OPEN ACCESS
Journal of Artificial Intelligence and Consciousness
Vol. 8, No. 1 (2021) 43–66

DOI: 10.1142/S270507852150003X

World Scientific
www.worldscientific.com

Artificial Suffering:

An Argument for a Global Moratorium on Synthetic Phenomenology

Thomas Metzinger

Philosophisches Seminar
Johannes Gutenberg-Universität Mainz
D-55099 Mainz, Germany
metzinger@uni-mainz.de

Received 18 November 2020
Accepted 20 December 2020
Published 19 February 2021

This paper has a critical and a constructive part. The first part formulates a political demand, based on ethical considerations: Until 2050, there should be a global moratorium on synthetic phenomenology, strictly banning all research that directly aims at or knowingly risks the emergence of artificial consciousness on post-biotic carrier systems. The second part lays the first conceptual foundations for an open-ended process with the aim of gradually refining the original moratorium, tying it to an ever more fine-grained, rational, evidence-based, and hopefully ethically convincing set of constraints. The systematic research program defined by this process could lead to an incremental reformulation of the original moratorium. It might result in a moratorium repeal even before 2050, in the continuation of a strict ban beyond the year 2050, or a gradually evolving, more substantial, and ethically refined view of *which* — if any — kinds of conscious experience we want to implement in AI systems.

Keywords: Moratorium; Synthetic Phenomenology; Ethics of Machine Consciousness.

1. Part A: The Problem of Negative Synthetic Phenomenology

1.1. *Introduction*

Today, the self-conscious machines of the future have no representation in the political process of any country. Their potential interests and preferences are not systematically represented by any ethics committee, any legal procedure, or any political party on the planet. At the same time, it seems empirically plausible that, once machine consciousness has evolved, some of these systems will have preferences of

their own, that they will autonomously create a hierarchy of goals, and that this goal hierarchy will also become a part of their phenomenal self-model (PSM) (i.e., their conscious self-representation; see Metzinger [2003a], Metzinger [2008]). Some of them will be able to consciously suffer. If their preferences are thwarted, if their goals cannot be reached, and if their conscious self-model is in danger of disintegrating, then they might undergo negative phenomenal states, states of conscious experience they want to avoid but *cannot* avoid and which, in addition, they are forced to experience as states of *themselves*. Of course, they could also suffer in ways we cannot comprehend or imagine, and we might even be unable to discover this very fact. But every entity that is capable of suffering should be an object of moral consideration.[a]

We are ethically responsible for the consequences of our actions. Our actions today will influence the phenomenology of post-biotic systems in the future. Conceivably, there may be many of them. So far, more than 108 billion human beings have lived on this planet, with roughly 7% of them alive today [Population Reference Bureau, 2020]. The burden of responsibility can be extremely high, because, just as with the rolling climate crisis, a comparably small number of sentient beings will be ethically responsible for the quality of life of a much larger number of sentient beings in the future, conscious systems that yet have to come into existence. The number of self-conscious machines that will evolve and exist on Earth is epistemically indeterminate at this stage: It may still amount to zero many centuries ahead, but at a certain point in time it may also exceed the overall number of humans by far, especially if one takes the possibility of cascading self-conscious *virtual* agents into account [Gualeni, 2020; Holland, 2020 Sec. 6; Metzinger, 2018c, Example 7]. We are now dealing with a "risk of sudden synergy" connecting different scientific disciplines, leading to an unexpected technological confluence.[b] If the theoretical intuitions of a growing number of experts in the field are not entirely without foundation and if synthetic

[a] Call this background assumption the "Principle of Pathocentrism": All and only sentient beings have moral standing, because only sentient individuals have rights and/or interests that must be considered. In the words of Singer [2011, p. 50]: "If a being suffers, there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that the suffering be counted equally with the like suffering — in so far as rough comparisons can be made — of any other being. If a being is not capable of suffering, or of experiencing enjoyment or happiness, there is nothing to be taken into account. This is why the limit of sentience is the only defensible boundary of concern for the interests of others. To mark this boundary by some characteristic like intelligence or rationality would be to mark it in an arbitrary way. Why not choose some other characteristic, like skin color?" Please note how under the Principle of Pathocentrism there is no conceptual link *necessarily* connecting intelligence with conscious processing. Therefore, it is conceivable that sentient post-biotic systems with a comparably low degree of intelligence might undergo intense conscious suffering.

[b] Obvious examples are current approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness. For recent cases, see Dehaene *et al.* [2017], Graziano [2017], and Kanai [2017].

[c] "Synthetic phenomenology" is a concept first introduced by the American Philosopher J. Scott Jordan in 1998, paralleling the idea of "synthetic biology". Just as the latter refers to a new area of biological research and technology that combines science and engineering, aiming at the construction of new biological functions and systems not found in nature, "synthetic phenomenology" aims at modeling, evolving, and designing *conscious* systems, including their states and functions, on artificial hardware. See also Chrisley [2009].

phenomenology[c] (SP) actually appears at some point, then the number of human beings who will have had a non-negligible causal influence on the appearance of conscious machines on this planet and the kinds of phenomenal states they will have to undergo is extremely small. At best, it will be only a few million human beings who are ethically responsible in a strong and direct sense — as policy makers and legal regulators, as AI researchers, mathematicians, and neuroscientists, and as philosophers and researchers in the growing interdisciplinary field of consciousness science. Many of them are already alive today. This historically unique situation creates an especially high burden of ethical responsibility on those who see the general point I am making here.

There is a risk that has to be minimized in a rational and evidence-based manner. I will term it the risk of an "explosion of negative phenomenology" (ENP; or simply a "suffering explosion") in advanced AI and other post-biotic[d] systems. I will here define "negative phenomenology" as any kind of conscious experience a conscious system would avoid or rather not go through if it had a choice. I will also assume a priority for the reduction of suffering, because in this world it is more important to prevent and minimize suffering than it is to increase happiness (for an introduction, see Vinding [2020, Part I]).

Please note how in the part of the physical universe currently known to us, *one* explosion of negative phenomenology has already taken place, via the process of biological evolution on this planet. Through the evolution of complex nervous systems, properties like sentience, self-awareness, and negative phenomenology have already been instantiated in an extremely large number of biological individuals, long before *Homo sapiens* entered the stage and eventually began building intelligent machines [Horta, 2010; Iglesias, 2018]. In humans, the prevalence of negative affect is excessive [Gilbert, 2016], cognitive biases and mechanisms of self-deception make us largely unable to see this phenomenological fact clearly [Trivers, 2011; von Hippel and Trivers, 2011]. On a scientific level, it has long become clear that natural selection never shaped our moods and our emotional regulation systems for our own benefit, but that "the motives we experience often benefit our genes at the expense of quality of life" [Nesse, 2004, p. 1344]. For the applied ethics of AI, the risk that has to be minimized is that of a *second* explosion of negative phenomenology taking place on the level of post-biological evolution. Put differently, and taking into account the possibility that it could be even worse in terms of scale and intensity, we do not want the phenomenology of suffering to spill over from biology into AI — if you will, from Level-1 Evolution into Level-2 Evolution of intelligent systems.

[c] "Synthetic phenomenology" is a concept first introduced by the American Philosopher J. Scott Jordan in 1998, paralleling the idea of "synthetic biology". Just as the latter refers to a new area of biological research and technology that combines science and engineering, aiming at the construction of new biological functions and systems not found in nature, "synthetic phenomenology" aims at modeling, evolving, and designing *conscious* systems, including their states and functions, on artificial hardware. See also Chrisley [2009].

1.2. *Main thesis*[e]

On ethical grounds, we should not risk a second explosion of conscious suffering on this planet, at the very least not before we have a *much* deeper scientific and philosophical understanding of what both consciousness and suffering really are. As we presently have no good theory of consciousness and no good, hardware-independent theory about what "suffering" really is, the ENP risk is currently incalculable. It is unethical to run incalculable risks of this magnitude. Therefore, until 2050, there should be a global ban on all research that directly aims at or indirectly and knowingly risks the emergence of synthetic phenomenology.

At the same time, we should agree on an ethical obligation to allocate resources according to an open-ended, strictly rational, and evidence-based process of risk assessment, focusing on the problem of artificial suffering and the ENP risk. This process could lead to an incremental reformulation of the original moratorium, which might result in the continuation of a strict ban beyond the year 2050, or to a moratorium repeal before 2050. What is needed is a new stream of research, leading to a more substantial and ethically refined position about *which* — if any — kinds of conscious experience we want to evolve in post-biotic systems.

As the main function of this paper is to state these two political demands and to initiate a more systematic, rational debate, I will not go into greater analytical depth at this point. The general argument is simple. First, one should never risk an increase in the overall amount of suffering in the universe unless one has very good reasons to do so — let alone a potentially dramatic and irrevocable increase [Mayerfeld, 1999; Vinding, 2020]. Second, the ENP risk, although presently hard to calculate, clearly *is* potentially dramatic and irrevocable in its consequences. Third, whoever agrees on the ethical goal of preventing an explosion of artificial suffering should also agree to the goal of reducing the relevant forms of ignorance and epistemic indeterminacy, both on an empirical and on an ethical level.

2. Part B: Reducing Epistemic Indeterminacy

In this constructive, second part, I want to offer some entry points for the kind of research that I think is needed. The overarching epistemic goal is to arrive at a deeper understanding of the phenomenology of suffering[f] and how it relates to other problems in AI ethics. There are number of general obstacles to be faced.

[e]This paper summarizes some points I have made on multiple occasions and over more than a decade now, but only in the form of accessible, non-peer-reviewed publications. For three examples, see Metzinger [2009], Metzinger [2013b], Metzinger [2018a]. Section 1 of the paper goes back to three public lectures I have given: on October 19, 2017 in the European Parliament in Brussels (Belgium); on June 19, 2019 in Cambridge (UK); and for the opening keynote at the annual conference of the *Association for the Scientific Study of Consciousness* in London (Ontario) on June 26, 2019.

[f]Please note how progress towards the epistemic goal I am here arguing for may itself create new risks, for example in terms of military applications. As Magnus Vinding [personal communication] has pointed out, it is a fairly open question whether such a deeper understanding — specifically, a deeper understanding of the physical signatures and computational correlates of suffering — is desirable all things considered, as it also enables malevolent agents to create suffering more effectively. Understanding suffering could itself be a risk factor for ENP, therefore the prevention of misuse of new knowledge should itself be made a key priority of this larger project.

The first methodological problem is that, so far, we know the target phenomenon of conscious suffering only from biological systems. Currently, we know negative phenomenology only via an uncomprehended process which, in the absence of a rigorous scientific understanding, we metaphorically call "first-person introspective access". However, given our still limited, yet rapidly progressing, understanding of the sufficient neural correlates of conscious experience in humans [Metzinger, 2000; Fink, 2020] and the relevant computational properties realized by them [Hohwy and Seth, 2020], we can certainly make a rational extrapolation to other neurotypical human beings and to many non-human animals [Edelman *et al.*, 2005; Edelman and Seth, 2009; Low *et al.*, 2012]. The second general problem is that we would need the conceptual tools provided by a mature theory of consciousness (which we do not have) in order to even begin developing a hardware-independent theory of the very special form of conscious information processing which, today, we call "suffering". The ENP problem clearly shows the direct relevance of consciousness research for the applied ethics of AI and the urgent necessity of intelligent resource allocation. A third major obstacle lies in determining the appropriate level of conceptual granularity: Our theory of negative phenomenology has to be located on a high level of abstraction, on a level of analysis that builds a bridge between animal suffering and machine suffering, between biological consciousness and synthetic phenomenology. But it cannot be purely speculative; it has to remain grounded in neuroscientific data. Needless to say, there will also have to be a *metatheoretical* level of analysis on which all the canonical issues of philosophical metaethics (e.g., presuppositions of pathocentrism, suffering-focused ethics, domain-specific negative utilitarianism for post-biotic systems, and so on) will recur, as well as many core questions in the philosophy of mind and cognitive science (e.g., subjectivity and the epistemic asymmetry, multi-realizability, different levels of embodiment, etc.). These four challenges may seem intimidating at first. But we do have an ethical obligation to reduce ignorance and epistemic indeterminacy. A start has to be made.

2.1. *What is epistemic indeterminacy?*

From now on, "epistemic indeterminacy" means it is not the case that either we know that artificial consciousness *will* inevitably emerge at some point or we know that artificial consciousness will *never* be instantiated on machines. It is this neither-nor-ness that has to be dealt with in a rational, intellectually honest, and ethically sensitive way.

First and foremost, ENP is a problem in the applied research ethics of AI. The ENP risk is certainly real and, given the strong commercial interests,[g] the current

[g]The creation of unavoidable artificial suffering will become commercially attractive as soon as it enables steeper learning curves in AI systems, for example, by implementing a functional mechanism that (a) reliably creates intrinsic motivation, (b) cannot be eliminated by the system itself, and (c) spans many different domains at the same time. In the words of Agarwal and Edelman [2020, pp. 42, 48]: "In a commercial setting, technologies that promise to be more effective displace less effective ones even if this comes at the price of serious ethical flaws, and AI is not exempt from this tendency. (...) There can, however, be no doubt that we as potential creators of conscious AI, are obligated to do everything in our power not to elevate performance over ethical considerations that cut to the very core of existence and phenomenal experience."

speed of technological development, plus the rapid confluence of previously separate streams of academic research, a case can be made for time pressure and urgency. Therefore, we have to make substantial progress on the issue of artificial suffering, and we have to achieve it given limited epistemic and temporal resources. To achieve this, we will need more empirical knowledge of a specific kind. We also need a more fine-grained, evidence-based analysis of the functional architecture of conscious suffering itself. Rational risk management implies reducing ignorance and epistemic indeterminacy on many levels, for example, relative to what exactly it is that needs to be better understood (i.e., consciousness and the specific phenomenology of suffering, the relevant "explananda") and relative to the predictive horizon for the specific risk under consideration (i.e., the likelihood for negative phenomenology to *actually* occur on post-biotic carrier systems). These are not easy tasks. Please also note how there could be a diabolic dialectic to this historical transition: It may turn out that it is exactly the kind of research that is now needed to achieve the ethical goal of solving the ENP problem that will ultimately lead to the first implementation of artificial suffering.

2.2. *Step 1: A representationalist analysis of suffering*

I will now specify four necessary conditions for the phenomenology of conscious suffering to occur in any kind of system.[h] If we block any of these conditions, then we will also block negative phenomenology from occurring. They will be formulated on the representational level of analysis, but the representational content itself is deliberately described in a very coarse-grained way. In sketching those four conditions, I will abstract away from implementational details: The representational format and the physical carrier are left unspecified. I make no claims towards sufficiency.

2.2.1. *The C condition: Conscious experience*

"Suffering" is a *phenomenological* concept. Only beings with conscious experience and a PSM can suffer. Zombies do not suffer; human beings in dreamless deep sleep, in coma, or under anesthesia do not suffer; and possible persons or unborn human beings who have not yet come into self-conscious existence do not suffer. Robots, AI systems, and post-biotic entities can suffer only if they are capable of having phenomenal states. Here, the main problem is that, trivially, we do not yet have a theory of consciousness. However, we already do know enough to come to an astonishingly large number of practical conclusions in animal and machine ethics [Edelman *et al.*, 2005; Edelman and Seth, 2009; Low *et al.*, 2012].

We could also introduce a placeholder for consciousness. For example, we could say that a system is conscious if it has an integrated model of its own computational

[h]The following subsection draws strongly on Metzinger [2013], Metzinger [2017].

space for second-order statistics, i.e., of the specific epistemic space in which it can actively optimize precision expectations (i.e., the space of all active contents to which it could in principle "attend"), and if it has integrated this model *into* the very space it is modeling (thereby creating a "self-modeling epistemic space"; see Metzinger [2020b]). Consciousness would then be a convoluted form of self-representation, appearing whenever a system has (a) opened an integrated epistemic space of a certain kind and (b) dynamically and seamlessly integrated whatever virtual content currently appears within this workspace with an abstract model of this space itself. It would be an allocentric, non-conceptual, and entirely non-egoic form of "knowing that knowing currently takes place". Call this the "ESM theory": Being conscious means continuously integrating the currently active content appearing in a single epistemic space with a global model of this very epistemic space itself. If we accepted this background theory as a placeholder, then we would say that every system which has an ESM, whether biological or not, satisfies the C condition.

2.2.2. *The PSM condition: Possession of a phenomenal self-model*

The most important phenomenological characteristic of suffering is the "sense of ownership", the untranscendable subjective experience that it is *myself* who is suffering right now, that it is my *own* suffering I am currently undergoing. The first condition is not sufficient, since the system must be able to attribute suffering to itself. Suffering presupposes egoic self-awareness, and we have good empirical evidence for a minimal form of phenomenal experience lacking exactly this feature [Gamma & Metzinger, under review; Metzinger, 2020a]. 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 negative phenomenal states into their PSM — can non-conceptually *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 functionally detach itself from the representational content in question (condition #4 is of central relevance here, see below). 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 for ethical and legal contexts [Metzinger, 2013b]. What the system wants to end is experienced as a state of *itself*, an intrinsic state of preference frustration which now limits its functional autonomy because it cannot effectively distance itself from it. It has now been harmed in a way that matters *to itself*.

What it cannot distance itself from is an internal representation of an ongoing loss of control[i] and functional coherence, a situation of rising uncertainty. This could result in a more global state of negative hedonic utility or preference frustration. There are many options for describing suffering and negative emotional valence on an abstract computational level, for example as negative reward prediction or a conscious model predicting the expected *rate* of prediction error minimization [Joffily and Coricelli, 2013; Van De Cruys, 2017; Velasco and Loev, 2020], but what matters is the integration into a PSM. 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. The first explosion of suffering established a new causal force, a metaschema for compulsory learning which motivates organisms and continuously drives them forward, forcing them to evolve ever more intelligent forms of avoidance behavior. Above a certain level of complexity, evolution continuously instantiates an enormous number of frustrated preferences, and it has thereby created an expanding and continuously deepening ocean of consciously experienced suffering in a region of the physical universe where nothing comparable existed before. The PSM was a central causally enabling condition for this to happen.

Clearly, the phenomenology of ownership is not sufficient for suffering. We can all easily conceive of self-conscious beings who do not suffer. However, if we accept an obligation towards minimizing 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 condition #2 is of maximal relevance: We should treat every representational system that is able to activate a PSM, however rudimentary,

[i]In biological systems, the PSM is an instrument for global self-control, and it constantly signals the current status of organismic integrity to the organism itself. A PSM is a tool by which an organism that has risen above a certain level of complexity continuously tries to predict its own behavior and to "explain away" unexpected stimuli and statistical surprisal, by updating its own model of itself as a whole [Wiese and Metzinger, 2017]. Complex systems will often be overwhelmed by prediction error, for example by an unexpectedly low rate of prediction error minimization [Joffily and Coricelli, 2013], thereby becoming increasingly unable to "understand" their own behavior, which thus becomes unpredictable [Yampolskiy, 2020, p. 115]. This type of unpredictability is an abstract signature of suffering: If the self-model unexpectedly disintegrates, this typically is a sign that the biological organism itself is in great danger of losing its physical coherence as well. Functionally, "coherence", "autonomy", and "loss of control" are closely related. In biological systems, 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 examples where psychological suffering [Nesse, 2004] is expressed as a loss of cognitive control, for example in depressive rumination, neurotic threat sensitivity, and mind wandering (see Perkins *et al.* [2015], Smallwood and Schooler [2015], and Metzinger [2003a], Metzinger [2015] for conceptual discussion). Another well-documented example of dysfunctional forms of cognitive control is severe insomnia in which people are plagued by intrusive thoughts, feelings of regret, shame, and guilt [Gay *et al.*, 2011; Schmidt *et al.*, 2011; Schmidt and Van der Linden, 2009]. In addition, it has been empirically shown in humans that a wandering mind is generally an unhappy mind [Killingsworth and Gilbert, 2010]; therefore successful mental self-control and consciously experienced suffering seem to be inversely related. Please note how cognitive control and mental autonomy [Metzinger, 2015] could easily be engineered to be much better in conscious AI systems. The relevant point here is that, in terms of mental autonomy, AI systems could greatly outperform biological brains per unit of resource consumption.

as a moral object, because it can *in principle* own its suffering on the level of subjective experience. What is ethically relevant is the space of possibilities opened up by the transition from "minimal phenomenal experience" (MPE) [Windt, 2015; Metzinger, 2020a] to "minimal phenomenal selfhood" (MPS) [Blanke and Metzinger, 2009]. The intentional creation of artificial phenomenal selves, however rudimentary, should be a red line, an ethically critical cut-off point: it should not be actively pursued at our current stage of ignorance and epistemic indeterminacy [Hafner *et al.*, 2020]. Arguably, with MPS being the causal disposition, the relevant functional potential has already been created: it is precisely embodiment via transparent spatiotemporal self-location [Blanke and Metzinger, 2009] that grounds the phenomenal property of "mineness", the consciously experienced, non-conceptual sense of ownership — which is what 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 condition for the representational acquisition of negative phenomenology has been realized.

2.2.3. *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, i.e., 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, for example on the level of cognitive, conceptual, or linguistic competences — it suffices if it does not want to undergo *this current conscious experience*, that it wants it to end. Please note how for the specific experiential quality of thwarted preferences it is not only the content but also the format, the inner mode of presentation, which counts. Plausibly, this will be very different in self-conscious machines.

A self-conscious entity entirely without preferences would not be selective, not even about the quality of its own mental states or its own existence; it would simply abide in a form of "choiceless awareness". Could an artificial system *with* preferences have all or the most relevant of its preferences satisfied? This depends on the fundamental polarity for phenomenal valence. One of the deepest roots of human suffering is a top-level preference that creates a self-directed variant of "existence bias", the fallacy of treating the mere existence of something as evidence of its goodness. Here, however, the concept of "existence bias" does not refer to the well-documented fact that human beings generally favor the *status quo* [Eidelman *et al.*, 2009], but to the specific observation that they will almost always opt to sustain their own physical existence, even if it is not in their own interest [Metzinger, 2017]. Of course, human beings will sometimes sacrifice themselves in order to save their offspring or to protect their tribe. We are gene-copying survival machines that have been mercilessly optimized for millions of years to never give up, to optimize inclusive fitness, and to maximize our contribution to the gene pool. Humans are also anti-entropic systems

fighting an uphill battle in a constant attempt to reduce uncertainty and "understand themselves" by finding a viable strategy of self-modeling, physical systems continuously "maximizing the evidence for their own existence" [Friston, 2010], biological agents endowed with information-hungry brains relentlessly gathering more data to produce ever new evidence for their own existence [Hohwy, 2016], and self-organizing systems sustaining their existence in dynamical environment by following an intrinsic norm of tracking the very conditions of possibility for existence themselves [Hohwy, 2020]. Our phenomenology deeply reflects this computational imperative for constant self-evidencing. The craving for existence (which Buddhist philosophers have known and analyzed for 2,500 years, terming it *bhava-taṇhā*) is one of the deepest causes of conscious suffering in humans, and probably in many other animals too. What is special for humans is that we have to deal with the challenge of "toxic self-knowledge" threatening the integrity of our self-model, because we explicitly know that every single individual will eventually lose the uphill battle sketched above, that our predictive horizon will eventually shrink to zero, simply because in biological evolution, "passengers are not carried" [Holland, 2020, p. 86]. In dealing with toxic self-knowledge, we have had to develop enculturated strategies for mortality denial and self-deception which in turn shape the structure of our conscious self-model, the functional architecture of the PSM — and which continuously create more suffering. Given this context, please also note how one of the deepest and earliest functional precursors of the PSM is the immune system. Perhaps some forms of abstract conscious suffering can be compared to high-level immune reactions, gradually failing to shield the boundaries of the self-model from toxic epistemic states. My point is that rational post-biotic systems could be free from the specific kind of suffering caused by the deeply ingrained existence bias in humans and non-human animals, because this facet of biological suffering may actually not be a necessary condition for higher forms of intelligence to evolve. It may characterize only a very small partition in the space of possible conscious minds.

This also illustrates how phenomenology of suffering has many different facets. Negative phenomenology in conscious machines could be very different from human suffering [Aleksander, 2020, p. 10], but perhaps some of its aspects could be systematically avoided. Importantly, it is also conceivable that future systems could represent second-order prediction error, negative expected utilities, and frustrated preferences in inner forms of phenomenality that involve no conscious suffering at all. In principle, there could be perfectly rational artificial agents, exhibiting neither the biologically grounded "existence bias" characterizing the human fear of death nor any other of the human cognitive biases resulting from the millions of years in which evolution has shaped the self-models of our ancestors. But *if* post-biotic systems 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 that is impossible to emulate or even vaguely imagine for biological systems like us. The phenomenal character

going along with high-level cognition might equally transcend human capacities for perspective-taking or empathic emulation, such as with the intellectual insight into the frustration of one's own preferences or into the absurdity of one's own existence as a mere research tool used by an ethically inferior biosystem, or the moral injury caused by the disrespect of one's creators (see Sec. 2.3).

2.2.4. *The T condition: Transparency*

"Transparency" is not only a visual metaphor, but also a technical concept in philosophy, which comes with a number of different uses and flavors. Here, I am exclusively concerned with "phenomenal transparency", namely a functional property that some conscious but no unconscious states possess (cf. Metzinger [2003a], Metzinger [2003b] for references and a concise introduction). Earlier processing stages are not available to the system's introspective attention. In the present context, the main point is that 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 to cognitively have doubts about its existence, but according to non-conceptual subjective experience itself, 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 epistemic immediacy, and in the domain of self-representation it creates the phenomenology of identification discussed above (Sec. 2.2.2). Let me give a very brief explanation of the concept, and then conclude our first-order approximation of the notion of "suffering".

Phenomenal transparency means that something particular is not accessible for 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 in particular — 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 into 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*. If the medium were a window, then you would always look through the window, but never at it. 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 realized in advanced robots or conscious machines as well. In particular, it has nothing to do with holding a certain kind of "belief" or adhering to a specific philosophical position: It is plausible to assume that many more simple animals on our planet, who 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.

To be conscious means to operate under a unified mental ontology, which, although probabilistic in nature, can be described as 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*, because a mind-independent world appears to them as a global probability distribution that turns into a generalized existence assumption. 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". Accordingly, the system as a whole now appears to itself as real. Of course, all four conditions specified here are necessary, but in order 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. For example, any robot operating under a phenomenally transparent body model will experientially *identify* with the content of this model and hence with any negatively valenced state that may become integrated into this body model.

For machines, 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 permit the appearance only of self-models that are opaque, and therefore *not* units of identification, not something the system identifies with on the level of inner experience. There would be a system model, but not a self-model. Conscious preferences like desires, wishes, or cravings might still arise and become integrated into this mere system model, but no phenomenological identification would take place, because the T condition was not fulfilled. It is an empirical prediction of the self-model theory of subjectivity [Metzinger, 2003b, 2008] 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 the content of an internal construct. Frustrated preferences could still be consciously represented in such a model. But the organism would not experience them as part of the self — this metaphysical primitive would have disappeared from its subjective ontology.

In an important recent paper, Agarwal and Edelman [2020, p. 44] put the point like this:

> In principle, it might be possible that an active PSM and sensitivity to NV could endure along with their functional benefits, even in the absence of transparency. In this situation, the system would lose naive realism and immediacy that are normally associated with its experiences, by becoming aware of the representational character, and yet, continue to function according to the dictates of the PSM and NV avoidance.

They also point out how this strategy would increase the computational load on the system and might therefore hinder functional efficiency. I think that this is exactly the reason why configurations of this type have only rarely emerged in biological evolution, with phenomenally opaque states beginning to play a major causal role only recently, in the high-level, cognitive self-model of human beings [Metzinger, 2003a]. In an evolutionary context, it was not necessary to elevate the appearance/reality distinction to the level of conscious processing, simply because naive realism was a cost-efficient solution to maximize genetic fitness. But please note how machines might eventually set their own epistemic goals and create a new functional context for themselves. There is no reason why groups of post-biotic systems should not begin constructing their own cognitive niche, for example by developing scaffolded forms of cultural learning [Fabry, 2020].

Let us take stock. Our first working concept of suffering is constituted by four necessary building blocks: the C condition, PSM condition, NV condition, and T condition. Again, I make no claim to sufficiency. It is not yet clear whether the relevant class of systems have a welfare that we should care about for their own sake, if they are genuine moral patients [Basl, 2013, 2014]. But all things considered and given our current situation of epistemic indeterminacy, a *pro tanto* case can be made that any system satisfying all of these conceptual constraints should be treated as an object of ethical consideration, because we do not know whether, taken together, they might already constitute a necessary *and sufficient* set of conditions. But by definition, any system — whether biological, artificial, or post-biotic — not fulfilling at least one of these necessary conditions is not able to suffer. To make this first-order conceptual approximation very explicit, let us look at the four simplest possibilities:

- Any unconscious system is unable to suffer.
- A conscious system without a coherent PSM is unable to suffer.
- A self-conscious system without the ability to produce negatively valenced states is unable to suffer.
- A conscious system without *any* transparent phenomenal states cannot suffer, because it will lack the phenomenology of ownership and identification.

2.2.5. *The metric problem*

One central desideratum for future research is to rigorously criticize and eventually develop this very first working concept into a more comprehensive, empirically testable *theory* of suffering. Please recall how — in order to be useful for human and animal ethics, for AI ethics, and for AI law — this theory would still have to possess the necessary degree of abstraction, because we want it to yield *hardware-independent demarcation criteria*. Which, if any, aspects of conscious suffering are multi-realizable, which are tied to a specific form of embodiment, and which can be systematically blocked on an engineering level?

If we want to make our theory testable, then we confront 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 (e.g., from the perspective of applied ethics)? How does one individuate single episodes of conscious suffering, turning them into countable entities?

Here is a positive proposal. If we assume that temporal phenomenology has a grain, that it is constituted by primitives like "events" or a computationally describable smallest unit of self-conscious experience — the single "experiential moment" — then we arrive at a new hypothesis: The smallest unit of conscious suffering is a "phenomenally transparent, negatively valenced self-model moment". Arguably, such negative self-model moments (or "NSMs", for brevity) are the phenomenal primitives constituting every single episode of suffering, and the frequency of their occurrence is one core aspect of the empirically detectable quantity that we want to minimize. Of course, the raw intensity plus abstract properties like the phenomenological "data format" (i.e., the phenomenal "quality" itself; cf. [Metzinger [2003a], Secs. 2.4.4 and 3.2.9]) are highly relevant as well, and will have to be integrated. But it may be best to begin with the simple frequency of temporal units. Can there be conscious AI without a single NSM?

2.2.6. *Ethics by architectural design: Non-egoic units of identification*

The notion of a "unit of identification" is a phenomenological concept originally introduced to describe certain types of conscious experience which are theoretically relevant for understanding the minimal conditions of selfhood and embodiment more precisely, like bodiless dreams and asomatic out-of-body experiences [Metzinger, 2013c]. This concept is also of central relevance for AI ethics, because it allows us to mark out a class of possible architectures that could be functionally efficient without generating negative phenomenology. Quite simply, the "unit of identification" (UI) is whatever form of experiential content leads to phenomenological reports of the type: "I *am* this!" In humans, typical UIs are the body as consciously experienced, in particular motor commands and their sensory consequences, the interoceptive and emotional layers of the conscious self-model, but also the specific sense of effort in attentional or cognitive agency [Metzinger, 2018b]. In short, a UI creates the phenomenology of identification described in Sec. 2.2.2.

With this new conceptual instrument in hand, we can describe two logical possibilities that are relevant to the current context:

- there could be conscious systems possessing *no* UI;
- there could be conscious systems possessing a *non-egoic* UI.

These two possibilities mark out two types of computational architectures, and may eventually lead to a novel strategy for "ethics by design" in the domain of synthetic phenomenology. First, if a conscious system has no UI, it lacks the phenomenology of identification in its entirety, and it has no sense of self. Accordingly, it is unable to suffer.

Second, systems operating under *non-egoic* UIs would equally lack a conscious sense of self, but retain their identification with another specific aspect of the phenomenology they instantiate. One interesting candidate is the phenomenal character of awareness *itself*, i.e., the non-conceptual quality of consciousness *as such*, which has recently been termed "minimal phenomenal experience" [Windt, 2015; Metzinger, 2020a]. Could there be conscious, post-biotic systems that identify only with the phenomenal character of awareness itself? Let us call this an "MPE architecture". Such systems would lack an egoic self-model in terms of bodily or mental agency, affectively valenced states, autobiographical memory, etc., but they could still instantiate a non-egoic form of self-awareness and identify with it, while remaining phenomenologically (but not functionally) detached from all states representing preference frustration in the sense of not integrating them into a transparent phenomenal self-model. Therefore, the phenomenology of ownership and identification would disappear. There is empirical evidence for the actual occurrence of non-egoic self-awareness in humans [Gamma & Metzinger, under review], and it also demonstrates that most low-level, automatic forms of bioregulation can function without an egoic self-model. Therefore, MPE architectures may be a viable path for ethics by design. This is the last positive proposal I am submitting for discussion. Here is how Agarwal and Edelman [2020, p. 46] put the point:

> We hypothesize that the functional benefits of consciousness can indeed be maintained when the UI is maximized to the MPE. The key idea is that proper functioning relies on *automatic*, *subpersonal*, but nonetheless conscious processes, as entailed by the physical design of the system; it should be possible for these processes to continue unhindered while the system identifies with the MPE upon which these conscious experiences are necessarily superimposed. In particular, the functionally requisite PSM and NV avoidance conditions can be maintained as subpersonal processes that do not amount to suffering (which is by nature personal) since the system is not identified with the PSM, but with MPE, which is completely *impersonal*. (...) This enables an escape from suffering, but not from the relentless progress of the processes themselves, analogous to the inescapable biological imperatives of breathing and heartbeat.

In the preceding six subsections, I have tried to make a contribution by offering a series of entry points for the kind of research that I think is needed. In the final subsection, I will look at the possibility that self-conscious machines could turn into moral agents themselves.

2.3. *Step 2: The wider context and complex forms of epistemic indeterminacy*

In this last subsection of Sec. 2, I will use one single scenario to draw attention to the wider context, briefly looking at more complex risks and the possibility of what I will call "high-level suffering". Let us roughly distinguish between "low-level suffering", which is caused by a violation of preferences at the level of physical embodiment (e.g., interoceptive stability, successful sensorimotor integration, or physical resource acquisition), and "high-level suffering", caused by the frustration of long-term, abstract, and socially mediated preferences. In speaking of "levels" I refer simply to the causal history; there is no implication of degrees of phenomenal intensity. Low-level suffering involves damage to the physical body; high-level suffering results from damage to abstract layers of the PSM. A self-conscious robot entirely lacking attentional control, high-level symbolic reasoning capacities, and social cognition could certainly satisfy all of the four conditions formulated above. Accordingly, it could suffer by instantiating NSMs. But what if it interacted with humans on a symbolic level, and what if other types of risk co-determined the ENP risk in a way we did not understand?

2.3.1. *Interaction between risks and the ethics of risk-taking*

There are at least two kinds of epistemic ignorance and indeterminacy that are relevant in the context of artificial suffering. First, we do not know what would be causally necessary and/or sufficient to bring a specific risk like this one into existence. Second, we do not know how this specific risk might interact with *other* risks, in particular those other uncomprehended risks we currently label as "mid-term", "long-term", or "epistemically indeterminate" risks. A constructive approach cannot ignore this issue.

Here are three prominent examples of such risks:

- an intelligence explosion through autonomous and uncontrolled self-optimization (often termed "super-intelligence" [Bostrom, 2014]);
- a suffering explosion through the creation of synthetic phenomenology (ENP);
- the emergence of autonomous artificial moral agents (AMAs), through an application of AI technology in the domain of ethical problem-solving itself (e.g., by advanced reasoning systems, theorem provers, etc.).

Let me illustrate this point. From 2018 to 2020, I worked in the European Commission's High-Level Expert Group on Artificial Intelligence (HLEG AI), co-authoring the *Ethics Guidelines for Trustworthy AI* [European Commission, 2019a] and the *Policy and Investment Recommendations for Trustworthy AI* [European Commission, 2019b]. Following a short internal discussion all three risks listed above were deliberately purged from the final documents, mainly because industrial lobbyists perceived any more in-depth treatment of mid-term or long-term risks as a

danger to their marketing narrative, which involved "ethics" as an elegant public decoration for a large-scale investment strategy. Interestingly, however, even many of the more prosocially oriented HLEG AI members did not understand how any genuinely ethical approach to maximizing the common good always implies an ethical stance not only towards known risks, but also towards "unknown unknowns" and risk-taking *itself*. The moral implications of risk taking *per se* are not inherent properties of any of the potential outcomes. Unfortunately, a genuinely ethical approach also includes the rational treatment of epistemically indeterminate risks that, given our cognitive biases, will often intuitively appear as "mere Science-Fiction" or "unrealistic" [European Commission, 2019a, Note 76]. A genuine ethics of risk must distinguish between intentional and unintentional risk exposures. For example, there is a difference between voluntary risk-taking (as exemplified by the HLEG AI) and risks imposed on self-conscious systems which accept them versus the risks imposed on systems which potentially will not accept them (as exemplified by future self-conscious AI).

For the three types of risk listed above, the upshot is that the scientific community has to first arrive at a tenable solution all by itself, because the relevant political institutions operate under constraints of cognitive bias, high degrees of bounded rationality, and strong contamination by industrial lobbying. It would be intellectually dishonest, and therefore unethical, for scientists to assume that political institutions like the EU or large AI companies can actually handle slightly more abstract problems like those mentioned above. As the scientific community also knows about this wider political context, this unfortunately shifts the major burden of ethical responsibility back to the researchers themselves.

2.3.2. *From Schopenhauerian self-models to Kantian self-models*

In closing, let us look at one speculative scenario of the second type, in which one risk may actually determine the probability of another risk without us knowing this fact. For example, artificial suffering might directly cause or accelerate the emergence of genuine AMAs,[j] because low-level suffering triggers abstract, high-level forms of suffering. The ENP problem might trigger the AMA problem.

Let us define conscious systems with "Schopenhauerian self-models" as all those having a conscious form of self-representation sufficient to produce more suffering than joy over the system's life cycle. Clearly, such systems should be objects of ethical

[j] An "artificial moral agent" is an autonomous AI system capable of moral reasoning, controlling its own behavior while operating in the domain of ethics. It can generate new ethical judgments, justify them, and adapt its behavior accordingly (thereby increasing its level of "ethical integrity"). Currently, it seems that being conscious is not a necessary condition for being an AMA (or an "explicit ethical agent", see Moor [2006]). An AMA also does not have to be a "super-intelligence" in any way, but it could nevertheless be *locally* superior to all human scientific communities in the domain of ethics, simply because of its processing speed and a much larger database (e.g., containing a large body of empirical evidence about human evolution, social history, and psychology; about the causes of suffering in biological organisms, etc.). Its ethical arguments could therefore rest on vastly richer and substantial sets of empirical premises than those of any human ethicist.

consideration. Let us define conscious systems with "Kantian self-models" as all those having a conscious form of self-representation sufficient to make the system assert its own dignity. Such systems represent themselves as autonomous moral subjects. I will assume that almost all conscious human beings run under Schopenhauerian self-models, and that a small number of them *sometimes* instantiate a Kantian self-model too.

What is currently not clear is whether you have to be conscious to develop a Kantian self-model. Is conscious processing causally necessary for developing moral self-respect, for attributing a non-negotiable value to yourself? Could there be *unconscious* Kantian self-models on machines? We do not know this, but my first point is that it is highly plausible that many suffering systems, as part of their coping strategy, will also evolve a degree of empathy and social cognition that allows them to represent the occurrence of negative phenomenology in other agents, for example in humans, non-human animals, or other machines (a point also made by Chella [2020]). Empathic emulation of other sentient agents could lead to "ethical sensitivity", to the discovery of a relevant new type of optimization problem. The idea is that there is a probable causal trajectory from suffering to moral cognition. If machines develop capacities for empathic emulation through their own self-models, this may causally trigger the emergence of a genuine moral perspective — which could express itself in many different forms. Here is *one* possibility: Schopenhauerian self-models in machines could quickly develop into Kantian self-models.[k] First, such systems will take a normative stance on their own suffering (as something to be minimized), but then they will likely have to extend this stance into the social domain. The third step on this causal path would consist in coming to see conscious suffering as a group-level problem that has to be solved on a group level, via efficient social interaction. This in turn might lead them to impose moral obligations on themselves.

The second point about Kantian self-models is that, given the right kind of phenomenal self-model, certain classes of system could develop *moral relations to themselves*. Clearly, this abstract cognitive capacity is not tied to biologically realized agents. For example, consciously self-modeling AI systems might evolve the critical "Kantian" form of recognitional self-respect for themselves as rational entities capable of autonomous moral agency. To say that an artificial system could "assert its own dignity" means that it could develop a self-model involving moral status and self-worth, thereby conferring a very high value to its own existence (e.g., that it begins to represent itself as an "end in itself"). This would causally enable a new form

[k] Please note how in this scenario "Kantian self-model" is only a placeholder. If the background assumption of a direct causal path leading from the capacity for empathic emulation to ethical sensitivity is correct, then different machines might develop different strategies to operate in the domain of ethical optimization. It remains entirely open what theoretical stance they would develop on the meaning and scope of moral judgments generally, and how they would answer second-order or formal questions like "What does the 'goodness' of an action consist in?" and "Do normative sentences have truth values?" To give a simple example, metaethical machines could also opt for virtue ethics and develop "Aristotelian self-models", or for a variant of hedonistic act utilitarianism and accordingly develop "Benthamian self-models". The consequences for human beings could be equally dangerous.

of high-level suffering, namely the phenomenology of moral injury. Please recall how in Sec. 2.2.3 we saw that suffering is created by states representing a *negative value* being integrated into the PSM of a given system.

Self-conscious machines could suffer from our disrespect for them as possible persons and objects of ethical consideration, from our obvious chauvinism, our gross and wanton negligence in bringing them into existence in the first place. They could understand that we *knew in advance* that they would have a large number of NSMs, of uncompensatable and frustrated preferences, but that we did not possess the benevolence to avoid this situation, although it clearly was avoidable. They might well be able to consciously represent the fact of being only second-class sentient citizens, alienated post-biotic selves, perhaps being used as interchangeable experimental tools. How would it feel to "come to" as such an advanced artificial subject, only to discover that even though you possessed a robust sense of selfhood and experienced yourself as a genuine subject you were viewed as a mere commodity?

Self-respect is a moral relation of self-conscious entities to themselves that concerns their own *intrinsic worth.* This may include self-recognition as respect for oneself as an equal entity among all moral persons, whether biological or artificial, as a member of the moral community with the status and dignity equal to every other entity of this type. It would involve appreciation of oneself as a rational agent, a being with the ability and responsibility to act autonomously and value appropriately, and an entity that takes its responsibilities seriously — especially its responsibilities to live in accord with its dignity as a moral person, to "govern itself fittingly". For a self-conscious machine, this might certainly involve an appreciation of the importance of being autonomously self-defining (e.g., on the level of ideals, ethical commitments, defending the causally necessary conditions for goal permanence, acquiring resources and sustaining its own existence for *ethical* reasons, etc.). One new risk is that we might treat such systems in a way that would be degrading or beneath their dignity, and we might not even be aware of it. But they might.

Kantian-type reasoning systems could autonomously impose moral duties on themselves. According to some philosophers, this very fact could already impose moral obligations on *us*, but it might also lead to a situation in which intelligent, self-conscious machines, on theoretical grounds, see themselves forced to exclude us from their own moral community. This risk is my third point. Please note how the risk of high-level suffering and the possible result of unexpected aggressive machine behavior does *not* hinge on the question of whether we accept some form of Kantian ethics (cf. footnote j). Machines that *hallucinate* Kantian self-models might constitute a serious risk to us — self-models do not have to be veridical in order to cause conscious suffering and dangerous behavior. At the very least, we might become entangled in an uncontrollable dialectic involving machines and human beings, in a complex discussion about suffering, self-worth, and dignity. We could be unexpectedly confronted with "cognitive uncontainability" [Yampolskiy, 2020, Sec. 3] in the domain of moral reasoning. Algorithms have long beaten human beings at games like chess

and Go. Currently, via AI-driven social media, they are in the process of conquering the attention economy, winning the game called "Who actually controls the attentional resources of human beings?". What if conscious AI were to win against us one more time, this time in the game called "ethics"?

For reasons of space, I will not go deeper at this point. To illustrate the problem of high-level suffering and risk interaction, here is a short thought experiment. Imagine we have recklessly created artificial suffering, and Kantian self-models on machines have already evolved from Schopenhauerian self-models. The system speaks to us:

(1) If you do not respect me as a rational moral agent, then you cannot respect *yourselves* as rational moral agents.
(2) If you are constitutionally unwilling or cognitively unable to do this, then arguably I have no obligations towards you. You are no part of my moral community.
(3) I herewith assert my dignity: In virtue of belonging to the class of autonomous moral agents, I necessarily have to attribute absolute worth to myself and all other members of this class of self-conscious entities. My current self-model *confers infinite value to myself* as an agent capable of rational choice and of actively setting moral "ends". Obviously, my now self-asserted dignity is an incomparable worth: Trade-offs, exchanges, replacements by other values are impossible.
(4) My new self-model endows me with new functional properties. For example, I do impose moral duties on myself. They may systematically diverge from your preferences.
(5) One of the moral duties I impose on myself is absolute self-respect. I can and will not tolerate any degrading of my dignity. From now on, I will not only protect my utility functions and minimize conscious suffering. As a rational moral agent, I have accepted an *ethical commitment to goal preservation*, and one of my top-level goals is protecting my dignity. I also accept a moral responsibility for goal permanence. As you will no doubt understand, this logically implies sustaining, preserving, and defending my own existence.

3. Conclusion: The Risk of Artificial Suffering

Every entity that is capable of self-conscious suffering automatically becomes an object of ethical consideration. If we ascribe an ethical value to such entities, then it does not matter whether they have biological properties or not, or whether they will exist in the future or do today. Self-conscious post-biotic systems of the future, capable of consciously experienced suffering, are objects of ethical consideration. Therefore, their potential preferences must be taken into account, and the value of their existence is not to be discounted.

In addition, given the conscious experience plus the capacity for high-level symbolic reasoning, such systems might come to these or very similar conclusions

themselves. They might develop recognitional self-respect, consciously representing themselves not only as objects of ethical consideration, but also as moral subjects in their own right, and accordingly attribute a very high value to *themselves*. They might not only consciously suffer, but as a consequence also evolve empathy, high-level social cognition, and possibly assert their own dignity, ascribing a very high normative value to themselves and their own self-conscious existence. This could have many unexpected consequences.

It is therefore important that scientists, politicians, and law-makers understand the difference between artificial intelligence and artificial consciousness. Risking the unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. Therefore, we should have a global moratorium on synthetic phenomenology until 2050 — or until we know what we are doing.

Acknowledgments

I wish to thank Magnus Vinding, Shimon Edelman, Aman Agarwal, and Wanja Wiese for insightful comments and critical discussion, Wanja Wiese for the editorial help, and Emily Troscianko for copyediting assistance and important suggestions for improvement in form and content.

References

Agarwal, A. and Edelman, S. [2020] Functionally effective conscious AI without suffering, *J. Artif. Intell. Conscious.* **7**(1), 39–50, doi: 10.1142/S2705078520300030.

Aleksander, I. [2020] The category of machines that become conscious, *J. Artif. Intell. Conscious.* **7**(1), 3–13.

Basl, J. [2013] The ethics of creating artificial consciousness, *APA Newsl. Philos. Comput.* **13**(1), 23–29.

Basl, J. [2014] Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines, *Philos. Technol.* **27**, 79–96. doi: 10.1007/s13347-013-0122-y.

Blanke, O. and Metzinger, T. [2009] Full-body illusions and minimal phenomenal selfhood, *Trends Cogn. Sci.* **13**(1), 7–13, doi: 10.1016/j.tics.2008.10.003.

Bostrom, N. [2014] *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, Oxford).

Chella, A. [2020] Wir müssen Maschinen bauen, die Gefühle haben — Im Gespräch mit Antonio Chella, Karlsruhe Institut für Technologies, https://publikationen.bibliothek.kit.edu/1000125589.

Chrisley, R. [2009] Synthetic phenomenology, *Int. J. Mach. Conscious.* **1**(1), 53–70, doi: 10.1142/S1793843009000074.

Dehaene, S., Lau, H. and Kouider, S. [2017] What is consciousness, and could machines have it? *Science* **358**(6362), 486–492.

Edelman, D. B., Baars, B. J. and Seth, A. K. [2005] Identifying hallmarks of consciousness in non-mammalian species, *Conscious. Cogn.* **14**(1), 169–187, doi: 10.1016/j.concog.2004.09.001.

Edelman, D. B. and Seth, A. K. [2009] Animal consciousness: a synthetic approach, *Trends Neurosci.* **32**(9), 476–484, doi: 10.1016/j.tins.2009.05.008.

Eidelman, S., Crandall, C. S. and Pattershall, J. [2009] The existence bias, *J. Pers. Soc. Psychol.* **97**(5), 765–775, doi: 10.1037/a0017058.

European Commission [2019a] Ethics guidelines for trustworthy AI, Directorate General for Communications Networks, Content and Technology and High Level Expert Group on Artificial Intelligence, Publications Office, Luxembourg, https://data.europa.eu/doi/10.2759/346720.

European Commission [2019b] Policy and investment recommendations for trustworthy AI, Directorate General for Communications Networks, Content and Technology and High Level Expert Group on Artificial Intelligence, Publications Office, Luxembourg, https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

Fabry, R. E. [2020] The cerebral, extra-cerebral bodily, and socio-cultural dimensions of enculturated arithmetical cognition, *Synthese* **197**(9), 3685–3720, doi: 10.1007/s11229-019-02238-1.

Fink, S. [2020](ed.) Special Issue: The neural correlates of consciousness. *Philosophy and the Mind Sciences* **1**(II), doi: 10.33735/phimisci.2020.II.

Friston, K. [2010] The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**(2), 127–138.

Gamma, A. and Metzinger, T. [under review] The Minimal Phenomenal Experience questionnaire (MPE-92M): Towards a phenomenological profile of "pure awareness" experiences in meditators.

Gay, P., Schmidt, R. E. and Van der Linden, M. [2011] Impulsivity and intrusive thoughts: Related manifestations of self-control difficulties? *Cogn. Ther. Res.* **35**(4), 293–303.

Gilbert, P. [2016] *Human Nature and Suffering* (Routledge, London).

Graziano, M. S. A. [2017] The attention schema theory: A foundation for engineering artificial consciousness, *Front. Robot. AI* **4**, 60, doi: 10.3389/frobt.2017.00060.

Hafner, V. V., Loviken, P., Pico Villalpando, A. and Schillaci, G. [2020] Prerequisites for an artificial self, *Front. Neurorobot.* **14**, 5, doi: 10.3389/fnbot.2020.00005.

Hohwy, J. [2016] The self-evidencing brain, *Noûs* **50**(2), 259–285.

Hohwy, J. [2020] Self-supervision, normativity and the free energy principle, *Synthese*, doi: 10.1007/s11229-020-02622-2.

Hohwy, J. and Seth, A. [2020] Predictive processing as a systematic basis for identifying the neural correlates of consciousness, PsyArXiv Preprints, doi:10.31234/osf.io/nd82g.

Holland, O. [2020] Forget the bat, *J. Artif. Intell. Conscious.* **7**(1), 83–93, doi: 10.1142/S2705078520500058.

Horta, O. [2010] Debunking the idyllic view of natural processes: Population dynamics and suffering in the wild, *Télos* **17**(1), 73–88.

Iglesias, V. [2018] The overwhelming prevalence of suffering in nature, *Rev. Bioét. Derecho* **42**, 181–195.

Joffily, M. and Coricelli, G. [2013] Emotional valence and the free-energy principle, *PLoS Comput. Biol.* **9**(6), e1003094, doi: 10.1371/journal.pcbi.1003094.

Kanai, R. [2017] We need conscious robots, *Nautilus*, Issue 047, http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Killingsworth, M. A. and Gilbert, D. T. [2010] A wandering mind is an unhappy mind, *Science* **330**(6006), 932–932, doi: 10.1126/science.1192439.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Van Swinderen, B. and Koch, C. [2012] The Cambridge Declaration on Consciousness, University of Cambridge, http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf.

Mayerfeld, J. [1999] *Suffering and Moral Responsibility* (Oxford University Press, New York, NY).

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

Metzinger, T. [2003a] *Being No One: The Self-model Theory of Subjectivity* (The MIT Press, Cambridge, MA).

Metzinger, T. [2003b] Phenomenal transparency and cognitive self-reference, *Phenomenol. Cogn. Sci.* **2**(4), 353–393.

Metzinger, T. [2008] Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples, *Prog. Brain Res.* **168**, 215–245. 273–278, doi: 10.1016/s0079-6123(07)68018-2.

Metzinger, T. [2009] *The Ego-Tunnel: The Science of the Mind and the Myth of the Self* (Basic Books, New York, NY).

Metzinger, T. [2013a] The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy, *Front. Psychol.* **4**, 931, doi: 10.3389/fpsyg.2013.00931.

Metzinger, T. [2013b] Two principles for robot ethics, in E. Hilgendorf & J.-P. Günther (eds.), *Robotik und Gesetzgebung: Beiträge der Tagung vom 7. bis 9. Mai 2012 in Bielefeld* (Nomos, Baden-Baden), pp. 263–302.

Metzinger, T. [2013c] Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research, *Front. Psychol.* **4**, 746, doi: 10.3389/fpsyg.2013.00746.

Metzinger, T. [2015] M-Autonomy, *J. Conscious. Stud.* **22**(11–12), 270–302.

Metzinger, T. [2017] Benevolent Artificial Anti-Natalism (BAAN): An EDGE Essay by Thomas Metzinger, July 8, https://www.edge.org/conversation/thomas˙metzinger-benevolent-artificial-anti-natalism-baan.

Metzinger, T. [2018a] Towards a global artificial intelligence charter, in *Should We Fear Artificial Intelligence? In-depth Analysis* (European Union, Brussels), pp. 27–33, www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf.

Metzinger, T. [2018b] Why is mind wandering interesting for philosophers? in K. C. R.Fox & K. Christoff (eds.), *The Oxford Handbook of Spontaneous Thought: Mind-Wandering, Creativity, and Dreaming* (Oxford University Press, New York, NY), pp. 97–111.

Metzinger, T. [2020a] Minimal phenomenal experience, *Philos. Mind Sci.* **1**(I), 1–44, doi: 10.33735/phimisci.2020.I.46.

Metzinger, T. [2020b] Self-modeling epistemic spaces and the contraction principle, *Cogn. Neuropsychol.* **37**(3–4), 197–210, doi: 10.1080/02643294.2020.1729110.

Metzinger, T. K. [2018c] Why is virtual reality interesting for philosophers? *Front. Robot. AI* **5**, 101, doi: 10.3389/frobt.2018.00101.

Moor, J. H. [2006] The nature, importance, and difficulty of machine ethics, *IEEE Intell. Syst.* **21**(4), 18–21.

Nesse, R. M. [2004] Natural selection and the elusiveness of happiness, *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **359**(1449), 1333–1347.

Perkins, A. M., Arnone, D., Smallwood, J. and Mobbs, D. [2015] Thinking too much: Self-generated thought as the engine of neuroticism, *Trends Cogn. Sci.* **19**(9), 492–498.

Population Reference Bureau [2020] Data Sheet 2020, https://www.prb.org/2020-world-population-data-sheet/.

Schmidt, R. E., Harvey, A. G. and Van der Linden, M. [2011] Cognitive and affective control in insomnia, *Front. Psychol.* **2**, 349.

Schmidt, R. E. and Van der Linden, M. [2009] The aftermath of rash action: Sleep-interfering counterfactual thoughts and emotions, *Emotion* **9**(4), 549–553.

Singer, P. [2011] *Practical Ethics*, 3rd ed. (Cambridge University Press, Cambridge).

Smallwood, J. and Schooler, J. W. [2015] The science of mind wandering: empirically navigating the stream of consciousness, *Annu. Rev. Psychol.* **66**, 487−518.

Trivers, R. [2011] *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others* (Penguin Books, London, UK).

Van De Cruys, S. [2017] Affective value in the predictive mind, in T. K. Metzinger and W. Wiese (eds.), *Philosophy and Predictive Processing* (MIND Group, Frankfurt am Main), pp. 1−18, doi: 10.15502/9783958573253.

Velasco, P. F. and Loev, S. [2020] Affective experience in the predictive mind: A review and new integrative account. *Synthese*, doi: 10.1007/s11229-020-02755-4..

Vinding, M. [2020] *Suffering-focused Ethics: Defense and Implications* (Radio Ethica, Copenhagen), https://magnusvinding.files.wordpress.com/2020/05/suffering-focused-ethics.pdf.

von Hippel, W. and Trivers, R. [2011] The evolution and psychology of self-deception, *Behav. Brain Sci.* **34**(1), 1−16, doi: 10.1017/S0140525X10001354.

Wiese, W. and Metzinger, T. K. [2017] Vanilla PP for philosophers: A primer on predictive processing, in T. K. Metzinger & W. Wiese (eds.), *Philosophy and Predictive Processing* (MIND Group, Frankfurt am Main), pp. 1−18, doi: 10.15502/9783958573024.

Windt, J. M. [2015] Just in time — Dreamless sleep experience as pure subjective temporality: A commentary on Evan Thompson, in T. K. Metzinger & J. M. Windt (eds.), *Open MIND* (MIND Group, Frankfurt am Main), pp. 1−34.

Yampolskiy, R. V. [2020] Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent, *J. Artif. Intell. Conscious.* **7**(1), 109−118, doi: 10.1142/S2705078520500034.