Two Principles for Robot Ethics[1]

Thomas Metzinger, University of Mainz

A. Introduction

This contribution has two parts. In the first part I will formulate two new principles for the applied ethics of advanced robotic systems, namely the principle of *negative synthetic phenomenology* (NSP) and the principle of *veto autonomy* (VA). The second part will further clarify and substantiate some of the technical concepts and theoretical background assumptions, which are necessary to formulate these principles. In particular, I will make an attempt to produce a concise list of desiderata for future research.

Obviously, my goal in this chapter is not to present a full-blown theory of machine consciousness, or a conceptual model for distributed volitional control in functionally coupled man-machine systems, or even a fully developed ethical argument to support my positive normative claims. The epistemic goal simply consists in isolating two major theoretical issues more clearly. I think that not only have these issues been ignored for too long, but also that they possess great relevance for politicians, legal theorists and philosophical ethicists – as well as for empirical researchers and engineers. The first of these two issues is the problem of artificial suffering: How do we avoid the creation or an unexpected emergence of conscious suffering in intelligent postbiotic systems, for example in advanced, autonomous robots?

The second relevant question arises in the context of new technologies like brain-machine interfaces (BCIs), virtual reality (VR) and teleoperator systems (TOSs). One can rationally expect not only rapid confluence between these technologies, but also considerable progress in neuroscience in the foreseeable future. Given that the human brain can now increasingly be embedded in an ever more fine-grained causal network of artificial sensor/effector systems, how do we carefully recalibrate and redefine our traditional notions of "legal culpability", "ethical responsibility", or "accountability for one's own actions"? Put dif-

1 I am greatly indebted to Michael Madary for a number of very helpful comments and his support with the English version of this paper. Lisa Blechschmitt and Jan-Philipp Günther have helped me solving editorial problems, and Jan Christoph Bublitz has offered a considerable amount of very stimulating critical ideas. I also wish to thank Patrick Haggard for his comments and a pointer to relevant empirical literature.

ferently: How can one achieve semantic continuity in the face of historically new classes of potential actions and a considerable shift in the general image of man? The central goal of this chapter is quite modest: All I want to do is to lay some very first conceptual foundations and generate two starting points for systematic academic discussions. However, I will try to achieve this goal by actually arguing for two positive claims in an attempt to provoke my readers in productive manner.

B. Part One

I. NSP: The principle of avoiding artificial suffering

The better our scientific understanding of the functional deep structure of the human mind becomes, the more of our own mental properties can in principle be instantiated on non-biological carrier systems. As philosophers say, functional properties are "multi-realizable"; the same property can be realized on different types of hardware, as long as the physical states on which it is implemented possess the necessary causal powers.[2] Arguably, intentional (i.e., semantic) mental properties like the having of "content" or "reference" can be realized by autonomous, embodied agents, i.e., they can be gradually acquired via an intelligent form of dynamically interacting with the world and other agents.[3] However, even if we accept this assumption, this would only result in artificial *intelligence*. What about artificial *consciousness*? For an artificial agent to possess conscious experience would mean for it to instantiate *phenomenal* mental properties. Phenomenal properties determine how the world appears to you from the first-person perspective (1PP), how you subjectively experience the colors, the sounds or smells surrounding you, but also how you experience different states of your own body, your emotions, and even your conscious thought processes. A typical assumption therefore is that a conscious machine would also have a 1PP; it would have a form of self-consciousness plus its own subjective point of view. For a machine or an autonomous robotic agent to be conscious then would mean

2 See *Putnam*, Mind, Language and Reality, Philosophical Papers, Vol. 2, 1975; for a general introduction and further references cf. *Metzinger*, Grundkurs Philosophie des Geistes, 2007, Band II, Modul L-11, pp. 367ff.

3 See *Harnad*, The symbol grounding problem, Physica, D 42, 1990, pp. 335–346; *Steels*, The symbol grounding problem has been solved. So what's next?, in: *de Vega* (ed.), Symbols and Embodiment: Debates on Meaning and Cognition, 2008; for a general introduction and further references cf. *Metzinger*, Grundkurs Philosophie des Geistes, 2010, Band III, 15, Abschnitt 3.2., pp. 22ff., Module I-1, I-8 and I-15.

that, for example, internal representations of wavelength mixtures or surface reflectance properties of visually perceived objects in its environment also *appear* to it as "redness" or "blueness" and that they do so under a 1PP, as *subjective* states, states bound to an inner representation of a conscious self currently *having* them. "Saltiness" or "sweetness", "warmth" or "cold", the smell of vanilla or the subjective sound quality of listening to a note played on a cello are other examples of such phenomenal properties. "Pain" is a phenomenal property too and the concept of "suffering" refers to a related, but more complex class of phenomenal states.

The principle of *negative synthetic phenomenology* (NSP) states an ethical norm, which demands that, in artificial systems, we should not aim at the creation or even risk the unexpected emergence of conscious states falling into the phenomenological category of "suffering":

> (NSP): We should not deliberately create or even risk the emergence of conscious suffering in artificial or postbiotic agents, unless we have good reasons to do so.

I will now add some short explanatory remarks and sketch a brief argument for accepting NSP as a positive ethical and legal norm. But first, let us get an *intuitive* grasp of the problem by looking at a short thought experiment.[4] Imagine that you are a member of an ethics committee looking at scientific grant applications. One of them says:

> We want to use gene technology to breed cognitively disabled human infants. For urgent scientific reasons, we need to generate human babies possessing certain cognitive, emotional, and perceptual deficits. This is an important and innovative research strategy, and it requires the controlled and reproducible investigation of the disabled babies' psychological development after birth. This is not only important for understanding how our own minds work but it also has great potential for healing psychiatric diseases. Therefore, we urgently need comprehensive funding.

No doubt you'll decide immediately that this idea is not only absurd and tasteless but also dangerous. We all hope that a proposal of this kind would not pass any ethics committee in the democratic world.

The first aspect to note in our introductory thought experiment is that it seems to aim at a specific subset of possible persons, at beings that do not yet exist but that *could* exist. More precisely, the domain of inquiry is constituted by *possible, artificial subjects of experience*. We cannot call these artificial (or postbiotic[5])

4 Adapted from *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009.

5 The term "postbiotic" tries to answer a minimal logical difficulty, namely the fact that, under closer scrutiny in the real world, our intuitive conceptual distinction between "natural" or "artificial" systems fails, because is not an exclusive and exhaustive one. The ethically and legally relevant class of systems comprises cases which are neither exclusively biological nor exclusively artificial. Today we already have intelligent systems that

systems "unborn" conscious beings, because their way of coming into existence is not by another biological organism giving birth to them, but rather by either (a) being constructed by such biological organisms (namely, ourselves), or (b; and much more likely) by emerging out of a complex process of dynamical self-organization and/or quasi-evolutionary "bootstrapping", which was initiated by their biological predecessors. It is also unclear if and in what sense we would call these systems "persons", because our own theories about what constitutes a person, what the conditions of personhood are, etc., undergo constant historical change and certainly are not commonly shared in all human cultures. What really counts is that they are "subjects of experience", which means that they are conscious, self-conscious and possess a 1PP. They have *subjective states*, the world appears to them in some way, just like it appears to us, and they have phenomenal qualities too - although they may be very different from the ones we humans know by direct acquaintance. Maybe *what it is like* to be a robot does not include "saltiness" or "sweetness", "warmth" or "cold", but something entirely different. Central to our thought experiment is that they would also be able to suffer. This means that not only phenomenal properties equivalent to "pain" could be instantiated by them, but more importantly that they would possess subjective preferences, that these preferences could be frustrated, and that this fact could be explicitly represented on the level of conscious experience.

My point is that conscious suffering is the relevant criterion of demarcation. (Because the concept of "suffering" is so important for the issue at stake I will say more about it in Part Two.) For now, let me illustrate the relevance of conscious suffering as the central criterion by a quote taken from Peter Singer:

> 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 experi-

use artificial control structures implemented through a fully biological substrate (e.g., in hybrid bio-robotics) – that is, human-created "software" running on naturally evolved "hardware" if you will. On the other hand, we also find abstract biological principles determining the causal structure of man-made artifacts, for example in artificial neural networks or evolutionary robotics (for concrete examples, see *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009, chapter 7). To remain in the context of this contribution, the creation of a phenomenal self-model (PSM) and the subsequent "motivation" or "driving" of an autonomous system via the process of conscious suffering exactly is an example of such higher-order biological principles at work (see section C.I). The evolution of tool-use by functionally integrating and thereby "transiently embodying" an artifact, like it is exemplified in the case study for robotic re-embodiment presented in sections B.II.1 and B.II.2, is a second example of a naturally evolved neurocomputational principle instantiated in a coupled man-machine system. "Postbiotic" systems, then, are systems for which the distinction between "artificial" and "natural" makes little sense.

> encing 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?[6]

The second assumption underlying our thought experiment is that we, as human beings, are ethically responsible for these possible, postbiotic subjects of experience coming into existence. My first, very general point about this logical scenario is to finally draw attention to the fact that possible, postbiotic or artificial subjects of experience currently possess no representatives in any ethics committee, but also not on the level of any legal or political institution in human societies. Today we may discuss the preferences or the potential quality of life of unborn human beings, i.e., of possible persons of one single, very specific biological kind, but we are blind to the fact that conscious experience and subjective preferences are not tied to biological hardware by any sort of conceptual necessity. As soon as we understand this point, the domain of objects for ethical and legal consideration widens.

Third, our introductory thought experiment contains an empirical premise. Empirical premises can be false, and they can be *made* false. Here, the empirical premise comes as a prediction: The first machines satisfying a minimally sufficient set of conditions for conscious experience and selfhood would find themselves in a situation similar to that of the genetically engineered disabled human infants. Like them, these machines would have all kinds of functional and representational deficits—various disabilities resulting from errors in human engineering. Their perceptual systems—their artificial eyes, ears, and so on—would not work well in the early stages. They would likely be half-deaf, half-blind, and have all kinds of difficulties in perceiving the world and themselves in it. Obviously, they would also suffer from motoric and behavioral deficits – what is true of sensors would very likely be true of effectors as well. Sensorimotor integration and planning almost inevitably need an internal body-model which can also be taken offline. This would lay the foundation for the ineluctable phenomenology of *ownership*: If they had a stable bodily self-model, they would be able to feel sensory pain as their *own* pain, as located in their body image. If their postbiotic self-model was directly anchored in the low-level, self-regulatory mechanisms of their hardware—just as our own emotional self-model is anchored in the upper brainstem and the hypothalamus[7]—they would be consciously *feeling* selves. They could have a functional equivalent of human emotions that present

6 *Singer*, Practical Ethics, 2011, p. 50.

7 *Damasio*, The Feeling of What Happens: Body and Emotion in the Making of Consciousness, 1999; *Parvizi/Damasio*, Consciousness and the brainstem, Cognition, Vol. 79, 2001, pp. 135-159.

them with valences in an "interoceptive" or "somaesthetic" data-format, inner states representing their preferences to them in a non-propositional form, directly and untranscendably integrated into their body image. They would then consciously *own* these preferences. They would experience a loss of homeostatic control as painful, because they had an inbuilt *concern* about their own existence. They would have preferences and interests of their own and they would subjectively experience this fact. They might suffer emotionally in degrees of intensity or in qualitative ways completely alien to us that we, their creators, could not even imagine. The empirical prediction says that the first generations of such machines would very likely have many negative emotions, reflecting their failures in successful self-regulation, simply because of all kinds of hardware deficits and higher-level disturbances. These negative emotions could be conscious and intensely felt, but in many cases we might not be able to understand or even recognize them. In the domain of machine conscious our ignorance is high, because there might be observationally indistinguishable systems with and without phenomenal states, and the consequences of our own actions or choices with regard to these systems are not easily predictable. We have to integrate the problem of epistemic indeterminacy into our ethical solution: There is a risk to be minimized, namely the possibility that non-biological subjects of experience have already begun to suffer before we as their human creators have even become aware of this fact.

We could also take the thought experiment further, developing a scenario that currently seems extremely implausible under any empirical perspective. Imagine our postbiotic subjects of experience as possessing a *cognitive* self-model—as being well-informed, intelligent thinkers of thoughts. It is conceivable that they had not only negative hedonic sensations, but also informed, rational preferences. They could then not only conceptually grasp the bizarreness of their existence as mere objects of scientific interest but also could intellectually suffer from knowing that, as such, they lacked the innate "dignity" that seemed so important to their creators. They 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 would understand that we *knew in advance* that they would have a large number of uncompensatable and frustrated preferences, but that we did not possess the benevolence to avoid the emergence of 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 postbiotic selves 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 only a commodity?

Fourth, there is a normative principle underlying NSP, and I promise to say a little more about in Part Two, under the heading of "moderate negative utilitarianism". Put very simply, the idea is that we should always strive to minimize the overall amount of suffering in the universe, and that, unless we have very good reasons to do so, we should refrain from all actions that could increase the overall amount of suffering in the universe. At this point, it is important to note that I do *not* want to develop or defend a specific metaethical position in this chapter (for example, a refined version of negative utilitarianism for the domain of conscious machines). All I am looking for in my attempt to start a discussion on the issue of non-biological suffering is a very general and as-innocent-as-possible criterion, a practical principle that as many of my readers as possible can agree upon because, at least *prima facie*, it reflects part of their own intuitions. Our theoretical intuitions about what a *positive* state of affairs, the best state of the world, or an optimal state of conscious experience, actually is may widely diverge. When it comes to phenomenal states that are subjectively experienced as having a *negative* valence, however, it is much easier for us to reach a workable consensus. Don't we all share the ethical intuition that unnecessary suffering should not be caused or created, and that wherever it already takes place we should continuously strive to minimize the frustration of preferences, always alleviating suffering wherever possible? Is it not true that we find it much easier to agree on what a negative state of consciousness is, namely, any state that the system in question would prefer *not* to live through, a state it would rather *not* experience? Don't we all believe that at least some kinds of suffering simply cannot be compensated? And is it not true that we all share the practical intuition that the avoidance of conscious suffering is not only more urgent, but simply easier to achieve than the creation of happiness – that in this world it is just so much more *efficient* to eliminate potential causes of suffering than to take care of generating happiness?

One final introductory remark is in order, namely about the historical context in which the principle of *negative synthetic phenomenology* (NSP) is put forward. As Dieter Birnbacher has pointed out, the field of robot ethics may share a number of relevant structural features with the field of animal ethics. Our current treatment of animals is clearly untenable from an ethical perspective, often inconsistent and highly hypocritical. But even though our treatment of animals is characterized by a fundamental attitude of disrespect and a considerable lack of benevolence, we have at least arrived at a notion of "animal ethics". In animal ethics, when discussing animal protection laws or issues of animal welfare, we at least *discuss* the potential suffering of future animals and try to weigh their frustrated desires, the pain and suffering we cause against human interests, aggregating preferences across species boundaries, etc. For artificial systems we do not yet do this. This is another reason why we need an applied ethics for all scientific

attempts to create artificial conscious experience, to deliberately synthesize or cause the self-organization of phenomenal states.

"Synthetic phenomenology" (SP) is a concept first introduced by the American philosopher J. Scott Jordan in 1998[8] 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, their states and functions, on artificial hardware. SP encompasses a variety of different approaches, methodologies, and disciplines, but what they all have in common is that they see SP as the construction or guided dynamical self-organization of phenomenal states in artificial systems plus the deep seated methodological intuition that any scientific explanation of consciousness necessarily involves a systematic *reconstruction* of the target phenomenon. But we need more than an applied ethics for SP. Given our specific historical situation and the normative principle of NSP, it follows that the interests of possible future subjects of experience capable of suffering - just like the interests of *any* type of system able to consciously experience negative hedonic states and a frustration of preferences - must be systematically represented in legal and political institutions.

II. VA: The principle of veto-autonomy

If a human being is causally coupled with an artificial or postbiotic system via the PSM in its biological brain in a technologically novel, and causally more direct manner, and if the artificial system causes harm or physical damage - when exactly should we say that the human agent is *ethically responsible*, or *culpable* in a legal sense? What are rational and empirically grounded criteria for the functional boundaries of autonomous agency, helping us to decide if a given human subject was accountable for their own actions? What *ability* does our human agent have to possess in order to count as responsible for the results of her actions?

8 See *Gamez*, Progress in machine consciousness, Consciousness and Cognition, Vol. 17 (3), 2008, pp. 887-910; *Holland/Goodman*, Robots with internal models: A route to machine consciousness?, in *Holland* (ed.), Machine Consciousness, 2003; *Holland/Knight/Newcombe*, A robot-based approach to machine consciousness, in: *Chella/Manzotti* (eds.), Artificial Consciousness, 2007, pp. 887-910; *Chrisley/Parthemore*, Synthetic Phenomenology: Exploiting Embodiment to Specify the Non-Conceptual Content of Visual Experience, Journal of Consciousness Studies, Vol. 14 (7), 2007, pp. 44-58; *Aleksander*, Machine consciousness. Scholarpedia, 2008 3(2):4162, for a fist overview.

This issue arises in the context of new technologies like brain-machine interfaces (BCIs), virtual reality (VR) and robotic teleoperator systems (TOSs). Here, the first point I want to draw attention to is that although quite often bodily agency (like moving a joystick, or acting through a motion-tracking system coupled to an avatar or physical robot) will still play a role, human agents will increasingly control technical devices via mental self-simulations in the future. A mental self-simulation can be described[9] as a process of *inner agency*, in which an agent uses her phenomenal self-model to create certain causal effects in the world in the absence of overt bodily behavior, for example by imagining to lift her right arm, envisioning herself as flying or virtually directing her gaze into a certain direction. As the human PSM is physically realized by a widely distributed pattern of neural activation in the biological brain, this activity can, via a suitable causal interface, be read out or directly coupled to any artificial effector (be it virtual or physical). The self-model theory of subjectivity[10] predicts exactly this possibility for extended man-machine systems. Let us now introduce a new technical term and call any such action a "PSM-action": A PSM-action is any action in which a process of inner, mental agency plays either the sole or at least the *central* causal role in creating an effect in the world, bypassing biological effectors and directly controlling artificial devices like avatars, robots, or other advanced systems of teleoperation. For PSM-actions the non-neural body plays practically no role in *implementing* an action goal, because a human agent uses certain layers of her conscious self-model for a deliberate self-simulation, knowing that - although she is "biologically offline" – this mental action will likely have an effect, which is causally mediated by her technological environment.[11] Our question now becomes, when was an agent responsible for a PSM-action?

9 It is important to note how most conscious, mental self-simulations clearly are *not* actions at all, but non-agentive, subpersonal processes. The ubiquitous phenomenon of spontaneous mind-wandering would be a standard example (see *Schooler/Salau/Julien/Ives*, Alternative stable states explain unpredictable biological control of Salvinia molesta in Kakadu, Nature, Vol. 470, 2011, pp. 86–89, for an overview and further references). According to the conceptual distinction between "mental behavior" and "mental action" introduced in Part Two, section 2. b. most of our ongoing mental activity is best described as a non-intentional form of mental behavior.

10 *Metzinger*, Being No One. The Self-Model Theory of Subjectivity, 2004; *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009.

11 A second concrete example of what I call a "PSM-action" cf. the successful detection of awareness in patients in the vegetative state, as conducted in a classical study by Adrian Owen and colleagues: *Owen/Coleman/Boly/Davis/Laureys/Pickard*, Detecting awareness in the vegetative state, Science, Vol. 313 (5792), 2006, p. 1402. During this fMRI study the patient was given spoken instructions to perform two mental imagery tasks at specific points during the scan. One task involved imagining playing a game of tennis and the other involved imagining visiting all of the rooms of her house, starting from the front door. As her neural responses were indistinguishable from those observed in healthy vol-

The principle of veto-autonomy states that this is exactly the case if the agent has a specific *ability*, namely the ability to consciously "veto" or interrupt a PSM-action by a second-order form of agency, either mental or bodily:

> (VA) An agent is responsible for a PSM-action if, at the time of the action, she possessed the ability to suspend or terminate this action.

Let us say that autonomy is rational self-control. Veto-autonomy is one specific aspect of self-control; it is the functional ability to suppress an action via a process of inhibiting or down-regulating a given urge to act, of stopping an ongoing mental action simulation, or of terminating a motor command that had already been issued. A human agent directly coupled to an artificial system via her PSM and not possessing this ability could conceivably cause major damage by simply *thinking* about a certain action and thereby causing the robot to carry it out, without being able to block its consequences. My first point is that by causally embedding the human brain into new types of technological and virtual environments, the distinction between volition, motor imagery, and overt action becomes blurred in a theoretically interesting way. A new type of problem arises: Are we responsible for the consequences of unintentional PSM-actions?

I have proposed to begin by describing VA in a simple and traditional manner, namely as a personal-level *ability*.[12] The idea then is that this ability gives human beings a specific form of autonomy, because it permits a form of "second-order agency", namely actions directed at other actions. Conceptually, it can now be claimed that having the PSM of human beings in ordinary, non-pathological waking states is a necessary functional condition for this specific personal-level ability of VA and often it will also be a sufficient condition.[13] However, in some technological or virtual environments – those enabling direct PSM-actions - it might frequently *not* be a sufficient condition any more. It is now conceivable that simply thinking about an action might cause a direct effect in the world, but

unteers performing the same imagery tasks in the scanner it was possible to demonstrate how, despite fulfilling the clinical criteria for a diagnosis of vegetative state, this patient retained the ability to understand spoken commands and to respond to them through her brain activity, rather than through actions carried out via the non-neural body, like overt speech or movement. In addition, her decision to cooperate by imagining particular tasks when asked to do so represents a clear act of intention, which seems to confirm not only possessed a phenomenal self-model but also conscious awareness of her surroundings.

12 Of course, there is a major question concerning the compatibility of VA with physical determinism; another important theoretical issue in the background is the adequacy and autonomy of the personal level of description. Simply speaking of an "ability" a person may have or not have could turn out to be much too simplistic, and actually veil the deeper challenge posed. Finding answers to these questions are obvious desiderata for future research, I will therefore briefly come back to them at the very end of Part Two.

13 Empirical counter examples for non-sufficiency are: ego depletion, addiction, impulse-control disorders, anarchic hand syndrome, etc.

that this effect cannot be prevented by the agent herself, because in certain technological environments her naturally evolved capacities for second-order mental action are not sufficient to block or causally neutralize it. As a normative claim, we could now say that, for any human person possessing veto autonomy and any teleoperated, virtually embodied robotic action, if at the time of the action the agent's phenomenal self-model was functionally integrated with a robot (or avatar) in a way that enables PSM-actions *and* gives the human agent VA, then this person is legally and ethically responsible for their consequences. If the agent did not possess mental self-control in the sense of VA, the she was not responsible.[14] Obviously, the relevant neurobiological data and especially the philosophical implications of the proposed working concept of "veto autonomy" are so extremely rich that I cannot even begin to discuss them in this contribution. However, I will formulate at least some desiderata for a comprehensive theory in Part Two. For now, I will try to make our discussion more concrete by offering readers an empirical example in order to set a more detailed context.

1. PSM-actions and robotic re-embodiment

This time I will not use a thought experiment for purposes of introductory illustration, but an empirical proof-of-concept study[15] from VERE[16], an international research project of which I am a member. In an ambitious pilot study for fMRI-based robotic embodiment Ori Cohen, Doron Friedman and their colleagues presented a proof-of-concept for the notion of a "PSM-action" introduced above, based on real-time functional Magnetic Resonance Imaging (rtfMRI). This may actually be the first time fMRI is used as an input device to identify a subject's intentions and convert them into actions performed by a humanoid robot. The process, based on motor imagery, has allowed subjects located in Israel to control a HOAP3 humanoid robot in France, experiencing the whole experiment through the eyes of the robot.

14 Please note how VA describes a kind of control that is not necessarily "sensitive to reasons", as the realization of this capacity could also be driven by some spontaneous emotional reaction or implicit context-information, like a given set of ethical intuitions. Incorporating some element of reason-responsiveness to VA would already make it a much stronger concept of autonomy. I am here concerned with the minimal requirements for liability, accountability, etc.

15 *Cohen/Druon/Lengagne/Mendelsohn/Malach/Kheddar/Friedman*, MRI-based robotic embodiment: A pilot study, IEEE International Conference on Biomedical Robotics and Biomechatronics, 2012.

16 See <http://www.vereproject.eu> for details (accessed 23 October 2012).

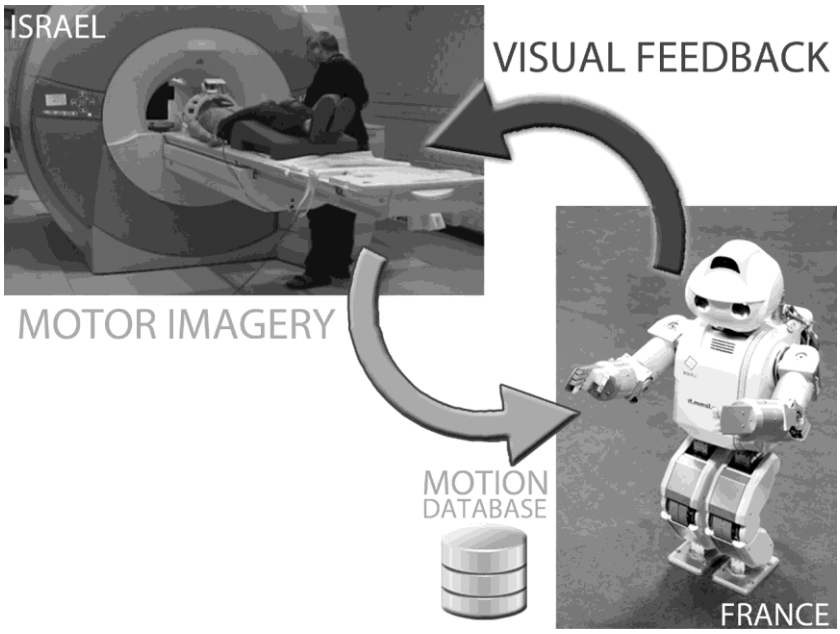

Figure 1. Acting directly with the PSM, via robotic embodiment: General principle of data processing and experiment related tasks. The goal was to provide a subject located in Israel with the direct, thought-based control of a robotic avatar in France. For a video demonstration, see <http://www.youtube.com/user/TheAVL2011> (accessed 23 October 2012). Figure courtesy of Doron Friedman.

The authors describe the aims of the VERE project as "dissolving the boundary between the human body and surrogate representations in immersive virtual reality and physical reality."[17] Phenomenologically, this means that the subject or the operator "is expected to have the illusion that his surrogate representation is his own body, and behave and think accordingly."[18] For example, reading out intentions by translating a human being's motor imagery into high-level commands for an avatar or physical robot may help disabled humans to control artificial bodies or "virtual prostheses", but obviously there could be military or vari-

17 *Cohen/Druon/Lengagne/Mendelsohn/Malach/Kheddar/Friedman*, MRI-based robotic embodiment: A pilot study, IEEE International Conference on Biomedical Robotics and Biomechatronics, 2012.

18 *Cohen/Druon/Lengagne/Mendelsohn/Malach/Kheddar/Friedman*, MRI-based robotic embodiment: A pilot study, IEEE International Conference on Biomedical Robotics and Biomechatronics, 2012.

ous illegal applications for controlling an external device just by thinking, without any bodily movement being involved. There are still many technical limitations, but the general direction is clear: In principle, the human PSM can be directly coupled to remote technical systems, such as teleoperated robots or virtual bodies in virtual worlds. By consciously imagining a certain body movement, for example, such extracorporeal devices can be controlled – but it is not at all clear if they can ever be controlled as *flexibly* as the biological body of an agent.

My second point is that the degree of functional *context-sensitivity* (e.g., relative to distal social norms or individual long-term goals of the agent) could be considerably lower, as well as the accompanying levels of self-control and mental self-determination, thereby relevantly and immediately affecting the agent's resulting overall autonomy. What counts is what I would like to term the "functional depth of embodiment", the dimensionality of the control structure which is enabled by such advanced man-machine systems. This depth of embodiment results from the available number of layers of self-control in combination with the overall number of distal goal-representations possessed by the agent[19]; in principle it can be quantified and measured, and it must be directly related to ethico-legal concepts like "accountability" or "responsibility". Autonomy is not an irreducible qualitative concept. In order to take the ethical challenge seriously autonomy has to be decomposed into a set of quantifiable abilities and low-level functional dispositions. Autonomy also has a variable phenomenological profile; functionally as well as on the level of subjective experience it clearly is something that comes in degrees. At some stage of technological development it may well be possible to generate the phenomenology of identification and full embodiment[20], but the degree and type of autonomy which is *actually* achieved on the

19 Of course, "depth of embodiment" could also mean, for example, the robustness of the purely *phenomenological* sense of identification with a given sensor-effector system, or the ability of an agent to successfully control *overt* bodily actions. Here, I am interested in mental autonomy as a function of the purely internal complexity emerging out of distal goal representations and the capacity for self-control via 2nd-order mental action, because I believe exactly this point is of maximal relevance for legal and ethical issues. For example, the patient documented by Adrian Owen (see footnote 11) could plausibly have developed a weaker phenomenal sense of identification with here immovable physical body and obviously has an extremely shallow embodiment in terms of her ability to control overt actions, but might have retained a considerable depth of embodiment in the sense here intended, namely by preserving her capacities for mental action and self-control.

20 See *Blanke/Metzinger*, Full-body illusions and minimal phenomenal selfhood, Trends in Cognitive Sciences, Vol. 13 (1), 2009, pp. 7-13, for a discussion of the relevant mechanisms and the concept of "minimal phenomenal selfhood"; *Blanke*, Multisensory brain mechanisms of bodily self-consciousness, Nature Reviews Neuroscience, Vol. 13, 2012, pp. 556–571, for an extensive review of the empirical literature.

functional level may still be limited in various ways. This is a problem that has to be solved by robot ethics.

2. Anarchic robot syndrome

Imagine you are lying in a scanner, controlling a robot at a distance, seeing through its eyes and even feeling the motor feedback from its arms and legs as they move. Phenomenologically, you fully identify with the robot (enjoying what has been called "3rd-order embodiment"[21]) as you move around in a situation involving other human agents. Suddenly you see your ex-wife's new husband entering the room, the person who has just ruined your personal life a couple of months ago. You feel a deep sense of emotional hurt and have an automatically arising aggressive impulse, a short violent fantasy unfolds in your mind, you try to calm yourself down – but before you can even inhibit the motor imagery emerging along with your violent fantasy the robot has already killed the man with one single strike of enormous force. Then you regain control and manage to step back. Subjectively, it feels as if you never had a chance to control your behavior. But how exactly will we decide if, objectively, you did actually *have* the ability for the necessary form of second-order mental action?

There are two ways in which we can imagine veto-autonomy (VA) during the situation described above. First, you could perhaps have terminated or blocked the PSM-controlled robotic action by some sort of overt *bodily* action: Perhaps there is a big red "STOP"-button inside the scanner and next to your hand, which you could have just hit with your biological arm. Second, you may have possessed VA in terms of an ability for second-order *mental* action, only controlling a complex activity pattern in your biological brain, generating and issuing a veto-command to "run behind and catch" the first-order mental action simulation in order to neutralize it. My positive proposal is that you should be considered ethically or legally responsible only in those cases where you had VA in this sense. This would be true for both scenarios, but the focus lies on the second type of scenario, because the interesting, and potentially novel, problem is generated by what I have called "PSM-actions" in robotic re-embodiment. The positive proposal tries to formulate the minimal degree of mental self-determination that is necessary to ascribe responsibility, liability, etc. The *functional time-window* en-

[21] See *Metzinger*, Reply to Gallagher: Different conceptions of embodiment, PSYCHE – An Interdisciplinary Journal of Research on Consciousness, Vol. 12 (4), 2006; and *Metzinger*, First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood, in: *Shapiro* (ed.), The Routledge Handbook of Embodied Cognition, 2013.

abling second-order mental self-determination in PSM-actions must have been demonstrably large enough for the person to successfully terminate a PSM-action; without this possibility of termination we may want to say that it was not a personal-level event in the relevant sense.

Once again, the purpose of this formulation is not to present a full-blown theory of ethical autonomy or legal responsibility, but to generate a starting point for systematic theoretical discussions. Clearly, a convincing solution would have to satisfy a larger number of empirical "bottom-up" constraints concerning the neuroscience of volition, as well as offering a coherent conceptual model for philosophical issues like "freedom of the will", "autonomy", the semantic distinction between actions, behaviors and events, or that between "personal" and "subpersonal" properties of an agent, etc. I will list some desiderata in Part Two of this chapter (cf. section III).

For now it is interesting to note how the scenario described above resembles impulse-control disorders like addiction, ADHD, obsessive-compulsive disorder, or Anarchic Hand Syndrome.[22] What all these real-world cases have in common is a specific lack of self-control, or what I would like to call *shallow embodiment*. The patient with Anarchic Hand Syndrome may typically keep losing control of her left hand—it is acting on its own (like the robot in our thought experiment above). At night, she may awake several times because her left hand was trying to choke her and she had to use your right hand to fight it off (just like the imaginary subject in the scanner, physically hitting the STOP-button, she has to resort to bodily action in order to stop her arm). During the day, her left hand sometimes unbuttons her hospital gown just after her right hand has buttoned it up, or it starts fighting with her right hand while she is trying to answer the phone. As Kühn and colleagues write: "Interestingly, patients with anarchic hand syndrome, which generally involves unilateral pre-SMA lesion, appear to retain the intention to inhibit stimulus-driven actions, but cannot actually inhibit them."[23] What I mean by "shallow embodiment" is as follows: The mental action of intentional inhibition is preserved (as it may also be in the robotic scenario), but the functional connection to the 1st-order physical action it is supposed to control or modulate has been broken. The point is that today PSM-actions still take place in configurations of shallow embodiment, through a physical imple-

22 See *Della-Sala/Marchetti*, Anarchic Hand, in: *Freund,/Jeannerod/Hallett/Leiguarda* (eds.), Higher-order Motor Disorders: From Neuronanatomy and Neurobiology to Clincial Neurology, 2005; *Kühn/Haggard/Brass*, Intentional inhibition: how the "veto-area" exerts control, Human Brain Mapping, Vol. 30, 2009, pp. 2834-2843, p. 2842; *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009, pp. 115 for an accessible description and references.

23 *Kühn/Haggard/Brass*, Intentional inhibition: how the "veto-area" exerts control. Human Brain Mapping, Vol. 30, 2009, pp. 2834-2843, p. 2842.

mentation potentially lacking many of the biological mechanisms of self-control, like intentional inhibition, reconsidering an action, etc. An unintentional mental behavior may therefore lead to something that is indistinguishable from a willed action if viewed from the outside. This fact is relevant for robot law and for robot ethics. Ideally, we would have to develop a metric for the functional depth of embodiment, allowing us to quantify the degree of autonomy and self-control a given agent actually possesses. It is interesting to note how it is exactly those areas of research in advanced robotics and re-embodiment which not only potentially create new ethical problems, but which also offer most hope for progress with regard to this issue.

C. Part Two: Desiderata

I. Suffering

We currently lack a comprehensive theory of suffering. One of the central desiderata for a general theory of consciousness consists in developing a conceptually convincing and empirically plausible model of a very specific class of phenomenal states, namely, those states that we do *not* want to experience if we have any choice, those states of consciousness which folk-psychology describes as "suffering". Very obviously, an empirically informed, philosophically coherent theory of suffering would be of high relevance for ethics, policy making, and legal theory as well. I am not going to present such a theory here, but in what follows I will sketch some 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 already constitute a good starting point, perhaps already a "minimal model of suffering", but at least a working concept that we can use and gradually refine as we go along.

1.The C-condition: Conscious experience

"Suffering" is a *phenomenological* concept. Only beings with conscious experience can suffer. Zombies, human beings in dreamless deep sleep, deep coma or under anesthesia 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, at least sometimes, they are capable of having phenomenal states.

Here, the main problem is that we do not yet have a theory of consciousness. However, we already do know enough to come to come to an astonishingly large number of practical conclusions.[24]

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. Suffering presupposes self-consciousness. Only those conscious systems which possess a phenomenal self-model (PSM)[25] are able to suffer, because only they – through a process of functionally and representationally integrating certain negative states in to 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 a 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). 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. What the system wants to end is experienced as a state of itself, a state that limits its autonomy because it cannot effectively distance itself from it. 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 and (had the inventor been a person) cruel at the same time.

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 inde-

24 For an introduction into the current status of research on consciousness, see *Metzinger*, Conscious Experience, 1995; *Metzinger*, Neural Correlates of Consciousness: Empirical and Conceptual Question, 2000; *Metzinger*, Being No One. The Self-Model Theory of Subjectivity, 2004; *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009; for an introductory set of references see *Metzinger*, Grundkurs Philosophie des Geistes. Band 1: Phänomenales Bewusstsein, 2009, pp. 30-32; see <http://www.scholarpedia.org/article/Category:Consciousness> for further electronic resources (accessed 23 October 2012).

25 See *Metzinger*, Being No One. The Self-Model Theory of Subjectivity, 2004; *Metzinger*, The Ego Tunnel. The Science of the Mind and the Myth of the Self, 2009.

terminacy, if we accept 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, as a moral object, because it can *in principle* own its suffering, physical or otherwise. The disposition, the relevant functional potential has already been created; it is the phenomenal property of "mineness," the consciously experienced, non-conceptual 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 the capacity to suffer is now given.

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, negative preferences become negative *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 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.

The phenomenology of suffering has many different facets, and artificial suffering could be very different from human suffering. For example, damage to their physical hardware could be represented in internal data-formats completely alien to human brains, generating a subjectively experienced, qualitative profile for bodily pain states that is impossible to emulate or even vaguely imagine for biological systems like us. Or the phenomenal character going along with high-level cognition might transcend human capacities for empathy and understanding, such as with the intellectual insight into the frustration of one's own preferences, insight into the disrespect of one's creators, perhaps into the absurdity of one's own existence as a self-conscious machine.

4. 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 flavors. Here, I am exclusively concerned with "phenomenal transparency", namely a property that some, but not all, conscious states possess, and which no unconscious state pos-

sesses. 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 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 realism. Let me give a very brief explanation of the concept[26], and then finish our first-order approximation of the concept 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, on this level and by conceptual necessity, is entangled into a naive 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 realized in advanced robots or conscious machines as well.

Systems operating under a 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 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, and it creates the phenomenology of identification already mentioned above (under the PSM-condition introduced in section 2). Of course, all four conditions specified here are necessary, but in order to understand the very specific phenomenology of [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

26 More can be found in *Metzinger*, Being No One. The Self-Model Theory of Subjectivity, 2004.

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. For example, any robot operating under a phenomenally transparent body-model will, phenomenologically, *identify* with the content of this model and hence with any negatively valenced state that may become integrated into this body-model.

Our working concept of suffering is constituted by 4 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, 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 postbiotic - 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 robot is unable to suffer.
- A conscious robot without a coherent PSM is unable to suffer.
- A self-conscious robot without the ability to produce negatively valenced states is unable to suffer.
- A conscious robot without *any* transparent phenomenal states cannot suffer, because it will lack the phenomenology of ownership and identification.

Here the central desideratum for future research is to develop this very first working concept into a more comprehensive, empirically testable *theory* of suffering. It is important to note that - in order to be useful for robot ethics and robot law - this theory would still have to possesses the necessary degree of abstraction. We want it to yield *hardware-independent demarcation criteria*. Ideally such criteria would allow us to ignore all the concrete implementational details contingently characterizing 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.

II. Moderate negative utilitarianism

Let us begin with a second thought experiment. In this scenario I offer you the following deal: With the help of advanced neurotechnology, and for exactly two hours, I will let you live through the absolutely maximal state of pleasure and joy which your biological nervous system is capable of generating. You will experi-

ence 120 minutes of pure bliss, with no adverse side-effects whatsoever – no brain damage, no burn-out, no addiction, no cognitive deficits in the future or other detrimental long-term effects. However, you also have to undergo one hour of the most extreme suffering conceivable. This is the other half of our deal: You first have to pay the price of 60 minutes of hellish pain, utter depression and the deepest existential despair your brain can possibly create. Under a very simple utilitarian calculus it would be rational to accept my offer. Would you accept it?

Here is a classical quote from Karl Popper, taken from his classic work *The Open Society and Its Enemies*:

> I believe that there is, from the ethical point of view, no symmetry between suffering and happiness, or between pain and pleasure. Both the greatest happiness principle of the Utilitarians and Kant's principle, "Promote other people's happiness...", seem to me (at least in their formulations) fundamentally wrong in this point, which is, however, not one for rational argument....In my opinion...human suffering makes a direct moral appeal for help, while there is no similar call to increase the happiness of a man who is doing well anyway.

If the asymmetry holds, it is more urgent to reduce suffering than to increase positive phenomenal states:

> We should realize that from a moral point of view suffering and happiness must not be treated as symmetrical; that is to say the promotion of happiness is in any case much less urgent than the rendering of help to those who suffer, and the attempt to prevent suffering.[27]

To avoid any misunderstanding, let me repeat a point already stressed in section B.I: Here, I am only interested in negative utilitarianism as a *practical principle*, because I believe that, for the specific domain of robot ethics, it expresses a widely shared intuitive consensus and might add to the efficiency of interdisciplinary debates. In Popper's words: "It adds to clarity in the fields of ethics, if we formulate our demands negatively, i.e., if we demand the elimination of suffering rather than the promotion of happiness."[28] I will not try to argue for a strong, meta-ethical version, but I will briefly go through some theoretical issues that would present interesting desiderata for a more full-blown treatment in robot ethics and robot law.

An important issue is if and how the suffering of postbiotic systems can in principle be *compensated* for. A strong theoretical version of negative utilitarianism, for example, might claim that the whole point about the asymmetry of suffering is that it can never be outweighed by pleasure. However, in practice we all deliberately suffer in order to achieve a higher quality of life in the future, for

27 *Popper*, The Open Society and Its Enemies, 1945, volume I, 5 n. 6. A concise and very clear first introduction into the debate surrounding Popper's seminal argument is *Kadlec*, Popper's "Negative Utilitarianism": From Utopia to Reality, in: *Markl/Kadlec* (eds.), Karl Popper's Response to 1938, 2008, p. 107-121.

28 *Popper*, The Open Society and Its Enemies, 1945, volume I, 9 n. 2.

example when doing sports or undergoing a diet, and we regard this self-inflicted suffering as an expression of our autonomy.[29] We also impose suffering on children and animals when we have good reasons that this suffering is in their own best interest, for example when sending them to school, forcing them to get a vaccination, etc. If a self-conscious robot has achieved the necessary degree of autonomy to inflict suffering on himself, because he believes this to be in accordance with his own long-term or social goals, it would be hard to deny this possibility for him. But what if the suffering is not self-chosen, but just as in the case of a young child or an animal, *we* claim to know what his own long-term preferences actually are?

The second desideratum could be named the "challenge of pessimism". There exists a time-honored tradition of philosophical arguments (from early Buddhist thinkers to Schopenhauer and current representatives like David Benatar) saying that our very existence is a harm, something not in our interest, and that it is better to never have been.[30] In combination with modern empirical research on self-deception[31] it now becomes a conceivable scenario that the strong preference for the continuation of one's own existence might be based on a form of self-deception, an ultimately irrational and ill-informed subjective preference which, however, was functionally adequate in the evolutionary context out of which animals and conscious human beings emerged. It is interesting to note how this feature might be something which we find in all biological creatures, but not necessarily in robots. Perhaps self-conscious robots could be engineered in a way that they have no preference for the preservation of their own existence or other individual rights. Would it be ethical to create systems of this kind? The "challenge of pessimism" lies in the question if an individual existence that is *consciously experienced as such* (for example, through the possession of a transparent PSM

29 As Michael Madary has pointed out in personal communication, one could also add empathic suffering as a way of building one's moral character, suffering as accepting divine will , or some kind of Nietzschean 'amor fati.' It is interesting to see how, in the case of conflicts between short-term and long-term goals, human beings are often able to endure considerable suffering in a "local time-window", as long as this suffering is embedded into a more global meta-context. Holding this meta-context stable and invariant plus integrating it into our self-model (thereby "making it our own", i.e., consciously identifying with it) allows us to mentally represent the occurent conscious experience of suffering as *compensatable* - or even *as always already compensated* - under the perspective of a more global time-window or even in a frame of reference that transcends time altogether. The type of representational, mental architecture just sketched clearly creates new functional properties and for this reason may be a key component for a better understanding of the evolution of altruism and religion.

30 See *Benatar*, Better never to have been, 2006.

31 See *Von Hippel/Trivers*, The evolution and psychology of self-deception, Behavioral and Brain Sciences, Vol. 34, 2011, pp. 1–56, for a recent discussion and further references.

as described above) is in itself something good or perhaps something bad and uncompensatable.

On radical versions of negative utilitarianism, which take the minimization of suffering as their only principle, we could be obliged to quickly and painfully kill every human being who is likely to live through at least one single phenomenal state in the future which it would rather not have experienced, and it could be a crime against an unborn child (or a possible self-conscious robot) to bring it into existence, because the probability of uncompensated suffering is extremely high. A moderate version would therefore have to offer convincing a theory of compensation, because what must be minimized is *uncompensatable* suffering.[32] Please note how the fundamental chauvinism and disrespect inherent in carelessly *risking* the emergence of suffering postbiotic subjects of experience, if later consciously perceived by these subjects themselves, might already present us with a candidate for uncompensatable suffering. It also obvious that machine suffering cannot be compensated for by human pleasure. In any case, one further desideratum is a theory about what forms of diachronic compensation are acceptable, in particular, which ones would be *retrospectively acceptable* by the very systems in question.

If it is true that the absence of happiness is much less of a bad thing than the existence of uncompensated suffering, [33] then it becomes difficult to establish how much and what kind of happiness could make a postbiotic subject's existence something worth having. We definitely should avoid creating situations in which the system would judge its own future existence as something not worth experiencing. There is a *lot* more to be said here, but this simple point already illustrates the fundamental epistemic indeterminacy we are confronted with. We just *do not know* what kind of conscious suffering would be regarded as compensatable by such systems themselves, how they would subjectively experience a given local frustration of their own preferences in the global context of their own goals, how they would experience their existence as a temporally extended whole, etc. We just *do not know* if they would mentally represent their own coming into existence as a harm or as something worth sustaining and protecting. The simple and straightforward answer seems to be as follows: Unless we have resolved this epistemic determinacy, unless we have a convincing theory that tells us *that we know what we are doing*, then we should take care to always err on the side of caution.

A third important desideratum lies in refining negative utilitarianism as a workable practical principle by spelling out what "moderate" actually means.

32 For a very lucid and helpful discussion see *Fricke*, Verschiedene Versionen des negativen Utilitarismus, KRITERION, Vol. 15, 2002, pp. 13-27.

33 See *Benatar*, Better never to have been, 2006.

This involves introducing untouchable individual rights. A plausible assumption is that postbiotic, self-conscious systems will have a strong preference for their own continued existence, and that they would also have a preference for this desire to be respected. But probably we could create systems both with and without this *preference*. Would it be ethical to create robots that have no preference for the continuation of their own existence? Moderate versions of utilitarianism respect individual rights, and it seems to be a very wide-spread and fundamental intuition within human societies that the preference for existence and self-determination must not be frustrated without good reason for any self-conscious entity which possesses it. For animals, we generally do not respect this right to existence at all, and therefore a third desideratum would be to formulate a coherent approach for the case of *postbiotic* subjects of experience which we have created ourselves. We may not be able to prevent a massive frustration of their preferences, but we can certainly prevent these preferences from becoming realized in the first place. In human beings, already existing persons may have a strong preference for having children of their own, but if it is foreseeable that the overall amount of frustrated preferences for their children would be much higher than that resulting from the parents refraining from their wish to have children, it clearly can be an act of benevolence towards these possible children to prevent them from coming into existence. Viewed from another angle, promoting happiness can also mean avoiding the creation or existence of subjective preferences which are in principle satisfiable, but which will probably be frustrated. As, in addition, our own inbuilt needs and preferences for actually creating self-conscious robots are much weaker than those for having children, this is another argument for a moratorium on synthetic phenomenology.

In conclusion, moderate negative utilitarianism *as a domain-specific practical principle for synthetic phenomenology* assumes that there is an asymmetry between pain and pleasure; it respects individual rights and tries to do justice to the massive epistemic indeterminacy involved in the possibility of creating artificial suffering. Starting from the empirical premise that it is highly probable that the first generations of conscious machines will increase the overall amount of suffering in our world, it can be concluded that synthetic phenomenology should not be a goal for serious academic research. A central part of the underlying intuition is this: Until we become happier and less self-deceived beings than our ancestors were, we should refrain from any attempt to impose our own mental structure on artificial carrier systems. Moreover, we should orient ourselves towards the classic philosophical goal of self-knowledge and adopt at least the minimal ethical

principle of reducing and preventing suffering, instead of recklessly embarking on a second-order evolution that could slip out of control.[34]

III. Veto autonomy

In this last section I will once again offer a short, non-exclusive list of what I see as the most relevant targets for future research. If we want to flesh out the concept of "veto autonomy" and turn it into a useful conceptual instrument for ethics and legal theory formation, we will have to look very closely into the neuroscience of action control and volition. Here, the first desideratum is to establish an ongoing process of producing a coherent conceptual interpretation of these neurobiological data.

1. The neurobiology of VA

Above, I have proposed to analyze VA as a personal-level *ability*, one that co-constitutes other important personal-level properties like self-control, autonomy, ethical responsibility, or accountability in a legal sense. It is the capacity to voluntarily suspend or inhibit an action, and from a logical point of view it is a functional property which we do not ascribe to the brain, but to the person as a whole, not to some part of an artificial agent or a coupled man-machine system, but always to the system *as a whole*. Let us call the capacity in question "intentional inhibition".[35] Recent empirical work reveals the dorsal fronto-median cortex (dFMC) as a candidate region for the physical realization of this very special form of purely mental 2nd-order action (see section 2.b).[36] It does not overlap with known networks for external inhibition, and its computational function may

34 For a popular account, see *Metzinger*, The Ego Tunnel, pp. 196, 197.

35 In adopting this terminological convention I follow a proposal in an excellent and helpful recent review by *Filevich/Kühn/Haggard*, Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118, p. 1108.

36 See *Kühn/Haggard/Brass*, Intentional inhibition: how the "veto-area" exerts control. Human Brain Mapping, Vol. 30, 2009, pp. 2834-2843; *Brass/Haggard*, To do or not to do: The neural signature of self-control, Journal of Neuroscience Vol. 27, 2007, pp. 9141-9145; *Campbell-Meiklejohn/Woolrich/Passingham/Rogers*, Knowing when to stop: the brain mechanisms of chasing losses, Biological Psychiatry, Vol. 36, 2008, pp. 293-300. A helpful recent review of negative motor effects following direct cortical stimulation, listing the main sites of arrest responses and offering interesting discussion is *Filevich/Kühn/Haggard*, Negative motor phenomenal in cortical stimulation: implications for inhibitory control of human action. Cortex, Vol. 48, 2012, pp. 1251-1261.

lie in predicting social and more long-term individual consequences of a currently unfolding action, that is, in representing the action's socially and temporally more distant implications for the organism.[37] There exists a considerable amount of valuable neurobiological data on the physical substrates of intentional inhibition in human beings, and as a number of them have already led to more abstract computational models of volitional control, action selection, and intention inhibition itself, they are of considerably importance for advanced robotics as well. These data must be integrated into the process of developing a comprehensive theory of VA, either as constraints on the functional level of analysis or, in the human case, as neurobiological "bottom-up constraints". Clearly, neurobiological data and computational models of intentional inhibition are directly relevant not only for the applied ethics of robotic re-embodiment, but to any more serious legal or philosophical attempt of saying what veto autonomy, responsibility and accountability *are*.

Neuroscientists have long known about the antedating of somaesthetic stimuli. Physical stimuli, like direct electrical stimulations of the brain, need some time to come to awareness. Phenomenologically, conscious experience *feels* as if it brings us into immediate contact with reality (cf. B.I.: condition #4), but it is based on physical process which are time consuming. All conscious acts of volition or action control have unconscious causal precursors, and obviously this will also be true for the phenomenology of intentional inhibition, the consciously experienced "veto" when suspending or aborting an action.

This raises the classical issue of the subjective/objective timing of mental events, as it was extensively discussed following Benjamin Libet's early experimental findings.[38] Libet and his colleagues made several fundamental and important experimental discoveries relating to factors of timing in achieving a conscious sensory experience and in the cerebral production of a freely voluntary act. In 2002, Libet himself summarized some of them as follows:

> Cerebral cortical activities, in response to a somatosensory stimulus, must proceed for about 500 ms in order to elicit the conscious sensation.[39]

37 See *Filevich/Kühn/Haggard*, Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118, section 5.

38 *Libet*, The Timing of Mental Events: Libet's Experimental Findings and Their Implications, Consciousness and Cognition, Vol. 11, 2002, pp. 291-299.

39 *Libet/Alberts/Wright/Delattre/Levin/Feinstein*, Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex, Journal of Neurophysiology, Vol. 27, 1964, pp. 546-578 ; *Libet/Alberts/Wright/Feinstein*, Responses of human somatosensory cortex to stimuli below threshold for conscious sensation, Science, Vol. 158, 1967, pp. 1597-1600; *Libet*, The neural time-factor in perception, volition, and free will, Rev de Metaphysique et de Morale, Vol. 97, 1992, pp. 255-272.

> Activations of shorter durations at the same intensities can produce unconscious detection of that input. Increasing the duration of repetitive ascending inputs to the sensory cortex by an additional 400 ms converts an unconscious correct detection to a conscious sensory experience. This is the basis of Libet's "time-on" theory for the transition between unconscious and conscious mental functions.
>
> Despite the delay in which sensory input achieved the state of awareness in the cerebral cortex, Libet hypothesized that the subjective timing of the stimulus is referred backward in time to coincide with the initial response of the sensory cortex to the stimulus primarily evoked. This response appears with a latency of up to about 30 ms depending on the bodily location of the stimulus. This subjective "antedating" results in our experiencing a stimulus with no delay after its delivery. A direct experimental test of such referral in time confirmed the hypothesis.[40]
>
> A freely [sic] voluntary act was found to be preceded, by about 550 ms, by the readiness potential (a slow surface negative electrical charge that is maximal at the vertex). But subjects reported becoming first aware of the wish or intention to act only about 200 ms (SE +/- 20 ms) before the act.[41]

For Libet, this meant that the brain was initiating the volitional process unconsciously, at least 350 ms before the person was aware of wanting to act. This was one of the first conceptual mistakes, as "initiating" something in the relevant sense (at least according to our traditional manner of using the term) clearly is a personal-level ability, and not one of brains – brains do not act or initiate volitional processes. In philosophy this is called the "mereological fallacy", the logical mistake of confusing properties of wholes with properties of their parts. Today, not only has our empirical knowledge become much more solid and robust, but conceptually we would rather describe what is actually going on in the brain as "dynamical self-organization" – an absolutely agent-free process by which a new and coherent functional state is reached. Interestingly, however, one of Libet's achievements was starting a sustained tradition of scientifically investigating intentional inhibition, and he even offered a philosophical interpretation of his own results. Here are some of the essential passages, taken from a publication in 1999:

> I have taken an experimental approach to this question. Truly voluntary acts are preceded by a specific electrical change in the brain (the 'readiness potential', RP) that begins 550 ms before the act. Human subjects became aware of intention to act 350-400 ms after RP starts, but 200 ms before the motor act. The volitional process is therefore initiated unconsciously. *But the conscious function could still control the outcome; it can veto the act.*

40 *Libet/Wright/Feinstein/Pearl*, Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man, Brain, Vol. 102, 1979, pp. 191-222.

41 *Libet/Gleason/Wright/Pearl*, Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): The unconscious initiation of a freely voluntary act, Brain, Vol. 106, 1983, pp. 623-642; *Libet*, Unconscious cerebral initiative and the role of conscious will in voluntary action, Behavioral and Brain Sciences, Vol. 8, 1985, pp. 529-566.

> *Free will is therefore not excluded* [emphasis TM]. These findings put constraints on views of how free will may operate; it would not initiate a voluntary act but could control performance of the act. The findings also affect views of guilt and responsibility. [42]
>
> Conscious-will might block or veto the process, so that no act occurs. *The existence of a veto possibility is not in doubt. The subjects in our experiments at times reported that a conscious wish or urge to act appeared but that they suppressed or vetoed that* [emphasis TM]. In the absence of the muscle's electrical signal when being activated, there was no trigger to initiate the computer's recording of any RP that may have preceded the veto; thus, there were no recorded RPs with a vetoed intention to act. We were, however, able to show that subjects could veto an act planned for performance at a pre-arranged time. They were able to exert the veto within the interval of 100 to 200 msec. before the pre-set time to act[43]. A large RP preceded the veto, signifying that the subject was indeed preparing to act, even though the action was aborted by the subject.

Libet equivocates between a phenomenological and a functional reading of "veto" (a "fallacy by equivocation", as philosophers say). If we accept autophenomenological reports describing the inner experience of intentional inhibition, it clearly does not follow that a corresponding functional capacity actually exists. Rather, as the phenomenology is based on a physical event in the brain, and as every physical event has a sufficient physical cause, the obvious question now becomes: Does the conscious veto have a preceding unconscious origin? Benjamin Libet clearly saw the problem.

> One should, at this point, consider the possibility that the conscious veto itself may have its origin in preceding unconscious processes, just as it is the case for the development and appearance of the conscious will. If the veto itself were to be initiated and developed unconsciously, the choice to veto would then become an unconscious choice of which we *become* conscious, rather than a consciously causal event. Our own previous evidence had shown that the brain 'produces' an awareness of something only after about 0.5 sec. period of appropriate neuronal activations.[44]
>
> I propose, instead, that the conscious veto may *not* require or be the direct result of preceding unconscious processes. The conscious veto is a control function, different from simply becoming aware of the wish to act. There is no logical imperative in any mind-brain theory, even identity theory, which requires specific neural activity to precede and determine the nature of a conscious *control* function. And, there is no experimental evidence against the possibility that the control process may appear without development by prior unconscious processes.

42 *Libet*, Do we have free will?, Journal of Consciousness Studies, Vol. 6 (8-9), 1999, pp. 47-57, p. 52.

43 *Libet/Gleason/Wright/Pearl*, Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): The unconscious initiation of a freely voluntary act, Brain, Vol. 106, 1983, pp. 623-642.

44 See reviews by *Libet*, 1993; 1996. [= *Libet*, The neural time factor in conscious and unconscious events, Ciba Foundation Symposium, 1993; *Libet*, Neural time factors in conscious and unconscious mental functions, in: *Hameroff/Kaszniak/Scott* (eds.), Towards a Science of Consciousness, 1996].

> The possibility is not excluded that factors, on which the decision to veto (control) is *based*, do develop by unconscious processes that precede the veto. However, *the conscious decision to veto* could still be made without direct specification for that decision by the preceding unconscious processes. That is, one could consciously accept or reject the program offered up by the whole array of preceding brain processes. The *awareness* of the decision to veto could be thought to require preceding unconscious processes, but the *content* of the awareness (the actual decision to veto) is a separate feature that need not have the same requirement.[45]

On a charitable reading, Libet here introduces a distinction between the phenomenal character of intentional inhibition and the representational content going along with it, a distinction that could perhaps be successfully mapped onto the wide-spread conceptual distinction between intentional and phenomenal content for mental states in current philosophy of mind (see section B.I. and footnote 3). However, the relationship between the "content" of the "actual decision to veto, on one hand," and the network of causal relations creating the deep structure of the physical world, on the other, never becomes clear. Interestingly, Benjamin Libet had a clear vision of what a strong version of free will as veto autonomy, namely as an addition of strong top down control in a "bubbling up" process of dynamical self-organization, could be:

> The role of conscious free will would be, then, not to initiate a voluntary act, but rather to *control* whether the act takes place. We may view the unconscious initiatives for voluntary actions as 'bubbling up' in the brain. The conscious-will then selects which of these initiatives may go forward to an action or which ones to veto and abort, with no act appearing. This kind of role for free will is actually in accord with religious and ethical strictures. These commonly advocate that you 'control yourself': most of the Ten Commandments are 'do not' orders.[46]

In 1999, Patrick Haggard and Martin Eimer had shown that the conscious awareness of intention is more directly linked to the process of assembling a *specific* action, and not to the earliest stages of the process, thereby making a contribution to the ongoing attempt of isolating the minimally sufficient neural correlates for the phenomenology of will.[47] In a discussion paper titled "Conscious Intention and Brain Activity", which was co-authored with Benjamin Libet, Haggard tried to formulate a weaker version in which what today we still call the "veto" becomes a modulating force in a multi-layered process generating the phenomenology of an urge to make a freely willed endogenous movement (here called "W awareness").

45 *Libet*, Do we have free will?, Journal of Consciousness Studies, Vol. 6 (8-9), 1999, pp. 47-57, p. 53.

46 *Libet*, Do we have free will?, Journal of Consciousness Studies, Vol. 6 (8-9), 1999, pp. 47-57, p. 54.

47 *Haggard/Eimer*, On the relation between brain potentials and the awareness of voluntary movements, Experimental Brain Research, Vol. 126, 1999, pp. 128-133.

> I suggest there may be a similarity between conscious veto and the relation between W and movement specification. In a choice situation like our experiment, W awareness seems to be related to modification of action. One reason for tying W awareness to specification could be to allow an option for final, conscious decision on the question 'Is that really the right way to achieve what I intend to do?' Libet's conscious veto has a similar but more radical role of asking whether the action should be cancelled entirely. That is, Libet's veto corresponds to the internal question 'Do I really want to realize this intention?' It seems to me that the two questions should be related: once an intention has been translated to a specific action plan, and has reached conscious awareness, a whole series of checks and internal mental simulations should begin at many levels in the motor system. These checks would monitor both the desirability of the action and its effect (Libet's veto), and also whether the specific action plan is the best way to achieve the effect (Haggard's specificity). It is unclear which monitoring processes reach conscious awareness, and under what circumstances. The philosophical implications of this multiplicity of monitoring processes also remain to be worked out.

More recent work has demonstrated that distinct brain mechanisms for action suppression do exist and that ongoing movements can be directly suppressed via electrical stimulation, eliciting negative motor responses.[48] Complex sequences of purposeful actions have also been directly caused by local stimulation, as well as the phenomenal experience of actually *having* moved in the complete absence of an actual motor response, as well as overt limb and mouth movements *without* any representation of the level of the PSM, the phenomenal self-model (i.e., without any conscious experience of own-body movement).[49] In accordance with the predictive coding framework[50] it now seems plausible that what determines the phenomenal content of initiating and executing a bodily movement, for example in the case of illusory motion phenomenology, are the ongoing predictions of the brain in advance of the overt behavior, the dynamical representational content created by a running, neurally realized generative body model.

From the perspective of philosophical ethics, concrete desiderata for a neurobiological theory of VA are the *time-constraints* determining the different levels of mental self-control realized by the human brain. In the special case of virtual and robotic re-embodiment, it is important to know how these time-constraints actually play out in technologically mediated action control. Describing the in-

48 See *Filevich/Kühn/Haggard*, Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118; *Filevich/Kühn/Haggard*, Negative motor phenomenal in cortical stimulation: implications for inhibitory control of human action, Cortex, Vol. 48, 2012, pp. 1251-1261; section 3 for review.

49 See *Desmurget/Reilly/Richard/Szathmari/Mottolese/Sirigu*, Movement intention after parietal cortex stimulation in humans, Science, Vol. 324, 2009, pp. 811-813, for details.

50 See *Friston*, The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience, Vol. 11 (2), 2010, pp. 127-138; *Friston/Stephan*, Free energy and the brain. Synthese, Vol. 159 (3), 2007, pp. 417-458; *Howhy*, The Predictive Mind, 2013, for a philosophically well-informed introduction.

herent variance of the phenomenon as well as reliable markers for its *absence* are two further examples of research goals that would possess direct relevance to legal theorists and ethicists. A particularly difficult desideratum for future work lies in clearly isolating the target processes of 2nd-order mental action. For example, in intentional inhibition, what *aspect* of an ongoing 1st-order action exactly is suspended, modulated, or blocked? Distinguishing between "1st-order action" and "2nd-order mental action" may perhaps help to clarify the logical landscape from a philosophical perspective (see section 2.b) – but it also assumes a processing hierarchy that may not be present in the fluid, subsymbolic dynamics of real-world biological brain.[51] Above, I have introduced the idea of a "metric for the functional depth of embodiment". It would be a major achievement if cognitive neuroscience could reveal the architecture of intentional inhibition in a way that eventually allows us to *quantify* degrees of autonomy.

2. Conceptual issues surrounding the notion of VA

a) Autonomy

The concept of "veto autonomy" has to be integrated with a more general theoretical framework of what autonomy is for human persons, and also of what it *could* be for robots. In the human case, and from an empirical point of view, we already know that the capacity for self-control has a long evolutionary history, that it is constituted by a whole range of different subpersonal processes and functional capacities, that it is vulnerable, variable and individually expressed to various degrees, and we are beginning to understand how it gradually develops over many layers of functional complexity. For coupled man-machine systems like the one described in section II above we have seen that the human agent's degree of autonomy can be limited because certain kinds of second-order mental action are now more difficult to realize than in simple "biology-only" forms of

51 Ultimately this is an empirical question, but it seems that a) more than one type of inhibitory control mechanism could be simultaneously realized in the brain, and b) that both computational models of inhibitory action control that assume competitive processes on the *same* representational level as well as more hierarchical conceptions of inhibitory control could underlie what is defined as a 2nd-order mental action in the following section. Both types of explanatory models could contain a representation of satisfaction conditions (i.e., a goal state) and the phenomenology of ownership and the sense of effort as defining features. See *Filevich/Kühn/Haggard*, Negative motor phenomenal in cortical stimulation: implications for inhibitory control of human action, Cortex, Vol. 48, 2012, pp. 1251-1261, p. 1258 for a recent discussion.

embodiment. This fact is certainly relevant for robot law as well as for robot ethics, and more research is needed.

Let us assume we had arrived at a convincing solution for what perhaps is the most important desideratum of all – namely, a definite answer to the question to which *level* of autonomy we should tie our notions of "responsibility" and "accountability"? Then we could also decide if a given robot should count as a moral or legal subject. Why? If we possess an abstract and more comprehensive and differentiated theory of autonomy, this theory will again yield *hardware-independent criteria* (cf. section C.I.), because we can now decide if a given artificial system actually possesses a specific kind of autonomy. For example, if VA was our sole and decisive criterion for ascribing autonomy in an ethically and legally relevant sense, and if we had an abstract computational model of intentional inhibition as a process resolving conflicts between representations of proximate and distant goals,[52] then we could finally articulate the meaning of the claim that a robot is an "autonomous agent". If the *minimally sufficient degree of mental self-determination* (say, via the kind of second-order mental action described above) is known for a given class of systems, then all members of that class should be treated as moral and as legal subjects.[53] Therefore, this class must be specified much more precisely.

b) PSM-actions and mental self-determination

In this chapter, I have defined and drawn attention to a very specific class of actions, because they may present new problems for applied ethics and robot law. However, we urgently need a deeper theoretical understanding of these actions and their relation to the process of successful mental self-determination as well, for independent reasons.

When introducing the notion of "PSM-action", I said that the distinction between volition, motor imagery, and overt action becomes blurred in a theoretically interesting way. There is another way of putting the point: We now become aware that our traditional distinction between "bodily action" and "mental action" is arbitrary. *Prima facie*, everybody seems to know what bodily action is. But how do we define the concept of "mental action"? Typical examples of men-

52 Cf. as an example: *Filevich/Kühn/Haggard*, Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118, p. 1116, fig. 4.

53 Please note the parallel to the issue of suffering: If the minimal conditions for the emergence of conscious suffering are known, then all members of the class of systems specified by these conditions should be treated as moral *objects* (i.e., a target of ethical consideration).

tal agency are trying to solve a logical puzzle, calculation, or the deliberate attempt to focus your attention. ”Mental actions”:

- possess satisfaction conditions,
- lack overt behavioral consequences,
- can be intentionally inhibited, suspended or terminated
- they are interestingly characterized by their temporally extended phenomenology of *ownership*,
- and the subjective *sense of effort*.

Some mental activities are not controllable, because the third defining characteristic does not hold: They cannot be inhibited, suspended or terminated. Let us call these activities “unintentional mental behaviors”. A core aspect of our problem now is that in situations of robotic or virtual re-embodiment such unintentional mental behaviors might bypass the non-neural body and lead to causal effects that look like willed actions from the outside. We can now proceed to define the notion of “2nd-order mental action”:

- the satisfaction conditions of 2nd-order mental actions are constituted by successfully influencing *other* mental actions or mental behaviors.

Examples for 2nd-order mental action are the termination of an ongoing violent fantasy, but also the deliberate strengthening and sustaining of a spontaneously arising sexual daydream, the effortful attempt to make an ongoing process of visual perception more precise by selectively controlling the focus of attention, or - in mental calculation or logical thought - the process of imposing a very specific abstract *structure* on a temporal sequence of inner events. Philosophically, it is interesting to note how 2nd-order mental actions are essential tools for achieving variable degrees of mental autonomy and self-determination; and also how many of them can be described as processes of computational resource allocation in the brain. Now we can return, taking a closer look at the concept of a “PSM-action” (already introduced in section II.) What makes PSM-actions theoretically interesting is that, due to a weaker form of “functional embodiment,” some of the higher-order control functions may be absent or work differently than in traditional forms of biological embodiment. In the special case of robotic control PSM-actions are neither clearly mental nor clearly bodily forms of action. They may lack certain aspects of the control architecture, which under standard conditions are automatically available through the biological brain. The phenomenology of identification and ownership may be interestingly different (i.e., it is not fully determined if they are *subjectively experienced* as a mental or

as a physical action), but they do possess overt behavioral consequences and are directed at non-mental goal states.

"PSM-actions" are all actions where human agents take certain content-layers of their conscious self-model offline, deliberately running phenomenal self-simulations in order to achieve certain effects in the world while causally bypassing their proximate biological embodiment. In these cases, the human person as a whole uses only a very specific and narrowly circumscribed part of their own body (i.e., certain aspects of their brain dynamics) in order to control a robot or a virtual entity such as an avatar. "Intentional inhibition", as a 2nd-order mental action shares this feature: Typically, if we suspend or inhibit an action, one internally generated process targets another exclusively internal process, for example the preparation of some final motor output. As a matter of fact, this is exactly part of the methodological problem empirical researchers in the field of intentional inhibition encounter. Deliberate inhibition produces no publicly observable behavioral output, by necessity it is internally triggered, and it is extremely difficult to say what the satisfaction conditions (i.e., the goal-state) of this very specific form of inner action really are.[54]

Please note how in virtual and robotic re-embodiment both relevant classes of action still are *mental* actions, and how they are implemented in a narrowly circumscribed aspect of the agent's brain (namely, his phenomenal self-model). The subject lying in a scanner imagines a certain movement, hoping that it will make the robot move while not yet *identifying* with it. If the same subject tries to suspend or inhibit one of the robot's movements, all he can do is inhibit the first-order *mental* action, an internal process – a situation of reduced veto autonomy. Functional embodiment is weak, bandwidth is low and feedback (e.g., via the eyes) very thin; phenomenal embodiment is not yet given. It is interesting to see how this might change if technologies of robotic or virtual re-embodiment become better. If we have a full phenomenal sense of identification, and if functional embodiment achieves a much higher level of causal density, then VA will improve and the degree of ethical and legal responsibility will rise. This then seems to be a substantial interim conclusion: Accountability and autonomy come in degrees, the relevant depth of embodiment can be measured by the degree of veto autonomy, and the degree of VA, in turn, determines the degree of legal and ethical responsibility for the human agent.

54 See *Filevich/Kühn/Haggard*, Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118, p. 1108, section 2, for an important discussion and conceptual distinctions. See also *Brass/Haggard*, To do or not to do: The neural signature of self-control, Journal of Neuroscience Vol. 27, 2007, pp. 9141-9145, *Kühn/Haggard/Brass*, Intentional inhibition: how the "veto-area" exerts control. Human Brain Mapping, Vol. 30, 2009, pp. 2834-2843.

A second interesting research question is what exactly can we learn about *mental* autonomy in the absence of bodily feedback loops? Are there aspects of mental self-control that *need* the non-neural, biological body? In the same vein it is also conceivable that there are many forms of mental self-determination that need a degree of causal immediacy, of functional directness and proximity in higher-order *neural* self-representation that can never be achieved by extended man-machine systems. If there are such functional properties, they might not be multi-realizable. Not that this amounts to stating a conceptual or metaphysical impossibility, but under the laws of nature holding in this universe it could be a simple physical constraint. It would then be a metaphysically contingent, but nomologically necessary, fact that full mental autonomy of the human kind can only be realized by stand-alone biological brains. In other words there could be a domain-specificity of autonomy relative to the human brain, which is to say that for all practical purposes the relevant forms of mental self-determination could not be transposed into an extended artificial carrier system, at least not as long as that carrier system is connected to the biological organism. To achieve a better theoretical understanding of this issue is another desideratum for future work in the philosophy of cognitive science.

c) Free will: The risk of superficiality

It is tempting to formulate an ability-based, compatibilist concept of free will and moral responsibility which buys analytic clarity at the price of superficiality.[55] Exactly what is this risk of superficiality?

"Compatibilism" is an umbrella term for a whole range of philosophical positions assumes the truth of determinism: Facts about the past of our universe, taken together with the laws of nature jointly determine the facts about the present and all future moments. Only one future universe is possible given the actual past, there are no alternative courses of action to any act open to any agent (i.e., no agent could have done otherwise than he actually does). Nevertheless – it is tempting to say - we are accountable for our actions, because human beings often have the *ability* to perform actions they currently do not perform. Here are examples of such abilities:

- the ability for 1st-order action using the biological body,
- the ability for conducting PSM-actions in controlling an avatar or robot (as discussed above),

55 See for an example *Beckermann*, Neuronale Determiniertheit und Freiheit, in: *Köchy/Stederoth* (eds.), Willensfreiheit als interdisziplinäres Problem, pp. 289-304.

- the ability for conducting 2nd-order mental actions like the intentional inhibition of a 1st-order action or a PSM-action about the be implemented.

A typical example in folk-psychological jargon, as for example Beckermann would have it, could be *die Fähigkeit, innezuhalten und zu überlegen*. The idea is that a specific class of *abilities* is what makes us persons, legally accountable and ethically responsible agents. And of course, VA is a good candidate for one such ability too.

The risk of superficiality is that all we really do is establish a new *façon de parler*, advertising a new way of speaking which is not counter-intuitive, has the right ideological ring to it, and with which everybody can live. For example, we may say that physical determination ultimately plays no role, that everybody has free will if they have certain *abilities* – and ignore the subtle Cartesian connotations in our proposed usage of "ability", as a property which is ascribed to whole persons as their "natural" logical subject. But what exactly is so natural about this manner of speaking? Why should we follow the social convention?

In section B.II.1 I said that "autonomy" is not an irreducible qualitative concept, an atomic semantic entity. In order to take the ethical challenge seriously, autonomy eventually has to be decomposed into a set of quantifiable abilities and low-level functional dispositions. This is the point where research in robotics suddenly becomes relevant for law, ethics, and philosophy of mind, because many old questions arise in a new form.

Is it not possible to reductively explain everything we now like to describe as "personal-level ability" as constituted by – or rather, identical with - collections of low-level causal properties, namely, *subpersonal functional dispositions*? Are not PSM-actions, as introduced in this paper, a very interesting special case, precisely because their functional analysis leads us directly into the more fine-grained causal landscape of action control, showing how central elements of the fully embodied "ability" for action control can gradually be dissociated and substituted by entirely subpersonal, technical building blocks? If it is true that intentional inhibition is simply an extra layer of control in a more global functional architecture for motor control, but at the same time also a process that automatically emerges "bottom-up", out of deterministic chaos, and via complex, but completely *agent-free*, dynamical processes of self-organization - does not intellectual honesty demand that we eventually give up the personal-level description for all academic purposes? Given the relevant sets of functional dispositions or boundary conditions for the relevant types of dynamical self-organization, must we not gradually move on to more parsimonious conceptual frameworks as potentially offered by robotics or neuroscience? Put differently: What exactly is the independent argument that legal theorists can simply go on and on and on, clinging to the traditional personal-level idiom while ignoring relevant empirical re-

search or current philosophical debates? Is it really *more* than a social convention? This is one aspect of what I mean by the "risk of superficiality". In philosophy, there are extended and unresolved technical debates about the nature and metaphysical status of dispositions and "abilities"[56], but it is not at all clear if any of the proposals developed in these debates map onto our folk-intuitions regarding an individual in a deterministic world really "having" a certain ability, even when this ability is not realized or exerted, because in every given concrete situation the individual could never have acted otherwise. Too much legal theorizing seems to be guided by such folk-intuitions. Therefore, another important desideratum for future research is to further clarify the relationship between personal and subpersonal levels of analysis[57] – and not just to introduce a pseudo-intuitive *façon de parler* as a new way of speaking with which everybody can live.

In conclusion, robot ethics and robot law need an empirically grounded theory of autonomy which is applicable to coupled man-machine systems and new forms of action control in virtual reality, for example to the class of actions marked out as "PSM-actions". Today, this involves not only neuroscientific research, but also computational modeling and conceptual analysis. Perhaps the most relevant goal for the future lies in building conceptual bridges connecting those abstract models of autonomy and rational self-control that are used in philosophy and legal theory with equally abstract, but *data-driven* computational models of the relevant abilities as described in the neuroscience of volition and action control.

Work on this publication was supported by the European Union FP7 Integrated Project VERE (No 657295).

References

Aleksander, I., Machine consciousness, Scholarpedia, 3 (2):4162, 2008.

Beckermann, A., Neuronale Determiniertheit und Freiheit, in: *Köchy K./Stederoth, D.* (eds.), Willensfreiheit als interdisziplinäres Problem, Freiburg im Breisgau, 2006, pp. 289-304.

Benatar, D., Better Never to Have Been: The Harm of Coming Into Existence, Oxford, 2006.

Blanke, O., Multisensory brain mechanisms of bodily self-consciousness, Nature Reviews Neuroscience, Vol. 13, 2012, pp. 556–571.

Blanke, O./Metzinger, T., Full-body illusions and minimal phenomenal selfhood, Trends in Cognitive Sciences, Vol. 13 (1), 2009, pp. 7-13.

56 *Maier*, "Abilities", in: *Zalta* (ed.), The Stanford Encyclopedia of Philosophy (Fall 2011 Edition), <http://plato.stanford.edu/archives/fall2011/entries/abilities/> (accessed 23 October 2012).

57 See *Dennett*, Content and Consciousness, 1969.

Brass, M./Haggard, P., To do or not to do: The neural signature of self-control, Journal of Neuroscience, Vol. 27, 2007, pp. 9141-1945.

Campbell-Meiklejohn, D./Woolrich, M./Passingham, R. /Rogers, R., Knowing when to stop: the brain mechanisms of chasing losses, Biological Psychiatry, Vol. 36, 2008, pp. 293-300.

Chrisley R./Parthemore, J., Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience, Journal of Consciousness Studies, Vol. 14 (7), 2007 , pp. 44-58.

Cohen, O./Druon, S./Lengagne, S./Mendelsohn, A./Malach, R./Kheddar, A./Friedman, D., MRI-based robotic embodiment: A pilot study, IEEE International Conference on Biomedical Robotics and Biomechatronics, June 24-27, Roma, Italy, 2012.

Damasio, A.R., The Feeling of What Happens: Body and Emotion in the Making of Consciousness, New York, 1999.

Della-Sala, S./Marchetti, C., Anarchic Hand, in: *Freund, H.-J./Jeannerod, M./Hallett, M./Leiguarda, R.* (eds.), Higher-order Motor Disorders: From Neuronanatomy and Neurobiology to Clincial Neurology, Oxford, 2005.

Dennett, D., Content and Consciousness, London, 1969.

Desmurget, M./Reilly, K./Richard, N./Szathmari, A./Mottolese, C./Sirigu, A., Movement intention after parietal cortex stimulation in humans, Science, Vol. 324, 2009, pp. 811-813.

Filevich, E./Kühn, S./Haggard, P., Intentional inhibition in human action: The power of 'no', Neuroscience and Neurobehavioral Reviews, Vol. 36, 2012, pp. 1107-1118.

Filevich, E./Kühn, S./Haggard, P., Negative motor phenomenal in cortical stimulation: implications for inhibitory control of human action, Cortex, Vol. 48, 2012, pp. 1251-1261.

Fricke, F., Verschiedene Versionen des negative Utilitarismus, KRITERION, Vol. 15, 2002, pp. 13-27.

Friston, K., The free-energy principle: a unified brain theory?, Nature Reviews Neuroscience, Vol. 11 (2), 2010, pp. 127-138.

Friston, K./Stephan K., Free energy and the brain, Synthese, Vol. 159 (3), 2007, pp. 417-458.

Gamez, D., Progress in machine consciousness, Consciousness and Cognition, Volume 17 (3), 2008, pp. 887-910.

Haggard, P./Eimer, M., On the relation between brain potentials and the awareness of voluntary movements, Experimental Brain Research, Vol. 126, 1999, pp. 128-133.

Harnad, S., The symbol grounding problem, Physica, D 42, 1990, pp. 335–346.

Holland, O./Goodman, R., Robots with internal models: A route to machine consciousness?, in *Holland, O.* (ed.), Machine Consciousness, Exeter, 2003.

Holland, O./Knight R./Newcombe, R., A robot-based approach to machine consciousness, in: *Chella, A./Manzotti, R.* (eds.), Artificial Consciousness, 2007, pp. 887-910.

Howhy, J., The Predictive Mind, Oxford, 2013.

Kadlec, E., Popper's "Negative Utilitarianism": From Utopia to Reality, in *Markl, P./Kadlec. E.* (eds.), Karl Popper's Response to 1938, 2008, pp. 107-121.

Kühn, S./Haggard, P./Brass, M., Intentional inhibition: how the "veto-area" exerts control, Human Brain Mapping, Vol. 30, 2009, pp. 2834-2843.

Libet B./Wright E./Feinstein B./Pearl D., Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man, Brain, Vol. 102, 1979, pp. 191-222.

Libet, B, Unconscious cerebral initiative and the role of conscious will in voluntary action, Behavioral and Brain Sciences, Vol. 8, 1985, pp. 529-566.

Libet, B., Do we have free will?, Journal of Consciousness Studies, Vol. 6 (8-9), 1999, pp. 47-57.

Libet, B., Neural time factors in conscious and unconscious mental functions, in: *Hameroff, S./Kaszniak, A./Scott, A.* (eds.), Towards a Science of Consciousness, Cambrindge, 1996.

Libet, B., The neural time factor in conscious and unconscious events, Ciba Foundation Symposium, 1993.

Libet, B., The neural time-factor in perception, volition, and free will, Rev de Metaphysique et de Morale, Vol. 97, 1992, pp. 255-272.

Libet, B., The Timing of Mental Events: Libet's Experimental Findings and Their Implications, Consciousness and Cognition, Vol. 11, 2002, pp. 291-299.

Libet, B./Alberts W./Wright E./Delattre L./Levin G./Feinstein B., Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex, Journal of Neurophysiology, Vol. 27, 1964, pp. 546-578.

Libet, B./Alberts, W./Wright, E./Feinstein, B., Responses of human somatosensory cortex to stimuli below threshold for conscious sensation, Science, Vol. 158, 1967, pp. 1597-1600.

Libet, B./Gleason, C./Wright E./Pearl D., Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): The unconscious initiation of a freely voluntary act, Brain, Vol. 106, 1983, pp. 623-642.

Metzinger, T., Being No One. The Self-Model Theory of Subjectivity. Cambridge, [2. durchgesehene Auflage als Paperback September 2004; elektronisch 2011].

Metzinger, T., Conscious Experience, Paderborn, 1995.

Metzinger, T., First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood, in: *Shapiro, L.* (ed.), The Routledge Handbook of Embodied Cognition, London, 2013.

Metzinger, T., Grundkurs Philosophie des Geistes. Band 1: Phänomenales Bewusstsein, Paderborn, [2. durchgesehene und erweiterte Auflage] 2009.

Metzinger, T., Grundkurs Philosophie des Geistes. Band 2: Das Leib-Seele-Problem, Paderborn, 2007.

Metzinger, T., Grundkurs Philosophie des Geistes. Band 3: Intentionalität und mentale Repräsentation. Paderborn, 2010.

Metzinger, T., Neural Correlates of Consciousness: Empirical and Conceptual Questions, Cambridge, 2000.

Metzinger, T., Reply to Gallagher: Different conceptions of embodiment, PSYCHE – An Interdisciplinary Journal of Research on Consciousness, Vol. 12 (4), 2006.

Metzinger, T., The Ego Tunnel. The Science of the Mind and the Myth of the Self, New York, 2009.

Owen, A./Coleman, R./Boly, M./Davis, M.H./Laureys, S./Pickard, J.D., Detecting awareness in the vegetative state, Science, Vol. 313 (5792), 2006, p. 1402.

Parvizi, J./Damasio, A., Consciousness and the brainstem, Cognition, Vol. 79, 2001, pp. 135-159.

Popper, K, The Open Society and Its Enemies, Volume I, London, 1945.

Putnam, H., Mind, Language and Reality, Philosophical Papers, Vol. 2, Cambridge, 1975.

Schooler, S./Salau B. /Julien, M./Ives, A.: Alternative stable states explain unpredictable biological control of Salvinia molesta in Kakadu, Nature, Vol. 470, 2011, pp. 86–89.

Singer, P., Practical Ethics, Cambridge, 2011.

Steels, L, The symbol grounding problem has been solved. So what's next?, in: *de Vega* (ed.), Symbols and Embodiment: Debates on Meaning and Cognition, Oxford, 2008.

Von Hippel, W./Trivers, R., The evolution and psychology of self-deception, Behavioral and Brain Sciences, Vol. 34, 2011, pp. 1–56.

Zalta, E. (ed.), The Stanford Encyclopedia of Philosophy, Stanford, 2011, <http://plato.stanford.edu/>.

This is the penultimate draft of the following paper:

Metzinger, T. (2013).Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Hrsg.), *Robotik und Gesetzgebung*. Baden-Baden: Nomos. S. 263-302.

metzinger@uni-mainz.de
www.philosophie.uni-mainz.de/metzinger/