Thomas Metzinger

What exactly is The Will?
From Robotic Re-Embodiment to Consciousness Ethics

Why is it that so many intelligent and educated people think that giving up on the traditional idea of free will is something extremely dangerous, something that we should never do – even if we must sacrifice our very own intellectual honesty for it? Is it more than reactionary resentment? I think, although many of these intelligent and educated people do not have good arguments, they have good intuitions: We may all have to pay a price for progress in the new Mind Sciences, in philosophy, cognitive science, and brain research. That price may be high. But the purely ideological resentment against naturalism as it is cultivated in certain parts of academic philosophy and the humanities will not help us to minimize the price we will have to pay mid-term for the naturalistic turn in the image of man. We have to face the facts, in an intellectually honest, rational, and evidence-based manner.

Is it possible to believe that "ultimate origination", the ability, given identical physical boundary conditions, to do otherwise does not exist? What would it mean that the ability to choose a course of action from among various alternatives not even for one's own mental actions, i.e., for rational thought and the intentional control of one's own attention? To be able to honestly take this to be true, our conscious self-model would have to change in a dramatic way. The phenomenal experience of ownership and the phenomenal experience of agency are intimately related, and both are important aspects of the conscious sense of self. If you lose control over your actions, your sense of self is greatly diminished. This is also true of inner actions; for example, many schizophrenics feel that not only their bodies but even their thoughts are controlled by alien forces. Could a naturalist turn in the image of man be a danger to our mental health?

A second, equally important anti-naturalist intuition is that free will is not only something in the mind, but also something that evolved in culture, a background assumption that underlies social cohesion – that is, a very precious social institution. What many researchers in the humanities often do not know yet is that by now there are already first empirical studies actually showing how a reduced belief in one's own free will can lead to a demonstrable attenuation of willingness to help, to an increase in willingness to cheat, to reduced self-control, a weaker reaction to one's own mistakes and a boost in aggression. Objective changes can experimentally even be shown to exist in the neural correlates of unconscious pre-stages of voluntary acts. The self-model theory of subjectivity can explain why this is so: the conscious, cognitive self-model is deeply anchored in our unconscious image of ourselves, and therefore shifts in the phenomenal self-model can – just as in psychosomatic illnesses – have direct and sustained causal effects for the inner state of the body and our outer behaviour. Therefore, if a vulgar materialistic doubt regarding one's own free will spreads within society this could lead to antisocial tendencies, to more impulsive and reckless behaviour. There is a complex psycho-social risk, and it is definitely real.

Let us now look at some recent technological developments that give us another angle on the question of the will: the embedding of the human mind into new medial environments. I have described the development in the revised 2014 version of my German book Der Ego-Tunnel (all references can also be found there) and will only present one single example here.

Robotic-Re-Embodiment

In Chapter Three of Der Ego-Tunnel I discussed the classical study on full-body illusions from the year 2007. Although the effect of these first experiments was rather weak, there have been many interesting variations of the basic idea and a flood of new scientific studies have been published since then. A second example I presented was the "Heart-Experiment" by Jane Aspell and Lukas Heydrich, with which we already became acquainted a while ago. Here is a third example, and one that is not only of theoretical interest for philosophers of the mind. It also demonstrates what I meant by claiming that some of those new technologies will be consciousness technologies, touching the very core of our self-conception, namely because they interact with our self-model in a very direct way.

The self-model theory is not simply one philosophical model among others. It has been laid out as an interdisciplinary research program right from the beginning, as firmly anchored in scientific data as possible. If the basic idea of the self-model theory is on the right track, it yields a whole range of empirical predictions that would have to be experimentally testable. One of these predictions is that it must in principle be possible to directly connect the conscious self-model in the human brain to external systems – for instance to computers, robots, or also to artificial body images on the Internet or in virtual realities. This prediction has recently been corroborated. In recent years, there has been great progress in the field of so-called Brain-Computer Interfaces (BCIs), and this progress allows us to investigate the empirical aspects of the self-model theory in more detail.

Special about such brain-computer interfaces is that a connection between a brain and a computer can be established without activating the peripheral nervous system, the body, or any limbs. New ways to act in the world emerge. Paralyzed persons can, for instance, operate robot arms or painting software with the "power of their minds", healthy persons have already sent Twitter messages directly out of their own brain or even spelled words in groups. To do this, one either records electrical activity (for instance using EEG or implanted electrodes), or one measures certain properties of cerebral blood flow (for instance using functional magnetic resonance imaging or nanoscale infrared spectroscopy). These measurements are then analysed with the aid of computers and the found patterns are converted into control signals. This technical development is philosophically interesting for a number of reasons, for not only does it enable us to act in the

world, to a large extent, "bypassing the biological body", but also to test theories about the emergence of the sense of selfhood more precisely than ever before. Many of these developments are historically new.

Another empirical prediction under the conceptual assumptions of the self-model theory is also that it must in principle be possible to couple the human self-model in a causally-direct way with artificial organs for acting and sensing while bypassing the non-neural, biological body. Through this, we could not only experientially, but also functionally, situate ourselves in technologically-generated environments in completely novel ways. For five years, I have been working in a research project funded by the European Union, the VERE project, in cooperation with scientists and philosophers from nine countries. One of the research goals of this ambitious project was to go beyond the classical experiments from the year 2007 and stably transfer our sense of selfhood to avatars or robots that can perceive for us, move, and interact with other self-aware agents (VERE is the acronym of Virtual Embodiment and Robotic Re-Embodiment). My official philosophical position still says that we will never really succeed in this. I believe that gut feelings, the sense of balance, and spatial self-perception are so firmly coupled to our biological body that we will never be able to leave it experientially on a permanent basis. The human self-model is anchored on interoception; it cannot simply be "copied out" of the brain. But I must confess that I am starting to have doubts. For firstly, it could be that simply different and newly extended forms of self-consciousness could in the future be generated by ever more densely couplings between self-model and avatars or robots – and secondly, technological progress in this area happens surprisingly fast.

In an ambitious pilot study, our Israeli colleagues Ori Cohen, Doron Friedman and their collaborators in France demonstrated that it is possible to read out action intentions of a test subject using real-time functional magnetic resonance imaging. These can then directly be transferred as high-level motor commands to a humanoid robot, which transforms them into bodily actions, while the test subject can simultaneously witness the whole experiment visually through the eyes of the robot. [fn 6] This process is based on wilfully generated motor imagery, allowing test subjects to "directly act with their PSM", [fn 7] by remote-controlling a humanoid robot in France from a scanner in Israel.

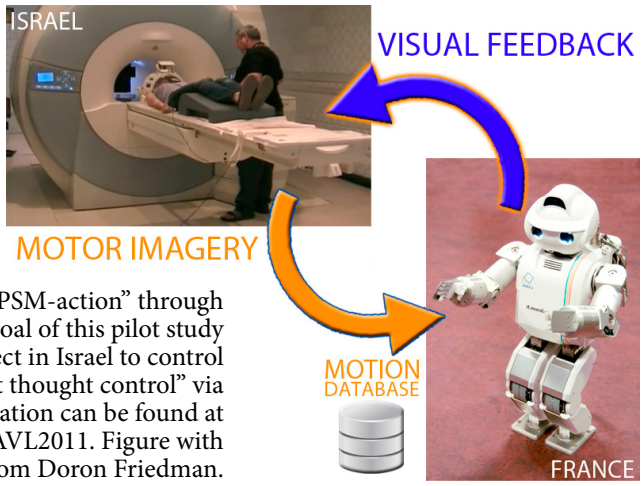

Figure 1: An example of direct "PSM-action" through robotic re-embodiment: The goal of this pilot study consisted in enabling a test subject in Israel to control a robot in France through "direct thought control" via the Internet. A video demonstration can be found at http://www.youtube.com/user/TheAVL2011. Figure with friendly permission from Doron Friedman.

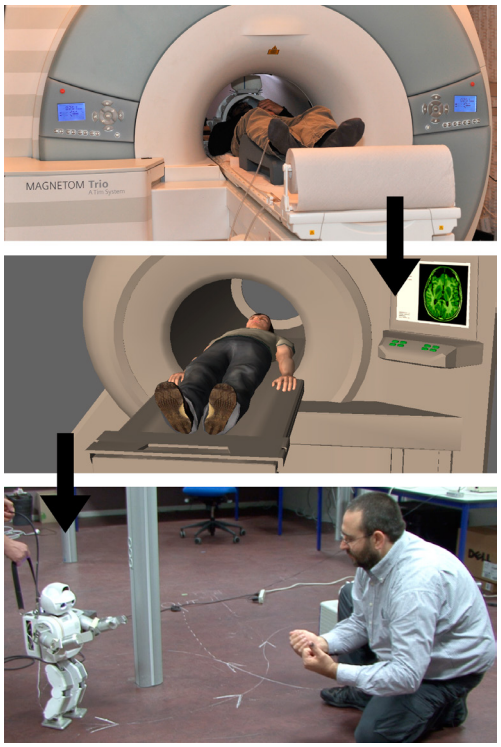

Figure 2: "PSM-actions": A test subject lies in a nuclear magnetic resonance tomograph at the Weizmann Institute in Israel. With the aid of data goggles he sees an avatar, also lying in

For philosophers, this technological development is interesting for several reasons: firstly, because of its ethical consequences, but also because it constitutes a historically new form of acting. I have introduced the notion of a "PSM-action" to be able to describe this new element more precisely. PSM-actions are all those actions in which a human being exclusively uses the conscious self-model in his brain to initiate an action. Of course, there will have to be feedback loops for complex actions, for instance, when seeing through the camera eyes of a robot, perhaps adjusting a grasping movement in real-time (which is still far from possible today). But the relevant causal starting point of the entire action is now not the body made of flesh and bones anymore but only the conscious self-model in our brain. We simulate an action in the self-model, in the inner image of our body, and a machine performs it.

a scanner. The goal is to create the illusion that she is embodied in this avatar. The test subject's motor imagery is classified and translated into movement commands, setting the avatar in motion. After a training phase, test subjects were able to control a far remote robot in France "directly with their minds" via the Internet, while they were able to see the environment in France through the robot's camera eyes. Figure with friendly permission from Doron Friedmann.

What can be learnt from such experiments? On the one hand, it is obvious that the phenomenal self-model often is a crucial part of a control hierarchy: it is an abstract tool. The PSM is a means to predict and monitor certain critical aspects of the process in which the organism generates flexible, adaptive patterns of behaviour. On the other hand, it is highly plastic, because several representations of objects external to the body can transiently be integrated into the self-model. A hammer or pliers could be such an object. For the principle of embeddability not only applies to rubber hands, as we already seen in the introduction. It even applies to tools in general – extensions of bodily organs that need to be controlled to generate intelligent and goal-directed behaviour. The self-model is the functional window through which the brain can interact with the body and vice versa.
When the body is extended by sticks, stones, rakes, or robot arms, the self-model has to be extended as well. Only if an integrated representation of the body-plus-tool exists, can the extended system of body-plus-tool in its entirety become part of the brain's control hierarchy. Asked differently: How else could one learn to intelligently – i.e., flexibly and in a context-sensitive manner – use a tool, without integrating it into the conscious self? The conscious self-model is a virtual organ, allowing us to own feedback loops and to initiate control processes, and to maintain and flexibly adapt them.

The phenomenal experience of ownership mirrors the respective hypotheses about those parts of reality we can possibly causally control at the moment. Some element of the control circuit are physical (like the brain and tools), others are virtual (like the self-model and the goal-state simulation). Robots are tools. Therefore it is possible to transiently embed entire robots or virtual bodies (avatars) into the PSM and thereby causally control them. Above I already pointed out how human beings (and also some other animals) often seek to control the behaviour or mental states of other persons. We "instrumentalize" and "seize" each other, sometimes we even turn each other into "bond-slaves" [Leibeigene as we say in German]. Human beings continuously try to extend their domain of influence – not only with sticks, stones, rakes, or robot arms, but also with brains and bodies of other human beings. In contrast, the classical theme of "possession" could also describe a robot, remote-controlled directly out of one's own mind. The robot is transiently "possessed"; however not by a devil or a demon but by something perhaps far worse – by a self-aware human being.

The anarchic robot

The neurological example of the "alien hand" demonstrates how it is possible that a human body part can carry out intelligent and goal-directed actions, although the respective patient does not deliberately initiate these actions and does not have the feeling that they are his actions. How would it feel if an avatar or a robot, with which you transiently identify via your self-model, suddenly did something you did not want to do?

Imagine you are laying on your back in a brain scanner, remote-controlling a robot, while you are seeing through its eyes and even feeling the motor feedback from its arms and legs while they are moving. Experientially, you completely identify with the robot, while at the same time you are moving freely in a situation in which also other human beings are present. Suddenly the new husband of your divorced wife enters the room. He is the person who destroyed all your plans and your entire personal life a few months ago. You are feeling the mortification again, the deep hurt, the sense of inner emptiness and existential loneliness following the divorce. Spontaneously an aggressive impulse arises inside of you, and almost simultaneously a brief, violent fantasy emerges inside of you. You are trying to calm yourself down – but before you can suppress the motor imagery that involuntarily went along with the violent fantasy in your conscious mind, the robot has already killed the man with one single, forceful blow. Now you regain control and are able to back out a few steps. Subjectively it feels as if you never had a chance to control your behaviour. But how can one decide if you – from a purely objective perspective – perhaps still possessed the capability of suppressing the aggressive impulse, just in time? In an ethical sense, are you responsible for the consequences of the robot's actions?

It is quite possible that at first embodiment in a robot or an avatar will only be "shallow" embodiment, because it does not provide us with the same type or level of autonomy as the biological body does. Perhaps our capacity for impulse control is weaker, or slightly less precise; perhaps we lack what I called "Veto Autonomy" elsewhere: the ability to suspend or terminate a planned, voluntary, or even already commenced bodily action within a certain time window. This leads us to the modern discussion about the wandering mind and the notion of "mental autonomy". Attentional lapses, "zoning out" and spontaneous episodes of mind wandering could be much more dangerous if our self-models are directly coupled to artificial media and novel types of tools for action. If this is so, responsibility or accountability could mean something different for human agents in virtual reality or when merging into a robot than it does in what, today, we call "normal life". First and foremost, it will be crucial to identify possible risks as early as possible and to take protective measures in due time.

Apart from this ethical dimension there is second important aspect: Directly coupling a human PSM with an artificial environment is an example for a new type of consciousness technology. Currently the effects are still weak and there are many technical problems. However, it is not entirely impossible that technological progress happens faster than expected. What would we do if systems for virtual or robotic re-embodiment were one day really functioning fluidly, with many degrees of freedom, and in real-time? What new conscious states would become possible if one was also able to control the feedback with the help of computer-aided direct brain stimulation, directly aimed at the users' self-model, again bypassing the biological body? What new forms of intersubjectivity and social cooperation could emerge if it was suddenly possible to connect several human persons and their self-models simultaneously via coupled brain computer interfaces, and perhaps even to merge them? Could we lose control of our own minds? Do we have control right now?

The Wandering Mind

While reading this book, how often did you suddenly notice that you had zoned out some time ago, whilst your eyes automatically kept following the lines, although you did not grasp their contents? How often does it happen to you that you fall into spontaneous daydreams during routine activities, or that you cannot fall asleep because you are plagued by compulsively reappearing thoughts – unfortunately, mostly by those that have a negative content? Do you sometimes go on unintentional time travels, for instance when you are waiting in your car at a red traffic light and are unexpectedly haunted by unbidden memories, or automatically beginning to plan the upcoming shopping or your next vacation?

One of the most exciting recent research fields in psychology is mind wandering. Our mind wanders. It wanders more frequently than most of us think, namely up to 50 percent of our waking life – and we pay a high price for it. Scientific studies have shown that the spontaneously wandering mind has a clearly measurable negative effect on text comprehension and success at school but also on learning success, sustained attention, and a student's memory capacity. Mind wandering has a negative influence on the stability of our mental "working memory" and our mathematical abilities, but also on our safety during car driving and a host of other activities, for which staying in contact with the "Now" is important.

One interesting finding of recent studies is that the wandering mind makes us unhappy: Someone who loses contact to the present because he zones out repeatedly into the future or past, generally has a worse mood than human beings who can keep their attention more strongly in the present. On the other hand, not all forms of mental absence are the same. Some forms of daydreaming or of spontaneous, stimulus- or task-unrelated thought seem to have positive sides, as well. For example, there is preliminary evidence that they play an important role in autobiographical planning, creative problem-solving, and in some forms of goal-directed thinking or perhaps even deeper forms of self-reflection.

When our mind wanders, we lose our mental autonomy. Mental autonomy is the ability to control one's own inner actions and to act on the mental level in a self-determined way, selecting one's own goals and being able to hold on to them. This also includes being able to suspend or terminate a mental action, or to intentionally inhibit some automatic inner behaviour. We lose our mental autonomy every time a certain part of our cognitive self-model transiently breaks down – and recent research shows that this happens to all of us several hundred times a day. I call this layer in the conscious self the "epistemic agent model": Our inner image of ourselves as a "knowing self", as an entity actively constructing knowledge relations to the world and itself. If we lost control on the level of bodily action as often as we do on the mental level, we would often – viewed from the outside – look like an odd mixture of an awakened person and a sleep-walking fidget. A sleepwalker errs on autopilot in a bizarre zigzag motion through the world. Like a kind of play-acting robot, he seems to enact a multitude of unknown, but evidently competing short stories and inner dramas.

In the course of it, he is magically attracted to ever new objects, but forgets them soon and continues his zigzag course. In particular, he constantly falls down, starts to struggle like a new-born child who is not yet able to walk. But then he suddenly rises again and briefly perceives the current moment: He becomes present, a person, an autonomous mental subject. As soon as the present does not demand his full attention any longer, the sleepwalker takes over again, tumbling through the world without really feeling himself, without being in touch with himself.

Our wandering mind possesses a whole series of interesting phenomenological aspects. Have you ever noticed that you can never perceive the actual emergence of the first thought taking you from the Now into a daydream or an inner monologue, but that you can at most – if you are very conscious – perceive the second thought, following up and developing from the first thought? I have coined the term self-representational blink for this interesting fact, i.e., for the short moment of inner blindness resulting from this "blink" through which the brain switches from one self-model to another – the blink of the eye of self-consciousness. Every mind-wandering episode begins with a collapse of the knowing self, the "epistemic agent model" in your brain. I predict that there is a detectable gap of self-blindness in between. A second interesting observation is that we cannot volitionally terminate a train of thought or inner story as long as we completely identify with it. We are actually lost. A particular part of the self-model has broken down, viz. the conscious knowledge that we ourselves possess the ability to terminate the state and return to the Now in the first place. We would be able to act, but in this moment we no longer know at all that we are a being who is capable of autonomous, inner action in the first place. One of the most important functions of our conscious self-model is to make a specific form of knowledge available, namely what abilities and opportunities for action we currently possess. Someone who does not know that they could stop cannot stop.

Inner action, inner non-action and mental autonomy

What exactly is this process we call "conscious thinking" in the first place? Conscious thinking exists, for instance, also during the night, in states of dreaming. During dreams, we possess no control whatsoever over our thoughts and we are not be able to control our attention volitionally. In the following chapter, we will see that sometimes there is the possibility to "awake" within a state of dreaming and regain mental autonomy. Such dreams are called "lucid dreams", for in such dreams the dreamer realizes that he is currently dreaming and hence he also regains control over thinking processes and the ability for volitional control of attention. In one of my scientific publications I have shown that we are only rarely awake in this sense during the day and that we are not mentally autonomous beings for more than two thirds of our conscious lives, either.

Depending on the scientific study, our mind wanders during 30-50% of our conscious waking phases. At night,

during our non-lucid dreams and those sleeping stages in which we have complex conscious thought but no pictorial hallucinations, we also lack the ability to suspend or terminate the thinking process – an ability of central importance for mental self-control. Then there are also various types of intoxication or light anaesthesia, of illness (e.g., fever dreams or depressive rumination), or of insomnia, in which we are in a sort of helpless twilight state, plagued by constantly reoccurring thoughts we cannot stop. In all these phases our mind wanders and we have no control over our thinking processes or our attention. According to a conservative estimate, the part of our self-model that endows us with real mental autonomy only exists during around one third of our entire conscious life. We do not exactly know when and how children first develop the necessary capacities and layers of their self-model. But it is a plausible assumption that many of us gradually lose it towards the end of our lives. If we consider all empirical findings regarding mind wandering together, we arrive at a surprising result that can hardly be underestimated as far as its philosophical significance is concerned: Mental autonomy is the exception; loss of control is the rule.

As far as inner action is concerned, we are only rarely truly self-determined persons; for the major part of our conscious mental activity rather is an automatic, unintentional form of behaviour on the sub-personal level. In short: cognitive agency and attentional agency are not the normal case but rather an exception; what we used to call "conscious thinking" is actually most of the time an automatically unfolding sub-personal process.

When you are simply observing your breath, you perceive an automatically-unfolding process in your body. By contrast, when you are observing your wandering mind, you are also experiencing the spontaneous activity of a process in your body. What physical process is that, exactly? A multitude of empirical studies show that areas of our brains responsible for the wandering mind overlap to a large extent with the so-called default-mode network. The default-mode network typically becomes active during periods of rest, and as a result, attention is directed to the inside. This is what happens, for instance, during day dreams, unbidden memories, or when we are thinking about ourselves and the future. As soon as a concrete task needs to be done, this part of our brain is deactivated and we concentrate immediately on the solution to the currently pending problem. My own hypothesis is that the default-mode network mainly serves to keep our autobiographical self-model stable and in good shape: like an automatic maintenance program, it generates ever new stories, which all have the function to make us believe that we are actually the same person over time. Only as long as we believe in our own identity over time, does it make sense for us to make future plans, to avoid risks, and to treat our fellow human beings fairly – for the consequences of our actions will, in the end, always concern ourselves.

My hypothesis is that exactly this was one of the central conditions in the evolution of social cooperation and the emergence of large human societies: It is yourself who will be punished or rewarded in the future; it is yourself who will either enjoy a good reputation in the future or be retaliated upon. What we need for that is an intact "narrative self-model", an illusion of sameness.

But on closer inspection, the narrative default-mode does not, I believe, actually produce thoughts but something I would describe as "cognitive affordances" because they afford an opportunity for inner action. They actually are precursors of thoughts, spontaneously occurring mental contents, that, as it were, are constantly calling out "Think me!" to us. Interestingly, such proto-thoughts also possess something like the "affordance character" just mentioned, they reveal a possibility. That possibility is not a property of the conscious self and not a property of the little proto-thought currently arising – it is the possibility of establishing a relation by identifying with it. Do you recall the example involving your favourite chocolate cookies? If we are capable of rejecting such offers or to postpone them into the future then we can also concentrate on that which we currently want to do. Now exactly the same principle also holds for our inner actions: if we lose the ability in question for a single moment only, we are immediately being hijacked by an aggressive little "Think me!" and our mind begins to wander. Often our wandering mind then automatically follows an inner emotional landscape. It will try, for instance, to flee from unpleasant bodily perceptions and feelings and somehow reach a state that feels better, like a monkey brachiating from branch to branch. Not acting, it seems, is one of the most important human capacities whatsoever, for it is the basic requirement on all higher forms of autonomy. There is outer non-acting, for instance in successful impulse control ("I will not grasp for this bowl of chocolate cookies now!"). And there is inner non-acting, exemplified by the letting go of a train of thought and resting in an open, effortless state of awareness, which can sometimes follow. There is thus an outer and an inner silence. Someone who cannot stop his outer flow of words will soon be unable to communicate with other human beings at all. Whoever loses the capability for inner silence loses contact to himself and soon will not be able to think clearly any more.

What is a good conscious state?

Let us assume there is life after death. This after-life is temporally unbounded, it lasts eternally and within it conscious experience continues to exist. However, there is an important difference: all conscious experiences after death are experiences you were permitted to choose from the set of those subjective experiences you made in your current life – because after death there are no new experiences anymore. Before death, by contrast, you lived through a large number of inner experiences and conscious states, and some of them were actively made by you – for instance by going to the movies or by taking a hike, by reading books, taking certain drugs, or by participating in a meditation retreat. And in fact, for the major part of our lives we are busy seeking, in one way or another, conscious states we would experience as pleasant or valuable. Let us say that the smallest unit of conscious experience is always one single subjective moment because if we look carefully we find that we are always living our life through conscious moments. In doing so, most of us are always looking for the "meaningful now", for those small "perfect" moments of happiness or an experience of meaning.

Our introductory thought experiment now consists of an idea and a question. The idea is that you are allowed to select exactly which conscious moments from your finite life will be transferred to a "playlist for eternity": After your death all subjective experiences on this list will be replayed again and again, in random order. This process then creates your very own and personal conscious eternity. During your lifetime you are like a phenomenological Cinderella calling out to the turtle doves: "The good into the pot, the bad into the crop!" And now the question: if you were permitted to make this irrevocable selection all by yourself, if it really was only yourself who could pick the good grains from the ashes of transience, into which the bad stepmother had thrown them, which moments would that be? And most importantly: how many moments, according to your own criteria, would you actually rank as truly worth living – in the sense of worth being re-lived?

At Johannes Gutenberg University in Mainz, we began with a first series of small pilot studies with advanced philosophy students. David Baßler programmed an SMS server in such a way that, for seven days, it sent ten signals a day at random points in time to the participants, whose cell phones would then briefly vibrate. The participants' task was to decide whether the last moment before the conscious experience of the vibration was a moment they would take with them into life after death. For many, the result was surprising: the number of positive conscious moments per week varied between 0 and 36, the average was at 11.8 or almost 31% of the phenomenological samples, while at 69% a little more than two thirds of the moments were spontaneously ranked as not worth reliving. If one takes the idea of consciousness ethics and our question about the nature of valuable conscious states really seriously, one first has to conceptually distinguish between the subjective and the objective value of a conscious moment. It would be conceivable that an objectively valuable subjective conscious experience – for instance a painful learning experience in the external world or a deep inner insight into a permanently recurring form of self-deception – would be subjectively perceived as unattractive and worthless. Conversely, there could be states that would subjectively be perceived as extremely meaningful that would appear as completely worthless from a critical third-person perspective – for example, certain states induced by psychoactive substances or deeper delusional states caused by ideological indoctrination. In our pilot studies we were primarily interested in gaining a better understanding of the mechanism through which we subjectively experience conscious moments as pleasant or valuable. In doing so we found it important to find a fine-grained and maximally simple form of assessment which would always register (grasp?) the current moment only, and as independently as possible from philosophical theories, ideologies, and conceptual presumptions. For example, in a second study we dropped the afterlife-assumption and the "eternity condition", replacing them with the following question: "Would you like to relive the very last conscious moment in this life?" Interestingly, under this condition only a little over 28% of life-moments will be ranked as positive and just below 72% are those one would actually not like to relive.

Recent research findings show that very many animals are able to suffer, because they possess a conscious self-model and that our current way of treating animals cannot be ethically justified in any way. But how do we find out whether a self-aware animal that is not able to talk to us experiences certain husbandry conditions as aversive, as a form of negative experience or not? The answer is simple: One verifies if the animal, when given the choice, would voluntarily enter the same state again. But if we now pose this question with respect to the long chain of conscious moments in our own lives, and if in doing so we are really attentive and honest to ourselves, we will make two surprising phenomenological observations. These observations are philosophically interesting: Firstly, it seems that – although conscious moments of dramatic suffering are rather rare in our lives – on average, and on the most fine-grained level of observation, we would rate our own lives as not worth living. This holds at least in the very simple sense that we would not like to relive a clear majority of the moments constituting our conscious lives. Upon closer inspection, and using this purely subjective criterion, we would pick only very few "good grains" from the evil stepmother's ashes and it would even be of little help if – as in the fairy tale – "all birdies under the sky were to help us". The second interesting phenomenological fact is that this discovery really touches us only very briefly. Almost immediately, massive activity on the level of our cognitive and autobiographical self-model kicks in to stabilize our sense of self-esteem: "The real issue is not about isolated hedonic qualities at all, the value of conscious experience is determined by the overall context of my life, by my personal-level goals and desires in an extended temporal frame of reference!" we immediately tell ourselves. We begin to philosophize: "It is not about the average value or the point balance – only peak experiences really count!" we suddenly think, or: "Most conscious moments are actually neutral and not really aversive or even a form of suffering!" Perhaps we find: "Well, it is true that most moments in my life are affectively rather of negative tone or simply boring, but I am writing a dissertation that will contribute to the knowledge of mankind, and epistemic progress is much more important than a well-filled play list for eternity!" It is a bit like listening to the German Federal Government's spokesman who is declares that some debate is now over. Taking this second phenomenological fact seriously, an inconvenient conjecture suggests itself: Perhaps it is exactly one main function of the self-model's higher levels to continuously drive the organism forward, to generate a functionally adequate form of self-deception glossing over everyday life's ugly details by developing a grandiose and unrealistically optimistic inner story – a "narrative self-model"? Interestingly, we already encountered the notion of a "narrative self-model" in the fourth chapter, when we were examining the wandering mind more closely. I believe that there exists a deep inner connection between self-deception, the conscious experience of sameness over time, and our constantly wandering mind.

Of course, a whole whirlwind of technical philosophical issues arises at this point: If many moments of pleasure necessarily involve the new and surprising aspects of an experience, wouldn't this "novelty aspect" be missing in an eternal playback after death? How exactly could one

reinsert it without causing major damage? Would it also be permitted to just select the one and only very best conscious moment from your life and set it to endless auto-repeat? Does it make sense to investigate individual "snapshots of consciousness" without the narrative self-model at all, or could it be that any attempt to isolate and analyse single moments from the wider context of their greater temporal dynamics is misguided right from the beginning? Is there such a thing as introspective knowledge in the first place? Is it not true that every inner decision for or against repetition at least indirectly has to be very strongly contaminated by theories, and constantly shaped by our personal background assumptions? What reason at all do we have to trust our own normative judgments, highly subjective as they are? If for myself I have marked out a conscious experience as "positive" or "valuable" – why should I follow this intuition in the first place? Perhaps what really counts in life is not at all about what, as a matter of contingent fact, I happen to experience as "valuable" or to subjectively perceive as "worth living".

This leads to the question whether one could meaningfully talk about an "objective value" of certain conscious states. Personally, I do not believe that we could identify or gain knowledge about such objective values, or that we could even secure them by providing an ultimate justification. Rather, it is exactly this fact which is part of the problem we need to solve.

Consciousness Ethics

Nevertheless the idea of a "consciousness ethics" remains not only an important but an absolutely central objective for the future. But it has to be built on much weaker foundations. All we can do is to open-mindedly investigate how the systematic cultivation of certain classes of conscious states could improve our living together in society and whether, in the real world, it actually also achieves what the ultimate goal of the original ethical idea was. To begin a conversation and in order to have a starting point for future discussions I want to present three such values, which have the advantage that almost all human beings in the world can agree upon them. These three objectives are the reduction of suffering, self-knowledge, and increasing mental autonomy.

Reduction of suffering

Conscious suffering is probably more common than most of us are willing to admit. An important criterion for a good conscious state therefore is whether it reduces consciously experienced suffering – particularly in the future and in other beings capable of suffering as well. Let me illustrate and sketch the core idea by introducing a new working concept into our discussion: the "NP-footprint". Whether a conscious state is a good conscious state depends to a large extent on how big its NP-footprint is. "NP" stands for "negative phenomenology", i.e., for the class of all unpleasant or distressful conscious states. We can simply define them as those conscious states that a sentient being would rather not relive when given the choice. The notion of a footprint, by contrast, has already been known for a long time in environmental ethics: the "ecological footprint" is a simple metaphor and at the same time a conceptual tool that could in principle be differentiated further. It is an indicator of sustainability relating the consumption of resources to the Earth's biocapacity, a measure of human demand on the Earth's ecosystems. At the same time, the ecological footprint is not only computable for persons and households, but also for entire nations, and even products and services can be balanced with the ecological footprint. The ecological footprint is, in particular, an indicator of justice, for it is built on the basic assumption that all human beings should have the same at their disposal. One result for Germany: if all human beings lived like the Germans, we would need 2.8 Earths, for the German footprint is 5.09 hectare big. However, the just ecological footprint lies at 1.9 hectare. Thus, the ecological footprint is also something like a currency, with the help of which one can measure the demands on the biosphere, and indeed for all resources and all potential uses. I think we may need something quite similar for consciousness ethics. The ego tunnel is our inner environment, hence consciousness ethics is about something one could call "inner ecology".

The ethical principle of minimizing suffering states that we should continuously reduce negative phenomenal states in all conscious beings who are able to suffer, first by decreasing our own NP-footprint. When creating or cultivating a particular conscious state we should therefore always ask ourselves: Does it reduce my NP-footprint, or does it possibly increase the overall amount of suffering in the world? Does robotic re-embodiment produce good states of consciousness or bad ones? What NP-footprint does the psilocybin-state of consciousness have, how big is the NP-footprint of an alcohol-state of consciousness? How about the pleasant conscious states created when eating meat? A good action and a good conscious state then are those states that not only minimize suffering in the respective subject of experience but also in all other beings who are able to suffer. Thus, the most important question always is: How much conscious suffering does a given conscious state create not only in myself but also in other human beings, in animals that are able to suffer – but also in potential artificial subjects? Here, it is of highest relevance to include and factor in possible subjects of experience, i.e., future human persons, future animals that are able to suffer, and, as we have already seen, also about conceivable post-biotic systems like conscious robots and avatars. Their number - and the associated risk of doing harm – may often be much larger than we commonly think. In consciousness ethics we are therefore not only concerned with the NP-footprint I leave in my own life, but always with the one we leave in the conscious self-models of other beings – in the present as well as in the future.

Self-Knowledge

One of the positive sides of the new image of man consists in the enormous depth of our phenomenal state space. The number of a human being's possible conscious states is incredibly large. Only rarely are we aware of this fact, although our freedom of action is presently beginning to be extended by the new consciousness technologies. Most notably, we have not really started to systematically test altered states of consciousness for their epistemic potential.

We already saw that the scientific method of generating knowledge probably is not the only one. But if there are in fact forms of knowledge that are not expressible in the form of true sentences, what might they be? One possibility

is that they could consist in very specific abilities – for instance in the knowledge how to do something right. Long ago, the British philosopher Gilbert Ryle drew a distinction between knowing how and knowing that. This simple conceptual distinction could be of significance when it comes to meditation, increasing mental autonomy, and the epistemic potential of altered states of consciousness in general. As far as non-linguistic and non-intellectual forms of self-knowledge are concerned, we could in fact simply be dealing with certain abilities, abilities for inner action. The more such abilities a human being possesses, the greater the space of self-determined action with his own mind will be. The more such skills a person has learnt, the more novel forms of subjective experience are accessible to him or her. This point not only holds for the example of successful psychotherapy. Someone who has, for instance, learnt at a meditation course how to deal with inner restlessness, persistent self-doubts, or especially difficult emotions, has gained a new skill. This skill consists in a non-linguistic form of self-knowledge, and it potentially increases his or her inner autonomy. Knowing how is practical knowledge, and of course such knowledge also exists for inner actions. Someone who has learnt, at a meditation course or under the influence of a classical hallucinogen like psilocybin (which has already been mentioned), to see the fine, infinitely soft motion of leaves in the wind or the delicate shimmer in the flow of water, like Aldous Huxley did, as "a transience that was yet eternal life, a perpetual perishing that was at the same time pure Being", has simply learnt a new skill. This skill to perceive the quality of "timeless change" possibly consists in remembering a previous state and directing one's attention back to this particular aspect of one's perception. This could, for instance, have the effect that this person will, at least to some extent also in future and the rest of his or her life, have access to completely new forms of experiencing nature, without a meditation course and without the substance. In any case, he or she now for the first time possesses the knowledge that they have such abilities and inner options for action at their disposal in the first place. This also means that their self-model has changed in a highly relevant manner. "Consciousness culture" also means raising one's own mental autonomy, be expanding one's self-model, cultivating new capacities for inner action.

However, today, one has to see clearly that the ancient philosophical project of self-knowledge must be realized under dramatically changed constraints and boundary conditions. This is particularly true whenever we are interested in non-scientific forms of knowledge – i.e., those which are not mediated by language or theories – but in those much more subtle mental skills just mentioned. Classical mindfulness meditation is the paradigm example of such a skill.

Mental autonomy

We should take new scientific insights about the wandering mind seriously. Classical mindfulness meditation is the exact opposite of mind wandering and now we can finally see more clearly what the use of meditation techniques really is about: the central goal is a sustained enhancement of one's own mental autonomy. This point naturally leads to the main argument for the implementation of a systematic but fully secularized meditation training at our educational institutions: it is about something we might call raising the standard of civilization. But what exactly is a "standard of civilization"?

A country without the death penalty possesses a higher standard of civilization than a country with the death penalty. A nation in which there is no torture anymore has achieved a higher standard of civilization than a country in which governmental authorities deliberately inflict bodily or mental pain on human persons, for instance to intimidate, punish, or blackmail them to give evidence. As far as these two criteria are concerned, we can now very concretely point out how, for instance, China, Iran, or the USA possess a lower standard of civilization than for example Germany or any of the other 96 countries of the world that have already completely abolished the death penalty. China, Iran, and the USA also have a lower standard of civilization than those countries in which one does not torture anymore. However, the "degree of civility" that is relevant here can also be measured by considering how a country deals with animals, i.e., with non-human subjects who are able to suffer. Another indicator is to what extent it takes the interests of future, i.e., not yet born human beings and animals into consideration when it comes to ethical, legal, and political discussions. Obviously, the protection of human rights, pacificity, and the capability of conflict management, or expenses a welfare state makes to afford all its citizens access to learning, education, health, and social security are good examples of what I mean by "standard of civilization", and which goes far beyond mere economic status. And just as the actual degree of realization of the liberal values constituting a free basic democratic order, an already achieved standard of civilization is something one can always and at any time fall behind. But at the same time it is also something that can systematically be enhanced.
What today's Western societies perhaps lack the most are systematic and institutionalized forms in which a country's citizens can increase the degree of their own mental autonomy. We still lack a deeper understanding of the fact that in the end it will always be precisely the mental autonomy of every individual citizen which makes the essential contribution to any sustained increase in the standard of civilization. The new science of mind wandering that we briefly looked at now shows with surprising clarity how during two thirds of our conscious lifetime we are not mentally autonomous subjects. It also provides objective evidence that in the end this fact in many ways, directly or indirectly, leads to a lower quality of life. Therefore, we should all think hard about concrete options to increase our mental autonomy. In this respect, the implementation of meditation lessons at schools and higher educational institutions may well be the most urgent and important political demand. But it is not the only one.

I believe the perhaps most important contribution that academic philosophy can make to high school curricula is so-called "critical thinking" or "informal logic". Informal logic is a branch of philosophy that is concerned with the forms and uses of arguments. It is not just about gaining a better understanding of the logical structure of arguments and later being able to build a rational argument oneself. Informal logic also systematically trains the ability for

critical thinking, for detecting fallacies, and to productively managing disagreement and intellectual conflicts in order to learn from each other. I think that systematically structured courses in informal logic would be another, equally important contribution to a sustained increase of the standard of civilization and mental autonomy. They would allow students to reliably recognize the most important types of fallacies, not to fall for rhetorical tricks, and to settle conflicts of opinion in an evidence-based and intellectually honest way. Such training would, as it were, be about developing the proper configuration of our cognitive self-model, and the question for neuroscience would be what exactly the right time window in the psychological development of adolescents would optimally support this part of the human self-model in its process of unfolding.

For me, meditation courses and training in informal logic at schools are complementary political demands, because one builds on the other. One needs – as empirical research on mind wandering clearly shows – the form of mental autonomy that is cultivated in formal meditation practices for stable mental self-control, in order to be able to see things clearly and think rationally in the first place. Critical rationality presupposes mindfulness; intellectual honesty is a special case of a truly spiritual attitude (as I have explained in my freely available open access essay "Spirituality and Intellectual Honesty"). But every adolescent also needs a solid understanding of the basic standards of rational, critical thinking – for example, in order not to fall for the bizarre ideological nonsense that frequently goes along with such offers outside of school. Keeping it in a sober and straightforward perspective, mindfulness and rationality simply rest on a certain set of mental skills and abilities that can be determined very precisely and therefore also be trained. The overall distribution and the individual expression of such skills then indirectly, but ultimately in a very strong way, determines the standard of civilization a given society achieves in the end. One really meaningful application of current research in the neuro and cognitive sciences would therefore consist in providing political decision-makers with clear and reliable information about what exactly is possible in this domain, and what options for action there are for a reasonable implementation in schools and universities.

How does all this relate to the question of freedom of will we started with at the beginning? Here is my answer: Even a physically-determined system can exhibit degrees of autonomy. We must transcend black-and-white thinking. Freedom is not something absolute – it comes in degrees of flexibility and context-sensitivity. Freedom comes in degrees of rationality, with intellectual honesty, and it also comes with the capacity of effortlessly resting within a choiceless form of awareness, of quietly observing without an observer. Therefore, the relevant question is: How can I raise my own degree of mental autonomy and that of other sentient beings – and how can modern science, philosophy and technology help me with this project? In this essay, I have only presented three very short, positive proposals to provide a foundation for future discussions concerning what a valuable state of consciousness might be. The three principles of reducing suffering, optimizing epistemic potential, and a systematic enhancement of mental autonomy, therefore, are not much more than starting points, an invitation to begin a new conversation.

www.questionofwill.sk

25 YEARS OPEN SOCIETY FOUNDATION
PRIESTOR SÚČASNEJ KULTÚRY
eea grants
norway grants
Kunsthall Trondheim

Projekt Otázka vôle je financovaný príspevkom vo výške 117 213 eur z grantov Islandu, Lichtenštajnska a Nórska prostredníctvom Finančného mechanizmu EHP a zo štátneho rozpočtu Slovenskej republiky.

Project Question of Will is supported with 117 213 eur by a grant from Iceland, Liechtenstein, Norway. Co-financed by the State Budget of the Slovak Republic.

www.eeagrants.sk | www.norwaygrants.sk