, human beings in dreamless deep sleep, as those in deep coma or under anaesthesia, do not suffer, just as possible persons or unborn human beings who have not yet come into existence are unable to suffer. Robots or other artificial beings can only suffer if they are capable of having phenomenal states. We do not yet have a theory

of consciousness.[18] But we already know enough to come to an astonishingly large number of practical conclusions in animal and machine ethics.[19]

The PSM-condition: possession of a phenomenal self-model

The most important phenomenological characteristic of suffering is the "sense of ownership", the non-transcendable subjective experience that it is *myself* who is suffering right now, that it is my *own* suffering I am undergoing. The first condition is not sufficient, since the system must be able to attribute suffering to itself. Suffering presupposes self-consciousness. We thus need to add the condition of having a conscious self-model: only those conscious systems which possess a PSM are able to suffer, because only they, through a process of functionally and representationally integrating certain negative states into their PSM, can *appropriate* the representational content of certain inner states on the level of phenomenology. Only systems with a PSM can generate the phenomenal quality of ownership, and this quality is another necessary condition for phenomenal suffering to appear.

Conceptually, the essence of suffering lies in the fact that a conscious system is forced to *identify* with a state of negative valence and is unable to break this identification or to detach itself functionally from the representational content in question (the fourth condition is of central relevance here). Of course, suffering has many different layers and phenomenological aspects. But it is the *phenomenology of identification* which is central for theoretical, as well as ethical and legal, contexts.[20] What the system wants to end is experienced as a state of *itself*, a state the content of which cannot be successfully integrated and which limits its autonomy because it cannot effectively distance itself from it. What it cannot distance itself from is an internal representation of loss of control and functional coherence, a situation of rising uncertainty.

If one understands this point, one also sees why the "invention" of conscious suffering by the process of biological evolution on this planet was so extremely efficient. It supports the active minimisation of uncertainty by elevating it to the functional level of global availability and simultaneously tying it to an individual first-person perspective.[21] Suffering is a new causal force, because it motivates organisms and continuously drives them forward. At the same time, and again just metaphorically,

had the inventor of conscious suffering been a person (like the "Evil Step-Mother" of our introductory thought experiment), we could describe the overall process as extremely cruel. Above a certain level of complexity, evolution continuously instantiates an enormous number of frustrated preferences; it has brought an expanding and continuously deepening ocean of consciously experienced suffering into a region of the physical universe where nothing comparable existed before.

Clearly, the phenomenology of ownership is not sufficient for suffering. We can all easily conceive of self-conscious beings who do not suffer. If we however accept an obligation towards minimising risks in situations of epistemic indeterminacy, and if we accept traditional ethical principles or legal duties demanding that we always "err on the side of caution", then this second condition is of maximal relevance: we should treat every representational system that is able to activate a PSM, however rudimentary, as a moral object, because it can *in principle* own its suffering on the level of subjective experience. The disposition, the relevant functional potential, has already been created: it is precisely the phenomenal property of "mineness", the consciously experienced, nonconceptual sense of ownership, which counts for ethical purposes. Without phenomenal ownership, suffering is not possible. With ownership, the capacity for conscious suffering can begin to evolve, because the central, necessary, functional condition for an *acquisition* of negative phenomenology is now given.

The NV-condition: negative valence

Suffering is created by states representing a *negative value* being integrated into the PSM of a given system. Through this step, thwarted preferences become thwarted *subjective* preferences, the conscious representation that one's *own* preferences have been frustrated (or will be frustrated in the future). This does not mean that the system itself must have a full understanding of what these preferences really are (in terms of cognitive, conceptual or linguistic competences): it suffices that it does not want to undergo *this present conscious experience*, that it wants it to end. Of course, to create the aforementioned phenomenal urgency of change, the mere representation of an *expected* negative utility may suffice. For the experiential quality it is not only the content, however, but also the format, the inner mode of presentation, which counts.

The phenomenology of suffering has many different facets. To give another example, artificial suffering in conscious machines could be very different from human suffering. It is also conceivable that, say, some kind of Bostromian "superintelligence"[22] could represent negative expected utilities and frustrated preferences in inner forms of phenomenality that involve no conscious suffering at all. There could be perfectly rational artificial agents, exhibiting neither the biologically grounded "existence bias" nor any other of the human cognitive biases which result from the millions of years in which evolution has shaped the self-models of our ancestors. But *if* they suffered, damage to their physical hardware could be represented in internal data formats completely alien to human brains – for example, generating a subjectively experienced, qualitative profile for embodied pain states which biological systems like ours could not emulate or even vaguely imagine. The phenomenal character going along with high-level cognition might transcend human capacities for empathy and understanding, such as through intellectual insight into the frustration of one's own preferences or into disrespect of one's creators – perhaps even into the absurdity of one's own existence as a self-conscious machine, a mere commodity or research tool used by an ethically inferior biosystem.

The T-condition: transparency

"Transparency" is not only a visual metaphor but also a technical concept in philosophy, which comes in a number of different uses and flavours. Here, I am exclusively concerned with "phenomenal transparency", a property which some conscious but no unconscious states possess.[23] Transparent phenomenal states make their representational content appear as irrevocably *real*, as something the existence of which you cannot doubt. Put more precisely, you may certainly be able cognitively to have doubts about its existence, but according to subjective experience this phenomenal content – the *awfulness* of pain, the fact that it is *your own* pain – is not something you can distance yourself from. The phenomenology of transparency is the phenomenology of direct realism and in the domain of self-representation it creates the phenomenology of identification discussed vis-à-vis the second condition.

Phenomenal transparency means that something particular is not accessible to subjective experience, namely the *representational character* of

the contents of conscious experience. This refers to all sensory modalities and to our integrated phenomenal model of the world as a whole – but also to large parts of our self-model. The instruments of representation themselves cannot be represented as such anymore, and hence the system making the experience, by conceptual necessity, is entangled in an illusion of epistemic immediacy, a naive form of realism. This happens, because, necessarily, it now has to experience itself as being in direct contact with the current contents of its own consciousness. What precisely is it that the system cannot experience? What is inaccessible to conscious experience is the simple fact of this experience taking place in a *medium*. Therefore, transparency of phenomenal content leads to a further characteristic of conscious experience, namely the subjective impression of immediacy.

Obviously, this functional property is not bound to biological nervous systems; it could be realised in advanced robots or conscious machines as well. In particular, it has nothing to do with holding a certain "belief" or adhering to a specific philosophical position; it is plausible to assume that many more simple animals on our planet, which are conscious but not able to speak or to entertain high-level, symbolic thoughts, have transparent phenomenal states – just as the first, simple, post-biotic subjects of experience in the future might have.

This may also provide us with a deeper understanding of what the process of conscious experience actually *is*. To be conscious means to operate under a unified mental ontology, constituted by an integrated set of assumptions about what kind of entities *really exist*. Systems operating under a single transparent world-model for the first time live in a reality which, for them, cannot be transcended. On a functional level they become *realists*. Again, this does not mean that they have to possess or even be able to form certain beliefs, or use explicit symbol structures in communication. It only means that the implicit assumption of the actual presence of a world becomes causally effective, because, as philosophers might say, non-intentional properties of their own internal representations are not introspectively accessible to them – they necessarily experience themselves as being in direct contact with their content. This is also true of the conscious self-model. A transparent self-model adds a new metaphysical primitive, a new kind of entity to the system's ontology: the "self".[24]

Of course, all four conditions specified here are necessary, but to

understand the very specific phenomenology expressed by self-reports such as "I am certain that I do exist and I am identical with *this*!", the conjunction of the PSM-condition and the T-condition is central. The transparent world-model allows a system to treat information as *factual* information, as irrevocably stemming from the real world; a transparent self-model creates the Cartesian phenomenology of being certain of one's own existence. Any robot operating under a phenomenally transparent body-model would, phenomenologically, *identify* with the content of this model and hence with any negatively valenced state that might become integrated into it.

At this stage, our working concept of suffering is constituted by four necessary building-blocks: the C-condition, the PSM-condition, the NV-condition, and the T-condition. Given our current situation of epistemic indeterminacy, any system that satisfies all of these conceptual constraints should be treated as an object of ethical consideration, simply because we do not know if, taken together, they might already constitute the necessary *and sufficient* set of conditions. By definition, any system, whether biological, artificial or post-biotic, not fulfilling at least one of these necessary conditions is not able to suffer. Here the central desideratum for future research is to develop this first working concept into a more comprehensive, empirically testable *theory* of suffering.

To be useful for human and animal ethics, for robot ethics and robot law, such a theory would still have to possess the necessary degree of abstraction. We want it to yield *hardware-independent demarcation criteria*. Which, if any, aspects of conscious suffering are multirealisable and which tied to a specific form of embodiment? Ideally such criteria would allow us to ignore all the concrete implementational details contingently characterising the relevant class of biological organisms on our planet, because we need to decide if a given artificial system is currently suffering, if it has the capacity to suffer, or if this type of system will likely evolve the capacity to suffer in the future.[25]

On our way towards a more universal theory of suffering, the second central problem is the "metric problem": if, say, for the purposes of an evidence-based, rational approach to applied ethics, we want to develop an empirically grounded, *quantifiable* theory of suffering, then we need to know what the phenomenal primitives in the relevant domain actually are. We have to determine the smallest units of conscious suffering. What exactly is the phenomenological *level of grain* that possesses explanatory

relevance (from a scientific point of view) and what level of granularity has maximal practical relevance (eg, from the perspective of applied ethics)? If we hold on to the background assumption made in the beginning, postulating that the smallest unit of conscious experience is a single "experiential moment", then we arrive at a positive conclusion: the smallest unit of conscious suffering is a "phenomenally transparent, negatively valenced self-model moment". Arguably, such negative self-model moments (NSMs) are the phenomenal primitives constituting every episode of suffering, and the frequency of their occurrence is the empirically detectable quantity that we want to minimise.

Six logical possibilities to minimise suffering

For reasons of space, I will present an unargued background assumption: it is much more important to reduce suffering than to maximise happiness, because frustrated preferences are ethically more relevant than satisfied preferences. "Negative utilitarianism" (NU)[26] says that we should concentrate on minimising suffering, because of a deep phenomenological asymmetry between positively and negatively valenced states (see Option 4 below). Should we, as I propose, decide to make a fresh start and turn the topic of conscious suffering into a target of modern philosophy of mind and cognitive science, we would very likely corroborate and confirm the asymmetry between happiness and suffering and be able to describe it in much greater detail. If so, this would be one reason to choose NU as a meta-ethical position.

Given the first working concept of suffering just sketched, we can already describe a number of possibilities to reduce the occurrence of our target phenomenon. Obviously, the richer and the more precise a phenomenological concept gets, the more possibilities and potential causal routes for changing the actual class of conscious states can be envisioned. If the concept in addition becomes empirically grounded (eg, by isolating its minimally sufficient correlates and analytically describing their common computational function) or if we increase its domain specificity (eg, by just looking into the suffering of pigs, cows or human beings), then technological interventions gradually become more feasible. And if, conceptually, suffering necessarily involves the phenomenal representation of a *gradient* (ie, of temporal properties like duration, change and succession), then creating a form of conscious experience lacking these

features would result in an absence of subjective suffering. But for now, accepting the four necessary conditions introduced above, let us set out the options arising.

Option 1: ending existence

The first option, quite obviously, would consist in painlessly and unexpectedly killing all sentient beings. No new "negative self-model moments" would be created in the process; no NSMs would be instantiated afterwards. If, however, the negative utilitarian is committed to effectively minimising suffering, and if she targets a specific population of experiential subjects, she is dependent on consensus within, and cooperation from, this population. She might then look for the optimal degree of adapting to the specific cognitive biases characterising her target population. For example, a more moderate version could say that we should respect the deepest cognitive bias of all currently known sentient beings – namely, existence bias – plus the fact that their transparent self-models not only express this bias on the level of inner experience but also force them consciously to experience themselves as indivisible wholes, as irreducibly individual entities. Such a moderate version might therefore respect an individual right to existence for all self-conscious biological systems already born, but prevent *future* sentient beings from coming into existence. Logical options like these have of course been explored in centuries of philosophical thinking, and they are related to what today is frequently referred to as "anti-natalism".[27]

It is interesting to see how, for many of us, intuitions diverge for biological and artificial systems. Imagine, as an elected member of a future Ethics Committee for Synthetic Phenomenology, it was your task to define and functionally constrain the "playlist" for the very first population of conscious machines. Your job would be to determine what kinds of phenomenal "self-model moments" would be allowed to evolve as such machines began to interact with the world and each other, and perhaps even to multiply themselves. You also know that, functionally speaking, suffering is necessary for autonomous self-motivation and the emergence of truly intelligent behaviour. Therefore, those conscious machines inevitably would have to instantiate a certain number of negatively valenced self-model moments in the enormous cascade of conscious experiences you are just about to trigger by setting off the evolution of artificial conscious subjects. Here, as opposed to the case of

biosystems like ourselves, a larger number of people would not want to unnecessarily *initiate* new chains of NSMs and therefore feel drawn to an ethical position one could term "anti-natalism for self-conscious machines".

Let me point to a concrete, pressing problem we face today. It is a problem within the newly emerging field of "robot ethics", the problem of successfully implementing moral cognition and ethical behaviour in artificial agents. As embodied artificial intelligence begins to operate in open environments populated by human beings, as robotic systems become ever faster and more autonomous, they will increasingly be confronted with situations in which it would be inefficient and irrational, perhaps even unethical, to waste precious temporal resources by waiting for a final decision from some human agent.[28] Many of these situations will be of a kind in which the consequences of the artificial agents' actions will directly affect the wellbeing of other sentient beings. Just think of three autonomous cars in a roundabout realising that they will soon be involved in a complex collision with a deer and two human-driven cars, having to decide (and to negotiate with each other) how to minimise intelligently the overall damage.

The best way to solve this problem of "synthetic morality" is by endowing machines with a formalised axiology and a computable value calculus. But then new problems appear: should the autonomy of such ethical robots be limited in a way that they could in principle never find a way to "flash" or in any other way alter their own "normative firmware"? This would preclude them from learning and becoming better in the domain of moral intelligence. And how would we prevent such systems from arriving at repugnant conclusions? Let us assume the robot's value set would be based on a negative utilitarian axiology.[29] "To always minimise involuntary suffering and the overall quantity of NSMs in all sentient beings" could therefore be its highest priority. It would probably be in our own interest to prevent such artificial moral agents from ever arriving at the general conclusion that option one is the best way to achieve this goal. Call this the "problem of the benevolent anti-natalist robot": almost nobody likes the idea of future artificial intelligences silently preparing for the extermination of all human beings and other sentient creatures on the planet – on purely ethical grounds, to be sure. This just illustrates some of the problems with option one. But it also makes it clear that, until such meta-ethical problems are solved, we

should strive to always err on the side of caution. In practice, we should take no unnecessary risks.

Option 2: Eliminating the C-Condition

Option 2 would consist in eliminating not the physical existence of conscious systems but all *phenomenal properties* in the universe. Ending all kinds of conscious experience would mean that the C-condition was not fulfilled. Trivially, as "suffering" is a phenomenological concept, there would be no suffering in a zombie world.

Option 3: Eliminating the PSM-Condition

A third option to minimise suffering would consist in eliminating not conscious experience as such but only *self*-consciousness. Here, the PSM-condition would not be fulfilled, resulting in the disappearance of all forms of phenomenal selfhood. Conscious states could still exist, but certain complex phenomenal properties would not be instantiated any more, most notably the qualities of selfhood, ownership and agency. Option 3 has been investigated in depth, for example by 25 centuries of Buddhist philosophy.[30]

Option 4: Eliminating the NV-Condition

A fourth option, equally looking back to centuries of philosophical discussion, would be to eliminate not self-consciousness but all *subjective preferences*. If one were to end all self-consciously experienced preferences, then no such preferences could ever be frustrated, because the NV-condition was not fulfilled. A being without preferences would not be selective, not even about the quality of its own mental states, so one might describe the resulting phenomenological configuration as a form of "choiceless awareness".

In accordance with the NU-assumption, I presuppose that satisfying preferences is ultimately not a valid option, because of impermanence and a deep phenomenological asymmetry between positive self-model moments and negative ones (NSMs). First, physical embodiment, impermanence and transience prevent any more permanent satisfaction of preferences (or a stable state in the self-model). In addition, the phe-

nomenology of suffering is not a simple mirror-image of happiness, mainly because it involves a much higher urgency of change. In most forms of happiness this centrally relevant subjective quality which I have termed the "urgency of change" is absent, because they do not include any strong preference for being even *more* happy. In fact, a lot of what we describe as "happiness" may turn out to be a relief from the urgency of change. The subjective sense of urgency, in combination with the phenomenal quality of losing control and coherence of the phenomenal self, is what makes conscious suffering a very distinct class of states, not just the negative version of happiness. This subjective quality of urgency is also reflected in our widespread moral intuition that, in an ethical sense, it is much more urgent to help a suffering person than to make a happy person even happier.

One can also state and dissolve another frequent misunderstanding. Human suffering is rarely dramatic suffering. Almost all negatively valenced states involve only a mild emotional sense of preference frustration – perhaps some weak impairment of bodily wellbeing or a diffuse background feeling of boredom, possibly accompanied by an unspecific, generalised worry about the future plus a subtle phenomenology of uncertainty. In addition, as it is plausible to assume that these frequent and much more subtly negative states are forming the majority of our conscious self-model moments, most of us may have long ago begun to perceive them as inescapable and uncontrollable. We may therefore operate under a "domain-general" version of learned helplessness with regard to our own suffering: we become unable or unwilling to avoid subsequent encounters with such inner situations, because on a deep functional level we already believe that we cannot effectively control the total probability of their occurrence. Consequently, we do not take action to avoid more subtle forms of negative everyday phenomenology.

Therefore, what we prematurely report as "neutral" states may often actually be the inner experience of subtle preference frustration plus learned helplessness. We report "neutral", but what we actually mean is "default". If one introspects carefully, truly neutral moments are something very rare, because some sort of affective valence accompanies almost all of our conscious moments. Zero pleasant intensity plus zero suffering describes a rare situation, but under NU it is a perfect and desirable state of conscious experience. In the history of philosophy, and in a large number of theoretical variations, this way of eliminating the

NV-condition has long been thought about. The ultimate goal was to attain a lucid and robust state of tranquillity, a state of equanimity not disturbed by the passions, as exemplified in classical philosophical notions like *ataraxia* or *upekkha*.

Option 5: Eliminating the T-Condition

Fifthly, it is conceivable that one might not eliminate self-consciousness *per se* but selectively target only the *phenomenology of identification* mentioned above. One would then only permit the appearance of self-models that are opaque and therefore *not* units of identification (UI).[31] There would be an organism-model, but not a self-model. Conscious preferences like desires, wishes or cravings might still arise and become integrated into this mere organism-model, but under this option no functional identification would take place, because the T-condition was not fulfilled. It is an empirical prediction of the self-model theory of subjectivity[32] that the property of "selfhood" would disappear as soon as all of the human self-model became phenomenally opaque, by making earlier processing stages available to introspective attention and thereby reflecting its representational nature as an internal construct on the level of its content. Frustrated preferences could still be consciously represented in such a model. But the organism would not experience them as part of the self – this entity would have disappeared from its subjective ontology.

Option 6: Maximizing the UI

A last logical possibility could be to maximise the unit of identification, thereby dissolving the phenomenally experienced first-person perspective and the underlying subject–object structure of phenomenal experience. Here, the idea is that there must always be one most general phenomenal property and that it is possible to conceive of a phenomenal configuration in which the mechanism of identification is tied to this property, not to the experience of selfhood. It is difficult to find a label for this most global and abstract form of phenomenal character but there are traditional candidates. For example, we might call the most general phenomenal property instantiated by the process of conscious experience "the unity of consciousness" or "the wholeness of the moment" or, perhaps best, we could simply speak of "phenomenality per se".

Option 6 then describes the possibility to keep even the phenomenology of identification but to tie it to phenomenality per se, by turning the process of conscious experience *itself* into the unit of identification. In this phenomenal configuration, which is clearly possible from a logical point of view, preferences might still arise and be frustrated but they would not be subjective ones, because the underlying subject–object structure of consciousness had been dissolved. In other words, the individual, first-person perspective would have disappeared, because its origin (the unit of identification) was now maximised. Phenomenologically, such an aperspectival form of consciousness would make suffering impossible, because it was not a subjective form of experience any more. One interesting, and remaining, question would be if for this class of states we would still want to say that the PSM-condition is fulfilled or not, if option six describes a form of conscious self-representation.

To conclude, I have argued that scientific research on consciousness has reached a stage of maturity in which we may slowly and carefully begin to depart from pure foundational research by imposing additional relevance constraints. I propose that to understand the deep structure of suffering on a more abstract level is one such highly relevant goal for future research. Going beyond the mere search for neural correlates in biological systems, it might involve a strategy of mathematical and computational modelling of conscious suffering, in turn leading to the formulation of new, testable hypotheses. To facilitate this process, I have offered a first set of ideas and conceptual instruments.

Notes

1. I want to thank Regina Fabry, Sascha Fink, Adriano Mannino, Iuliia Pliushch, Lisa Quadt, Wanja Wiese and Jennifer Windt for their comments on earlier versions. I am also greatly indebted to Robin Wilson for excellent and substantial editorial help with the English version of this essay.

2. T. Metzinger (ed.), *Conscious Experience*, Thorverton, Imprint Academic, 1995; T. Metzinger (ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, Cambridge, MA, MIT Press, 2000.

3. R.T. Hurlburt, *Investigating Pristine Inner Experience: Moments of Truth*, Cambridge, Cambridge University Press, 2011; R.T. Hurlburt & C.L. Heavey, 'Investigating Pristine Inner Experience: Implications for Experience Sampling and Questionnaires', *Consciousness and Cognition*, no. 31, 2015, pp. 148–59.

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

T. Metzinger, *The Ego Tunnel*, New York, NY, Basic Books, 2009.

5. J. Hohwy, *The Predictive Mind*, Oxford, Oxford University Press, 2013; A. Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York, Oxford University Press, 2016; T. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing*, Frankfurt am Main, MIND Group, 2017; T. Metzinger & J. M. Windt (eds.), *Open MIND*, Frankfurt am Main: MIND Group, 2015; K. Friston, 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, vol. 11, no. 2, 2010, pp. 127–38; T. Metzinger, 'The Myth of Cog-nitive Agency: Subpersonal Thinking as a Cyclically Recurring Loss of Mental Autonomy', *Frontiers in Psychology*, no. 4, 2013, p. 931; T. Metzinger, 'M-Autonomy', *Journal of Consciousness Studies*, vol. 22, nos. 11–12, 2015, pp. 270–302.

6. See N. Grahek, *Feeling Pain and Being in Pain*, 2nd edn., Cambridge, MA, Bradford Books, MIT Press, 2007, for a notable exception.

7. S.B. Fink, 'Independence and Connections of Pain and Suffering', *Journal of Consciousness Studies*, vol. 18, nos. 9–10, 2011, p. 62.

8. M.A. Killingsworth & D.T. Gilbert, 'A Wandering Mind is an Unhappy Mind', *Science*, vol. 330, no. 6006, 2010, p. 932; A.M. Perkins et al., 'Thinking Too Much: Self-Generated Thought as the Engine of Neuroticism', *Trends in Cognitive Sciences*, vol. 19, no. 9, 2015, pp. 492–98; Metzinger, 'The Myth of Cognitive Agency'; Metzinger, 'M-Autonomy'.

9. B. Lenggenhager et al., 'Video Ergo Sum: Manipulating Bodily Self-consciousness', *Science*, vol. 317, no. 5841, 2007, pp. 1096–99; Metzinger, 'Empirical Perspectives from the Self-Model Theory of Subjectivity'; T. Metzinger, 'Why are Out-of-Body Experiences Interesting for Philosophers? The Theoretical Relevance of OBE Research', *Cortex*, vol. 45, no. 2, 2009, pp. 256–58; O. Blanke & T. Metzinger, 'Full-Body Illusions and Minimal Phenomenal Selfhood', *Trends in Cognitive Sciences*, vol. 13, no. 1, 2009, pp. 7–13.

10. K. Friston, 'The Free-Energy Principle', p. 128; J. Hohwy, 'The Self-Evidencing Brain', *Noûs*, 2014.

11. S. Eidelman, C.S. Crandall & J. Pattershall, 'The Existence Bias', *Journal of Personality and Social Psychology*, vol. 97, no. 5, 2009, pp. 765–75; S. Eidelman & C.S. Crandall, 'Bias in Favor of the Status Quo', *Social and Personality Psychology Compass*, vol. 6, no. 3, 2012, pp. 270–81.

12. See T. Pyszczynski, S. Solomon & J. Greenberg, 'Thirty Years of Terror Management Theory', *Advances in Experimental Social Psychology*, no. 52, 2015, pp. 1–70, for a recent review.

13. F. Nietzsche, *Beyond Good and Evil*, ed. R.-P. Horstmann & J. Norman, Cambridge, Cambridge University Press, 2002, aphorism 146, p. 69.

14. Quantitatively speaking, and under practically every conceivable metric, wild animal suffering exceeds human suffering and the suffering inflicted by humans on other animals by factory farming and so on by many orders of magnitude.

See, eg, B. Tomasik, 'The Importance of Wild Animal Suffering', *Relations: Beyond Anthropocentrism*, vol. 3, no. 2, 2015, pp. 133–52.

15. Metzinger, *Being No One*; T. Metzinger, 'Précis: Being No One', *Psyche*, vol. 11, no. 5, 2006, pp. 1–35; Metzinger, 'Empirical Perspectives from the Self-Model Theory of Subjectivity'.

16. Perkins et al., 'Thinking Too Much'; J. Smallwood & J.W. Schooler, 'The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness', *Annual Review of Psychology*, no. 66, 2015, pp. 487–518; Metzinger, 'The Myth of Cognitive Agency'; Metzinger, 'M-Autonomy'.

17. R.E. Schmidt & M. van der Linden, 'The Aftermath of Rash Action: Sleep-Interfering Counterfactual Thoughts and Emotions', *Emotion*, vol. 9, no. 4, 2009, pp. 549–53; P. Gay, R.E. Schmidt & M. van der Linden, 'Impulsivity and Intrusive Thoughts: Related Manifestations of Self-Control Difficulties?', *Cognitive Therapy and Research*, vol. 35, no. 4, 2011, pp. 293–303; R.E. Schmidt, A.G. Harvey & M. van der Linden, 'Cognitive and Affective Control in Insomnia', *Frontiers in Psychology*, no. 2, 2011, 349.

18. A. Seth, 'Models of Consciousness', *Scholarpedia*, vol. 2, no. 1, 2007, p. 1328; Metzinger (ed.), *Conscious Experience*; Metzinger (ed.), *Neural Correlates of Consciousness*.

19. A.K. Seth, B.J. Baars & D.B. Edelman, 'Criteria for Consciousness in Humans and Other Mammals', *Consciousness and Cognition*, vol. 14, no. 1, 2005, pp. 119–39; P. Low et al., 'The Cambridge Declaration on Consciousness', 2012, http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf (accessed January 6, 2016); T. Metzinger, 'Two Principles for Robot Ethics', in E. Hilgendorf & J.-P. Günther (eds.), *Robotik und Gesetzgebung*, Baden-Baden, Nomos, 2013, pp. 263–302.

20. Metzinger, 'Two Principles for Robot Ethics'.

21. Metzinger, Being No One.

22. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

23. See T. Metzinger, 'Phenomenal Transparency and Cognitive Self-Reference', *Phenomenology and the Cognitive Sciences*, vol. 2, no. 4, 2003, pp. 353–93, for a concise introduction.

24. See T. Metzinger, 'The No-Self Alternative', in S. Gallagher (ed.), *The Oxford Handbook of the Self*, Oxford, Oxford University Press, 2011, pp. 279–96.

25. See the thought experiment in Metzinger, *The Ego Tunnel*, p. 194.

26. See F. Fricke, 'Verschiedene Versionen des Negativen Utilitarismus', *Kriterion*, no. 15, 2002, pp. 13–27, for introductory references. An interesting discussion is B. Contestabile, 'Negative Utilitarianism and Buddhist Intuition', *Contemporary Buddhism*, vol. 15, no. 2, 2014, pp. 298–311.

27. D. Benatar, *Better Never to Have Been: The Harm of Coming into Existence*, Oxford & New York, Clarendon Press, Oxford University Press, 2006.

28. A. Mannino et al., 'Künstliche Intelligenz: Chancen und Risiken', *Diskussionspapiere der Stiftung für Effektiven Altruismus*, no. 2, 2015, pp. 1–17.

29. Fricke, 'Verschiedene Versionen des Negativen Utilitarismus'; B. Contestabile, 'Negative Utilitarianism and Buddhist Intuition'.

30. M. Siderits, *Buddhism as Philosophy: An Introduction*, Indianapolis, IN, Hackett Publishing, 2007; M. Siderits, 'Buddhist Non-Self: The No-Owner's Manual', in S. Gallagher (ed.), *The Oxford Handbook of the Self*, Oxford, Oxford University Press, 2011, pp. 296–315.

31. Metzinger, 'The Myth of Cognitive Agency'; T. Metzinger, 'Why are Dreams Interesting for Philosophers? The Example of Minimal Phenomenal Selfhood, plus an Agenda for Future Research', *Frontiers in Psychology*, no. 4, 2013, p. 746.

32. Metzinger, *Being No One*; Metzinger, 'Phenomenal Transparency and Cognitive Self-Reference'; Metzinger, 'Précis'; Metzinger, 'Empirical Perspectives from the Self-Model Theory of Subjectivity'.

References

Benatar, D., *Better Never to Have Been*: *The Harm of Coming into Existence*, Oxford & New York, Clarendon Press, Oxford University Press, 2006.

Blanke, O. & T. Metzinger, 'Full-Body Illusions and Minimal Phenomenal Selfhood', *Trends in Cognitive Sciences*, vol. 13, no. 1, 2009, pp. 7–13.

Bostrom, N., *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

Chalmers, D.J., 'What is a Neural Correlate of Consciousness?', in T. Metzinger (ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, Cambridge, MA, MIT Press, 2000, pp. 17–39.

Clark, A., *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York, Oxford University Press, 2016.

Contestabile, B., 'Negative Utilitarianism and Buddhist Intuition', *Contemporary Buddhism*, vol. 15, no. 2, 2014, pp. 298–311.

Eidelman, S. & C.S. Crandall, 'Bias in Favor of the Status Quo', *Social and Personality Psychology Compass*, vol. 6, no. 3, 2012, pp. 270–81.

Eidelman, S., C.S. Crandall & J. Pattershall, 'The Existence Bias', *Journal of Personality and Social Psychology*, vol. 97, no. 5, 2009, pp. 765–75.

Fink, S.B., 'Independence and Connections of Pain and Suffering', *Journal of Consciousness Studies*, vol. 18, nos. 9–10, 2011, pp. 46–66.

Fricke, F., 'Verschiedene Versionen des Negativen Utilitarismus', *Kriterion*, no. 15, 2002, pp. 13–27.

Friston, K., 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, vol. 11, no. 2, 2010, pp. 127–38.

Gay, P., R.E. Schmidt & M. van der Linden, 'Impulsivity and Intrusive Thoughts: Related Manifestations of Self-Control Difficulties?', *Cognitive Therapy and Research*, vol. 35, no. 4, 2011, pp. 293–303.

Grahek, N., *Feeling Pain and Being in Pain*, 2nd edn., Cambridge, MA, Bradford Books, MIT Press, 2007.

Hilgendorf, E. & J.-P. Günther (eds.), *Robotik und Gesetzgebung*, Baden-Baden, Nomos, 2013.

von Hippel, W. & R. Trivers, 'The Evolution and Psychology of Self-Deception', *Behavioral and Brain Sciences*, vol. 34, no. 1, 2011, pp. 1–56.

Hohwy, J., *The Predictive Mind*, Oxford, Oxford University Press, 2013.

Hohwy, J., 'The Self-Evidencing Brain', *Noûs*, vol. 50, no. 2, 2016, pp. 259–285, DOI: 10.1111/nous.12062 (accessed January 10, 2016).

Hurlburt, R.T., *Investigating Pristine Inner Experience: Moments of Truth*, Cambridge, Cambridge University Press, 2011.

Hurlburt, R.T. & C.L. Heavey, 'Investigating Pristine Inner Experience: Implications for Experience Sampling and Questionnaires', *Consciousness and Cognition*, no. 31, 2015, pp. 148–59.

Killingsworth, M.A. & D.T. Gilbert, 'A Wandering Mind is an Unhappy Mind', *Science*, vol. 330, no. 6006, 2010, p. 932.

Lenggenhager, B. et al., 'Video Ergo Sum: Manipulating Bodily Self-Consciousness', *Science*, vol. 317, no. 5841, 2007, pp. 1096–99.

Low, P. et al., 'The Cambridge Declaration on Consciousness', 2012, http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf (accessed January 6, 2016).

Mannino, A. et al., 'Künstliche Intelligenz: Chancen und Risiken', *Diskussionspapiere der Stiftung für Effektiven Altruismus*, no. 2, 2015, pp. 1–17, http://ea-stiftung.org/s/Kunstliche-Intelligenz-Chancen-und-Risiken.pdf (accessed January 11, 2016)

Metzinger, T. (ed.), *Conscious Experience*, Thorverton, Imprint Academic, 1995.

Metzinger, T. (ed.), *Neural Correlates of Consciousness: Empirical and Conceptual Questions*, Cambridge, MA, MIT Press, 2000.

Metzinger, T., 'Faster than Thought', in T. Metzinger (ed.), *Conscious Experience*, Thorverton, Imprint Academic, 1995, pp. 425–61.

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

Metzinger, T., 'Phenomenal Transparency and Cognitive Self-Reference', *Phenomenology and the Cognitive Sciences*, vol. 2, no. 4, 2003, pp. 353–93.

Metzinger, T., 'Précis: Being No One', *Psyche*, vol. 11, no. 5, 2006, pp. 1–35.

Metzinger, T., 'Empirical Perspectives from the Self-Model Theory of Subjectivity: A Brief Summary with Examples', *Progress in Brain Research*, no. 168, 2008, pp. 215–78.

Metzinger, T., *The Ego Tunnel*, New York, NY, Basic Books, 2009.

Metzinger, T., 'Why are Out-of-Body Experiences Interesting for Philosophers? The Theoretical Relevance of OBE Research', *Cortex*, vol. 45, no. 2, 2009, pp. 256–58.

Metzinger, T., 'The No-Self Alternative', in S. Gallagher (ed.), *The Oxford Handbook of the Self*, Oxford, Oxford University Press, 2011, pp. 279–96.

Metzinger, T., 'The Myth of Cognitive Agency: Subpersonal Thinking as a Cyclically Recurring Loss of Mental Autonomy', *Frontiers in Psychology*, no. 4,

2013, p. 931, DOI: 10.3389/fpsyg.2013.00931 (accessed January 6, 2016).
Metzinger, T., 'Why are Dreams Interesting for Philosophers? The Example of Minimal Phenomenal Selfhood, plus an Agenda for Future Research', *Frontiers in Psychology*, no. 4, 2013, p. 746, DOI: 10.3389/fpsyg.2013.00746 (accessed January 6, 2016).
Metzinger, T., 'Two Principles for Robot Ethics', in E. Hilgendorf & J.-P. Günther (eds.), *Robotik und Gesetzgebung*, Baden-Baden, Nomos, 2013, pp. 263–302.
Metzinger T. & Windt J. M. (eds.), *Open MIND*, Frankfurt am Main: MIND Group, 2015, http://open-mind.net
Metzinger T. & Wiese W. (eds.), *Philosophy and Predictive Processing*, Frankfurt am Main, MIND Group, 2017, http://predictive-mind.net
Nietzsche, F., *Beyond Good and Evil*, ed. R.-P. Horstmann & J. Norman, Cambridge, Cambridge University Press, 2002.
Perkins, A.M. et al., 'Thinking Too Much: Self-Generated Thought as the Engine of Neuroticism', *Trends in Cognitive Sciences*, vol. 19, no. 9, 2015, pp. 492–98.
Pyszczynski, T., S. Solomon & J. Greenberg, 'Thirty Years of Terror Management Theory', *Advances in Experimental Social Psychology*, no. 52, 2015, pp. 1–70.
Schmidt, R.E., A.G. Harvey & M. van der Linden, 'Cognitive and Affective Control in Insomnia', *Frontiers in Psychology*, no. 2, 2011, 349, DOI: 10.3389/fpsyg.2011.00349 (accessed January 6, 2016).
Schmidt, R.E. & M. van der Linden, 'The Aftermath of Rash Action: Sleep-Interfering Counterfactual Thoughts and Emotions', *Emotion*, vol. 9, no. 4, 2009, pp. 549–53.
Seth, A., 'Models of Consciousness', *Scholarpedia*, vol. 2, no. 1, 2007, 1328, DOI: 10.4249/scholarpedia.1328 (accessed January 6, 2016).
Seth, A.K., B.J. Baars & D.B. Edelman, 'Criteria for Consciousness in Humans and Other Mammals', *Consciousness and Cognition*, vol. 14, no. 1, 2005, pp. 119–39.
Siderits, M., *Buddhism as Philosophy: An Introduction*, Indianapolis, IN, Hackett Publishing, 2007.
Siderits, M., 'Buddhist Non-Self: The No-Owner's Manual', in S. Gallagher (ed.), *The Oxford Handbook of the Self*, Oxford, Oxford University Press, 2011, pp. 296–315.
Smallwood, J. & J.W. Schooler, 'The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness', *Annual Review of Psychology*, vol. 66, 2015, pp. 487–518.
Tomasik, B., 'The Importance of Wild Animal Suffering', *Relations: Beyond Anthropocentrism*, vol. 3, no. 2, 2015, pp. 133–52. DOI: http://dx.doi.org/10.7358/rela-2015-002-toma.