CHAPTER

9

Why Is Mind-Wandering Interesting for Philosophers?

Thomas Metzinger

Abstract

This chapter explores points of contact between philosophy of mind and scientific approaches to spontaneous thought. While offering a series of conceptual instruments that might prove helpful for researchers on the empirical research frontier, it begins by asking what the explanandum for theories of mind-wandering is, how one can conceptually individuate single occurrences of this specific target phenomenon, and how one might arrive at a more fine-grained taxonomy. The second half of this contribution sketches some positive proposals as to how one might understand mind-wandering on a conceptual level, namely, as a loss of mental autonomy resulting in involuntary mental behavior, as a highly specific epistemic deficit relating to self-knowledge, and as a discontinuous phenomenological process in which one's conscious "unit of identification" is switched.

Key Words: mental autonomy, mind-wandering, philosophy of mind, self-knowledge, consciousness, mental behavior

The main goal of this chapter is to isolate and draw attention to a number of conceptual issues in the now burgeoning empirical literature on mind-wandering. This is only a first outline of such issues and is not intended to be an exhaustive list. However, I am convinced that such philosophical questions are not only relevant, but that—on a different level of analysis—they may reveal potentially rewarding targets for future research. One way to view philosophy of mind is as a *meta-theoretical* enterprise: While first-order experimental work tries to get as close as possible to the target phenomenon itself, the second-order theoretical work of philosophers focuses on methodological issues and the conceptual structure of first-order, data-driven theories. A philosophical approach will therefore mostly be not about mind-wandering itself, but rather about the *concepts* that we use and develop to understand it. In practice, of course, both levels of investigation are deeply intertwined; empirical researchers have a strong interest in methodological issues and a tendency to make implicit conceptual background assumptions, while philosophers may not only offer novel conceptual instruments to experimentalists, but also formulate empirical hypotheses themselves.

Philosophers of mind often try to contribute to a more comprehensive and unified general framework that can guide and inspire empirical research, and of course they always have their own, more abstract, questions in the back of their minds. From a philosophical perspective, it has proven to be interesting and fruitful to select specific, increasingly well-researched target phenomena and to see what can be learned about them from the empirical literature. By always keeping an eye on the value of findings as "bottom-up constraints" for a more general, comprehensive model of the human mind, while at the same time, in a critical vein, trying to uncover conceptual inconsistencies or unwarranted background assumptions that may block further progress, the philosopher can make important contributions to the field. Interdisciplinary cooperation works best

when focused on a specific target phenomenon like conscious experience in virtual reality (Metzinger, forthcoming), dreaming (Metzinger, 2013b), out-of-body experiences (Metzinger, 2009), or identity disorders (Metzinger, 2004), and this short chapter can be seen as another installment in a series of attempts to lay some conceptual foundations for an empirically grounded theory of self-consciousness.

In the first three sections of this chapter I will ask many questions. In doing this, my aim is to flag some epistemic targets of potential interest for an interdisciplinary readership. I will ask what the *explanandum* for theories of mind-wandering is, how we can conceptually *individuate* single occurrences of our target phenomenon, and how we might arrive at a more fine-grained *taxonomy*. In the three subsequent sections (constituting the second half of the chapter), I will suggest some answers, sketching a series of positive proposals as to how we might understand mind-wandering on a conceptual level: as a loss of *mental autonomy* resulting in involuntary mental behavior; as a highly specific *epistemic deficit* relating to self-knowledge; and as a *phenomenological* process in which our conscious "unit of identification" is switched ("UI-switching" is a new technical term I will introduce in the final section).[1]

The Explanandum

What, exactly, is it that we want to explain in scientific research on mind-wandering? Is the target phenomenon really a unitary phenomenon, for instance a distinct type or class of mental processes? And what would therefore *count* as an explanation of the phenomenon in its entirety?

Let us take the widespread notion of "spontaneous, stimulus-independent or task-unrelated thought" (Antrobus, 1968; Giambra, 1989) as a starting point. It is important to note how, in its origin, "spontaneity" is an exclusively phenomenological concept, because it is based on the introspective experience of an apparently uncaused, subjectively unexpected, and sudden onset of conscious thought. "Spontaneity" is therefore not an objective property, but rather an entirely subjective characteristic of certain thoughts. It therefore cannot function as an empirical demarcation criterion to define the boundaries of our target domain. Empirically, it is plausible to assume that there will always be unconscious neural precursors of mind-wandering. These could, for example, be specific introspectively inaccessible goal representations that drive the high-level phenomenology of mind-wandering (Klinger, 2013), such as postponed goal-states that have been environmentally cued by goal-related stimuli under high cognitive load (Cohen, 2013; McVay & Kane, 2009). To understand the overall process, we may have to adopt the "dolphin model of cognition": Just as dolphins cross the boundary between water and air, thought processes often cross the border between conscious and unconscious processing, and in both directions. For example, "spontaneously occurring" chains of cognitive states may have their origin in unconscious goal-commitments triggered by external stimuli, then transiently become integrated into the conscious self-model for introspective availability and selective control, only to disappear into another unconscious "swimming bout" below the surface. Conversely, information available in the conscious self-model may become repressed into an unconscious, modularized form of self-representation where it does not endanger self-esteem or overall integrity (Pliushch & Metzinger, 2015).

Conceptually, to take the "spontaneity" characterizing the onset of a mind-wandering episode seriously as an objective property of the human mind would mean to accept it as causally indeterminate—and therefore inaccessible to standard experimental methods.[2] On a physical or functional level of description, to call something "spontaneous" means to describe it as "uncaused" and "lawless"—and doing so could even be seen as a form of hand-waving. For this reason, the first semantic element in "spontaneous, task-unrelated thought" will not help us in isolating the explanandum.

A more moderate and nuanced account could attempt to describe *degrees* of spontaneity and analyze them as degrees of constraint satisfaction on different levels of analysis. For example, one traditional philosophical distinction is between the "content" and the "vehicle" of a mental process. We could then separately investigate constraints governing the content of thought, as well as the constraints determining the dynamic neural carriers (i.e., the "vehicles") of this content, and we could accordingly distinguish different degrees of constraint-satisfaction. As Christoff et al. (2016, p. 2) have proposed, "spontaneous thought" could then be characterized by the absence of strong content-constraints imposed either by deliberate cognitive control and/or "automatic" constraints. Automatic constraints would presumably influence the neural carriers in a functionally more direct way, for example by mechanisms implicitly processing affective or sensory salience.

Importantly, however, fundamental methodological issues remain, because one has to distinguish between the representational content of a given neurodynamical state as ascribed from a third-person perspective and as introspectively reported from a first-person perspective. First, the content of a mind-wandering episode might be described as "unconstrained" relative to some theory of mental representation or under a specific mathematical model of neural computation; it would then be a property ascribed by an external observer. On the other hand, if the "content" is what can be introspectively accessed and reported by experimental subjects—for example, by asking, "Was your mind moving about freely?"—then subjectivity is back in. Researchers get a statistical measure of self-reports and can fruitfully and legitimately employ what Daniel Dennett calls the "heterophenomenological method" (Dennett, 1991, pp. 72–81), but are ultimately still dealing with a phenomenological construct.

One path toward a solution may consist in analyzing "spontaneity" not as a phenomenological or a metaphysical property of certain thought processes, but as an *epistemic* feature: Perhaps it is a systematic lack of knowledge, a specific form of introspective blindness characterizing a very large portion of conscious thought. As Smallwood and Schooler (2015, p. 491) put it, "the spontaneous occurrence of mind wandering means that the causal path that links the experience to ongoing processes and outcomes is opaque." If this is correct, then one way to transform "spontaneity" into a proper, experimentally tractable explanandum for research would be to isolate exactly those causal conditions in the brain that make the causal precursors of a given cognitive event functionally available for introspective attention and verbal report. "Opaque" then means that there is no internal model of the causal path that can be introspectively accessed, and this would also give us a first functional-level notion of "spontaneity." We could then ask what exactly the neural mechanisms creating an internal model of what philosophers call "horizontal mental causation" (i.e., the linear causation of one mental event by another) are. What is the domain or the subset of cognitive processes on which these mechanisms operate? How do they break down? Can they be experimentally modulated?

Perhaps it is *never* the case that what one ascribes or introspectively reports as the "content" of an episode actually causes the "content" of the next episode. Alternatively, perhaps it is only sometimes the case; possibly our internal, introspectively accessible model of horizontal mental causation simply is a misrepresentation most of the time—a high-level confabulation that has proved to be adaptive? It is important to understand that all we can ever introspectively attend to is a *model* of our own cognitive processing. We never have a mysterious "direct" form of access to the cognitive processes in our heads, because all knowledge—including self-knowledge—always is knowledge *under a representation*. Therefore, the core target for research may actually be the way in which our brains *model* the causal relations between their own inner states, the way the system creates an internal model of itself by trying to predict and "explain away" its own mental behavior. A brain that was functionally unable to generate dynamic models of horizontal mental causation could only support a phenomenology of one "spontaneously occurring" mental content after another. An organism with such a brain could experience mental causation together with a high degree of "freedom," subjective unpredictability, or a lack of cognitive control in terms of rational connection between content elements, but in the complete absence of an internal self-model explicitly portraying horizontal causal chains connecting mental events, the cognitive first-person perspective would simply dissolve. For such an organism, there would never be a coherent *train* of thoughts, only events, and never a unified process in terms of a temporally extended cognitive *Gestalt*. The self-conscious mind would be largely unintelligible to itself, a constant source of surprise and uncertainty—it would be hard for the organism harboring it to experience it as its own mind. It could therefore never develop an inner image of the system carrying it as an embodied "thinker of thoughts," as an entity that is not only a bodily, but also a cognitive agent. Perhaps some animals have a cognitive phenomenology of exactly this kind, unfolding on a more robust and stable platform of bodily and emotional self-consciousness. Our brains are different.

Now take the second semantic element, "task-unrelatedness." Philosophers will immediately see that there is a very strong implicit background assumption hiding behind this idea: that human beings pursue one and only one task at a time. But we know that this is false. A biological organism has multiple tasks and many problems to be solved at the same time—it is continuously faced with a multitude of challenges that have to be met. Any higher biological organism is a paradigmatic example of parallel processing, and there are many levels

of functional granularity on which it must continuously operate—sustaining its existence; preserving homeostatic stability; continuous prediction error minimization relative to the dynamic, internal self-model created by the brain; successfully achieving procreation; rising in a social dominance hierarchy by effectively deceiving self and others (Pliushch & Metzinger, 2015; von Hippel & Trivers, 2011). At any given point in time, in the organism's central nervous system, there will be multiple goal-representations competing for the control of overt behavior, for the focus of attention, and for high-level cognitive. In addition, for embodied agents like ourselves who constantly refine and update the interoceptive layers of their self-model, there never really is anything like an absence of internal constraints (*pace* Andrews-Hanna, Irving, Fox, Spreng, & Christoff, Chapter 13 in this volume).

An analogous point holds for the notion of "stimulus-unrelated thought." As the human self-model is functionally anchored in elementary processes of bioregulation (Metzinger, 2003a, 2014), no form of cognitive processing is ever fully disembodied or independent of the continuous bombardment by stimuli originating in the *interior* of the body. Interoceptive input, proprioception, the continuous flow of vestibular information, or the "background buzz" generated by autonomous activity in the input-independent layers of the body-schema are examples of permanent sources of stimulation. These internal sources of stimulation and constant perturbation influence not only bodily self-awareness, but also our emotional self-model, thereby setting an internal context. We may not be subjectively aware of this context at all times, and we may be even less aware of the detailed causal pathways by which it shapes and relates to the contents of our thoughts, but it certainly exists. Upon closer inspection, the notion of "stimulus-unrelatedness" really relates to a phenomenal property: the conscious experience of our ongoing thoughts as not being caused by something subjectively represented as belonging to the extracorporeal environment—to an *external* stimulus. In order to bridge the gap between an implicitly phenomenological concept and a productive functional analysis, we could proceed from looking at the content to focusing on the physical dynamics on which this content "rides"—for example, by looking at the way in which the brain *predicts* incoming stimuli by "canceling out" sensory input.

Similarly, "task-unrelatedness" may also ultimately only be a phenomenological property, one that is derived from the high-level introspective experience of only being able to solve one problem at a time. A strong metaphysical interpretation of the second element of "task-unrelatedness" would have to say that a large portion of human cognition is actually aimless—an arbitrary process that in a fundamental way cannot count as goal-directed, perhaps not even as a form of intelligence. This would make task-unrelatedness difficult to understand from an evolutionary perspective, because it would involve paying a high metabolic price for a ubiquitous dynamic feature that ultimately doesn't serve any of the organism's needs (but see Simonton, Chapter 10 in this volume).

Again, an epistemological perspective may prove to be fruitful in defining the explanandum more clearly: Mind-wandering is an inner process experienced by an organism for which the organism possesses no conscious knowledge of the goals the process serves, simply because there is no introspectively available model of the goal-state. In the generation of intelligent behavior, when exactly is it necessary to have an internal model of the goal-state? When is it a superfluous waste of resources? Is there a specific functional advantage of explicit goal-representation (e.g., veto control or the creation of an illusion of trans-temporal identity)? Elsewhere (Metzinger, 2013a, 2015) I have argued that mind-wandering is an unintentional form of mental behavior.[3] Unintentional mental behaviors may surprise the organism in which they emerge and may be basically inexplicable to it from its limited inward perspective, while still being a very efficient and adaptive form of intelligence. For example, one speculative but perhaps novel hypothesis is that a considerable portion of mind-wandering actually is "mental avoidance behavior": an attempt to cope with adverse *internal* stimuli or to protect oneself from a deeper processing of information that threatens self-esteem. There is nothing wrong with the idea of a cognitive system whose behavior is driven by a multiplicity of goal-representations, the content (and the continuous hierarchical restructuration) of which it does not consciously know or understand. Before claiming the existence of an objective property like "task-unrelatedness," it may therefore be more interesting to look at the dynamic mechanisms of task-*representation* first. This leads to genuinely philosophical questions: How are "goal-states" or "tasks" individuated in the first place? What is different for exclusively *mental* forms of goal-relatedness, and are there specific sets of satisfaction conditions characterizing cognitive actions,

and only cognitive actions? Such questions provide further reasons why mind-wandering can be interesting for philosophers, especially as one of the many reasons why mind-wandering is interesting for philosophers is that it directs our attention to the problem of mental action (the "contrast class," if you will); mind-wandering poses the interesting challenge of describing the deeper principles of goal-state selection and action initiation while subtracting the non-neural body and abstracting from issues of motor implementation (Metzinger, 2017).

"Thought" is the third semantic element in our notion of "spontaneous, task-unrelated thought." On the one hand, it is difficult to define what "thought" is in the first place; on the other hand, one of the greatest contributions of the field of mind-wandering research to cognitive science may exactly lie in finally introducing a massive and empirically grounded taxonomical differentiation for the term "cognition." Philosophers have, of course, thought about this third semantic element characterizing mind-wandering for centuries. I will not even begin to sketch the theoretical landscape. Instead, I will confine myself to pointing out that terms like "cognition" or "cognitive" have long become empty buzzwords in neuroscience and empirical psychology, and that this problem has to be solved in a principled manner—at least if a conceptual construct like "spontaneous, task-unrelated thought" is to be used by serious people wishing to treat it as referring to a potential explanandum for rigorous empirical research.

But here are some questions one might ask to get started: Is "conscious thought" simply a folk-psychological term that should be eliminated in favor of a fine-grained neuroscientific theory? Are there necessary conditions, such as agentive direction, for verbally reportable types of mental activity to count as "thought"? Is the target phenomenon tied to the wakeful state, or does conscious cognition in the dream state similarly present us with an example of "thought"? Philosophers individuate thoughts by their contents, by what they are *about*. Can mind-wandering be about anything, or are there specific forms of content characterizing the target phenomenon? The problem to be solved is that in developing a systematic catalogue of explananda, we might end up with a very long disjunction ("Mind-wandering is *a* or *b* or *c* or . . .") and risk the danger of widespread fallacies of equivocation. In informal logic, the "fallacy of equivocation" refers to the misleading use of a term with more than one meaning or sense. Therefore, one needs to be able to say clearly what one *single* and what *one and the same* occurrence of the target phenomenon are. As I will explain in the next section, in order to do proper science, mind-wandering episodes have to be turned into countable entities, and we need criteria to determine their identity. I take it that empirical researchers currently are unable to do this.

Individuation

What are the temporal boundaries of a given, single episode of mind-wandering or a specific period of "spontaneous, task-unrelated thought"? When exactly does such an episode begin, and when does it end? Putting the question slightly differently, if we conceive of an individual episode as a chain of mental events, what counts as the *first* event in this chain and what is the *last*? Such questions raise further important issues. For example, could there be episodes constituted by one single mental event only? Or is there a minimal number of events—say, the attentional lapse, the appearance of the first retrospectively reportable (e.g., "task-unrelated") content, plus the terminal moment of meta-awareness?

To "individuate" episodes means to turn them into single, countable entities. To turn mind-wandering into a proper target for empirical research, we do not want to ask, "How *much* mind-wandering was there, during a given period of time?" but rather "How *many* individual occurrences of our target phenomenon could we experimentally detect?" In principle, it must be possible to say, "During the last 300 seconds, subject *s* had 14 distinct episodes of mind-wandering, namely, episode *a*, which lasted 2,834 milliseconds and began just after 5,398 milliseconds, episode *b*, [. . .], and finally episode *n*, which lasted 4,793 milliseconds and ended precisely 2.5 seconds before the end of the experimental period." In order to achieve this, one needs not just testable, objectively viable criteria marking the onset and the end of each episode, but also a criterion for counting psychological items of this newly introduced kind, as well as a criterion helping us to decide on identity or non-identity among items of that kind. For example, would our future theory of mind-wandering allow that a patient has *one and the same* recurring negative thought pattern again, at multiple points or intervals in time? Are there context-invariant "cognitive atoms," distinct units of mental content that can be activated in the subject's conscious mind, again and again? There are deep and complex conceptual questions lurking in the background. Here is another one: We do not want individual mental episodes to

possess proper parts that themselves are of that psychological kind we call "mind-wandering" or "spontaneous, task-unrelated thought"—else we may have problems counting them. What, then, is the smallest explanatory unit? Perhaps most of all, some of us may also want to know what the "essence" of our target phenomenon actually is, *what* that kind of phenomenon is.

In earlier work, I have made some positive proposals. One proposal is that the essence or inner nature of mind-wandering is "UI-switching," a sudden, subjectively unpredicted, and often unnoticed change in the phenomenological unit of identification (see the last section of this chapter). Recall that a unit of identification simply is whatever is currently experienced as the conscious self, whatever conscious content would give rise to reports of the type "I *am* this!" I have grave doubts that "essences" in a strong metaphysical sense really exist, but framing an answer in this more modest manner could perhaps help us to specify what, in our world and under the laws of nature that hold in it, is common to all occurrences of mind-wandering—what constitutes their inner nature. If I am correct, mind-wandering occurrences can be characterized by a single UI on the level of their content, and in time they are "bracketed" by shifts in the UI (again, see the last section). Second, I have also formulated an empirical hypothesis saying that the onset of every single episode must be characterized by a discontinuity in phenomenal self-awareness—an experimentally detectable "self-representational blink" (SRB; Metzinger, 2013a, p. 9). Third, I have proposed that the end of every single episode that leads to a regaining of cognitive self-control is marked by another shift in the self-model involving the reappearance of an explicit representation of the ability for mental veto control, typically accompanied by a voluntary termination of the ongoing mental chain of events. This creates a new unit of identification, namely, the "meta-aware self"—an internal model of an active entity that has the ability to end an ongoing chain of task-unrelated thought and to return the focus of attention to what is now consciously remembered as "the" original task.

Taxonomy

Imagine you are sitting in a boring lecture and have drifted off into a pleasant erotic fantasy. After you become aware of the fact that you have just completely zoned out, you carefully tune back into the fantasy while paying some attention to the lecture. Whenever you notice that you have had another full attentional lapse and completely zoned out (the lecture still hasn't got any better), you deliberately tune back into the fantasy again, afterward "letting it go," observing it as it unfolds by itself. Is this interplay between mental action and the ensuing loss of cognitive control, the recurring cycle between "zoning out," "coming to," and "tuning out" again a form of mind-wandering? Only one-half to two-thirds of it can really be characterized as "task-unrelated thought." Perhaps the real "task" here is not actually listening to the lecture. Maybe the highest-priority task consists in keeping up the outer appearance of being an interested listener and in being a well-rested and relaxed conversant at the conference dinner afterward, remembering just a minimally sufficient number of keywords to (in a social emergency) be able to fake an intelligent question or two? How much of the contents of your erotic fantasy was "spontaneous" in the sense that there really was a strong introspective experience of sudden onset, novelty, and unexpectedness? Was it caused by unconscious interoceptive stimuli? Was its introspectively available content really "conceptual" in the sense of high-level symbolic cognition, something that can be called "thought" in a stricter and narrower sense, or was the experiential content rather one constituted by visual and motor imagery, an embodied and increasingly complex simulation of the tactile, kinesthetic, or interoceptive stimuli to be expected and of the emotional arousal they would cause?

From a traditional philosophical perspective, a systematic taxonomy would have to individuate mind-wandering episodes by their verbally reportable content and by the functional context in which they occur. This may already lead to helpful taxonomical differentiations. It also draws attention to another methodological constraint that empirical research has to satisfy: namely, always clearly distinguishing between "content" as experienced and ascribed from a first-person perspective (1PP) and "content" as ascribed from the scientist's third-person perspective (3PP). For example, there are highly developed and centuries-old 1PP approaches to what, today, we like to call "online experience sampling" (namely, classical mindfulness meditation of the "open monitoring" type), and there are much more recent 3PP approaches involving external cueing and a statistical analysis of the results related to whole groups of individuals (see Andrews-Hanna et al., Chapter 13 in this volume). These approaches lead to diverging results, as 3PP content is a much more abstract theoretical construct than

1PP content. However, intuitions anchored in 1PP content may contaminate theory formation on the level of science, and scientifically informed subjects will at some point re-import 3PP content into their individual phenomenological reports.

Take the example of depressive rumination. Is it really "task-independent"? What about the diminished attentional control caused by worry and self-referential negative thoughts in anxiety disorders (Forster et al., 2015)? Maybe it actually reflects, from a neurocomputational perspective, a close-to-optimal form of processing, just as many visual illusions can count as "optimal percepts" if analyzed mathematically. We must never forget that first-person description and third-person functional analysis may greatly diverge. For example, for certain taxonomical categories, there may be hidden epistemic benefits that are "invisible" from a subjective, first-person perspective.

Here is another example of a philosophical concept that I think might possess great heuristic fecundity for future research on "stimulus-independent, task-unrelated thought": "epistemic innocence." Philosopher Lisa Bortolotti (Bortolotti, 2015a, 2015b) has recently introduced this concept to articulate the idea that certain mental processes such as delusion and confabulation (which may count as suboptimal from an epistemological perspective) may actually have not just psychological, but also epistemic benefits. Their psychological benefits may not simply be purchased with epistemic costs—what superficially appears as an imperfect cognitive process may actually epistemically innocent, causally enabling not only coherence, but also mental knowledge acquisition. I believe this philosophical idea can be fruitfully applied in the domain of mind-wandering. For example, does depressive rumination perhaps serve the interests of the individual or its group, fulfilling a task that is introspectively inaccessible to the patient and which, scientifically, we simply haven't understood yet? Are the comparably perseverative forms of thinking happening during NREM sleep (including slow-wave sleep) a form of conscious thought, too? In what sense are they "task-unrelated"? Are there perhaps hidden epistemic benefits to dreamless sleep experience? Mentation during sleep as well as cognition during ordinary and lucid dreams are examples of candidates for our taxonomy of states of mind-wandering, which are characterized by a special functional context (shallow levels of embodiment and situatedness) and, with the exception of lucid dreams, are generally absent cognitive self-control (Metzinger & Windt, 2007). Interestingly, there is a strong overlap between theoretical issues in empirical research on dreaming and mind-wandering (Fox et al., 2013; Metzinger, 2013b; Windt, 2015; see also Windt & Voss, Chapter 29 in this volume). Again, here are some examples: What exactly is the relationship between mental self-control, the occurrence of dream lucidity, and what researchers in mind-wandering call "meta-awareness"? Can both lucid lapses and mind-wandering lapses plausibly be interpreted as the disintegration of an internal epistemic agent model (see later discussion in this chapter)? Are there common *positive* functionalities connecting dreaming and mind-wandering during wake states, such as the encoding of long-term memory, complex, preparatory motor planning, or creative incubation? Philosophically, it is also interesting to look at "false lucidity" and the phenomenology of insight (Kühle, 2015; Voss & Hobson, 2015): in becoming lucid at night and during daytime mind-wandering, is the experience of *oneself* having actively regained meta-awareness (and thereby mental autonomy) an illusion of control over a mental event that was really triggered by an unconscious process? Currently, this is only a conceptual possibility; What kind of experimental design could settle the issue?

An even more radical approach could take the step from a content-taxonomy to a vehicle-taxonomy. We could stop speaking of the "content" of mind-wandering episodes altogether, for example by exclusively focusing on dynamical properties of its minimally sufficient neural correlates. Is depressive rumination "spontaneous, task-unrelated thought"? Kalina Christof and colleagues arrive at a negative answer (cf. Christoff et al., 2016, p. 8), by proposing a wider dynamicist framework for understanding mind-wandering under which depressive rumination would be understood as a member of a family of spontaneous-thought phenomena. Drawing on Ludwig Wittgenstein (e.g., *Philosophical Investigations* I: 66), a philosophical analysis of this point could say that "mind-wandering" actually is a cluster concept (i.e., a term that is defined by a weighted list of criteria, such that no one of these criteria is either necessary or sufficient for membership). Recall how a more nuanced account of mind-wandering could attempt to describe *degrees* of spontaneity and analyze them as degrees of constraint satisfaction at different levels of analysis. For depression, we clearly find rigidity and an involuntary fixation on symptoms of distress at the content level. In addition, one level of description below, there is a

diminished degree of constraint satisfaction for the functional property of M-autonomy (Metzinger, 2015; see later discussion in this chapter), because in depressive rumination patients have great difficulties in disengaging from their own involuntary behavior, as their capacity for veto control on the mental level is weakened. Here, the overall functional context would be a clinical one, with a local microfunctional correlate on the molecular level (e.g., a dysbalance of certain neurotransmitter systems characterizing the "vehicle" or carrier), while the content might be redundant, repetitive, and characterized by negative affective valence. This opens the radical possibility of increasingly proceeding without the ascription of "content" at all, semantically enriching our initial cluster concept of "mind-wandering" by *exclusively* defining it by criteria on the vehicle level. In their synthesis of the new interdisciplinary field of spontaneous thought, Jessica Andrews-Hanna and colleagues (Chapter 13 in this volume) propose exactly this—breaking out of the "flashlight" of 1PP content into the rich darkness of lower levels of description. But what exactly is it that we are trying to find in the dark, and is it something that should still be called "thought"?

I hope that my readers will agree that more than enough new questions have been asked in the three preceding sections—it is now time to offer some answers. The general point I have been trying to make should be clear by now: to sustain its great initial success, experimental research on spontaneous thought needs a much more systematic and fine-grained taxonomy of its research targets. Such a taxonomy cannot be constructed in a purely data-driven, bottom-up manner, because it rests on implicit conceptual assumptions and on our epistemic interests. What exactly is it that we want to *know*? I will not discuss any further examples in the remainder of this chapter. Instead I will present a series of positive proposals for developing a conceptual framework, plus some first conceptual tools that empirical researchers could operationalize and apply in the design of experiments. I will begin with the subjective sense of agency and the possibility of illusions of control on the mental level.

Losing and Regaining Mental Autonomy: Mental Action Versus Unintentional Mental Behavior

Philosophers have thought long and hard about what distinguishes "actions" from other kinds of events in the physical world (Davidson, 1988/2001; Dretske, 1988; Wilson & Shpall, 2016). Indeed, "action theory" can be considered a small subfield within the discipline of academic philosophy. There are, however, also *mental* actions—and this is another point of contact where mind-wandering becomes interesting to philosophers (Metzinger, 2017; O'Brien & Soteriou, 2009; Pezzulo, 2017). Perhaps some elements of the philosophical toolkit can prove to be interesting for experimentalists as well.

Deliberately focusing one's attention on a perceptual object and consciously drawing a logical conclusion are examples of mental actions. Just like physical actions, mental actions possess satisfaction conditions (i.e., they are directed at a goal state). Although they mostly lack overt behavioral correlates, they can be intentionally inhibited, suspended, or terminated, just like bodily actions can. Additionally, they are interestingly characterized by their temporally extended phenomenology of ownership, goal-directedness, a subjective sense of effort, and the concomitant conscious experience of agency and *mental* self-control.

Let me distinguish the two most important types of mental action:

- *Attentional agency* (AA): the ability to control one's focus of attention;
- *Cognitive agency* (CA): the ability to control goal/task-related, deliberate thought.

AA and CA are functional properties that are gradually acquired in childhood, can be lost in old age or due to brain lesions, and whose incidence, variance, robustness, and so on, can be scientifically investigated. However, they also have a subjective side. Attentional agency (Metzinger, 2003a, 6.4.3; 2006, Section 4; 2013a; 2013b; 2015) also possesses a phenomenal signature, as is the case for other forms of subjective experience, like pain or the subjective quality of "blueness" in a visual color experience. For this reason, AA also has a phenomenological reading: as the conscious experience of actually initiating a shift of attention, and of controlling and fixing its focus on a certain aspect of reality. AA involves a sense of effort, and it is the phenomenal signature of our functional ability to actively influence what we will come to know, and what, for now, we will ignore.

Consciously experienced AA is theoretically important because it is probably the earliest and simplest form of experiencing oneself as a knowing self, as an epistemic agent. Human beings learn to control the focus of their attention long before they can control symbolic, high-level cognition.

Research into animal intelligence and human phenomenology shows that AA can and does exist without cognitive control, but modern dream research and various psychiatric syndromes demonstrate that cognitive agency causally depends on and might actually be a functional derivative of attentional agency (e.g., Windt, 2015). To consciously enjoy AA means that you (the cognitive system as a whole) currently identify with the content of a particular and highly specific type of mental self-representation, an "epistemic agent model" (EAM; Metzinger, 2013a, 2013b, 2015, 2017). Whenever such an EAM is active in your brain, you experience yourself as a knowing self, an agent searching to improve its knowledge about the world. AA is fully transparent:[4] The content of your conscious experience is not one of self-*representation* or of an ongoing process of self-modeling, of depicting yourself as a causal agent in certain shifts of "zoom factor," "resolving power," or "resource allocation," in actively "optimizing precision expectations" or engaging in a "selective sampling of sensory data that have high precision (signal to noise) in relation to the model's predictions" (Feldman & Friston, 2010, p. 17). Rather, you directly experience *yourself* as, for example, actively selecting a new object for attention or trying to "see things more clearly." This is interesting because although during many types of mind-wandering episodes we do not have AA, these episodes can of course be *about* having been an attentional agent in the past, or *about* planning to control one's attention in the future.

Analogously, a closely related point can be made for CA. Conceptually, cognitive agency is not just a complex set of functional abilities such as the capacities for mental calculation; consciously drawing logical conclusions; engaging in rational, symbolic thought; and actively constructing new arguments. Again, there is a distinct phenomenology of currently being a cognitive *agent*, which can lead to experiential self-reports like "I am a thinking self in the act of grasping a concept," "I have just actively arrived at a specific conclusion," and "I am attempting to build an argument." There is a functional analysis ("autonomous cognitive self-control") and a phenomenological reading, based on verbal self-reports. The classical meta-theoretical issue, of course, is in what sense autophenomenological reports can or should inform the process of functional analysis and decomposition. But most important, what AA and CA have in common is that in both cases, we consciously represent ourselves as epistemic agents: According to subjective experience, we are entities that actively construct and search for new epistemic relations to the world and ourselves. We are information-hungry, and there is something we want to *know*.

Empirical research programs on spontaneous, apparently task-unrelated thought are interesting for philosophers, because they demonstrate (a) that epistemic mental agency is a *much* more vulnerable and *much* rarer phenomenon than many philosophers of mind may have intuitively assumed, and (b) that what we traditionally call "conscious thought" or "high-level symbolic cognition" may, more often than not, be a *subpersonal* process (as I have argued elsewhere; see Metzinger, 2013a, 2015). Such programs raise the need for conceptual demarcation criteria allowing us to distinguish between intentional mental action and unintentional mental behavior, as well as between personal-level thought, and forms of conscious cognitive processing that are better described as automatic, sub-personal chains of events. Nevertheless, the wealth of existing philosophical literature on action and thought may provide many helpful conceptual tools for empirical researchers to use to sharpen their hypotheses and predictions. For example, it would be excellent if empirical investigators always carefully distinguished between personal-level mentation and subpersonal processes, between properties of the person as a whole and properties of his or her brain. My own more specific positive proposal is this: the beginning of every mind-wandering episode is marked exactly by the collapse of our epistemic agent model (a conscious self-representation of *now* possessing the ability for epistemic self-control), and the end of every episode is marked by the re-emergence of a new epistemic agent model (the "meta-aware self"). How can we spell this point out on the functional level?

I think it could be heuristically fruitful to analyze mind-wandering as a *loss of mental autonomy*. The topic of mental autonomy (M-Autonomy hereafter; cf. Metzinger, 2015) is an excellent example of an area in which empirical research into mind-wandering makes a contribution to issues possessing great relevance in other fields, not only philosophy, but also law and psychiatry.[5] Very generally speaking, autonomy is the capacity for rational self-control, whereas the term "mental autonomy" refers to the specific ability to control one's own mental functions, like attention, episodic memory, planning, concept formation, rational deliberation, and decision-making. Mental autonomy includes the capacity to impose

rules on one's own mental behavior and to explicitly select goals for mental action, as well as the ability for rational guidance and, most important, the intentional inhibition, suspension, or termination of an ongoing mental process. M-autonomy is a functional property,[6] which any given self-conscious system can either possess or lack. Its instantiation goes along with new epistemic abilities, a specific phenomenological profile, and the appearance of a new layer of representational content in the phenomenal self-model (Metzinger, 2003a). In humans, first insights into its neuronal realization are now beginning to emerge. From a philosophical perspective, this functional property is interesting for a whole range of different reasons. One of them is that it is directly relevant to both our traditional notions of a "first-person perspective" (1PP) and of "personhood" (Metzinger, 2015). If one cannot control the focus of one's attention, then one cannot sustain a stable first-person perspective, and for as long as one cannot control one's own thoughts, one cannot count as a rational individual. In other words, spontaneous thought is a subpersonal process, like respiration or heartbeat.

Biological systems produce different kinds of observable output, which can in turn be characterized by different degrees of autonomy and self-control. There are actions and behaviors, and both kinds of output are conceptually individuated by their satisfaction conditions—that is, they are directed at goal states. However, for actions, conscious goal-representation plays a central causal role: actions are typically preceded by a selection process; they can be terminated, suspended, or intentionally inhibited; and they exhibit a distinct phenomenological profile involving subjective qualities like agency, a sense of effort, goal-directedness, global self-control, and ownership. Behaviors, on the other hand, are purposeful, but possess no explicit form of conscious goal-representation. They are functionally characterized by automaticity, decreased context-sensitivity, and low self-control; we may not even notice their initiation, but they can be faster than actions. While their phenomenological profile can at times be completely absent, behaviors typically involve the subjective experience of ownership without agency, the introspective availability of goal-directedness varies, and there is frequently a complete lack of meta-awareness.

We find a parallel situation if we look at our inner life; some mental activities are not deliberately controllable, because one centrally important defining characteristic does not hold: they cannot be inhibited, suspended, or terminated. Let us call these activities "unintentional mental behaviors." Mind-wandering can therefore be conceptualized as a form of unintentional behavior, as an involuntary form of mental activity. Viewed in this way, research on mind-wandering is a subfield of human ethology; it belongs within the field of cognitive ethology for *Homo sapiens* (Allen & Bekoff, 1999; Marler & Ristau, 2013).

Of course, the fact that a given mental or bodily behavior is unintentional in no way implies that this behavior is unintelligent or even maladaptive. It is plausible to assume that many animals' minds wander, perhaps a lot of the time. For example, low-level, saliency-driven shifts in attentional focus are unintentional mental behaviors, not inner actions, and in standard situations, they cannot be inhibited. They are initiated by unconscious mechanisms, but may well result in a stable, perceptually coupled first-person perspective as their final stage. Stimulus-independent, task-independent thought, however, normally begins as a form of uncontrolled mental behavior, a breakdown of consciously guided epistemic autoregulation (the active control of one's own epistemic states on the level of high-level cognition). Just like an automatic, saliency-driven shift in the focus of attention, stimulus-independent, task-independent thought may be caused by unconscious factors like introspectively inaccessible goal representations that drive the high-level phenomenology of mind-wandering (Klinger, 2013), for example representations of postponed goal-states that have been environmentally cued by goal-related stimuli under high cognitive load (Cohen, 2013; McVay & Kane, 2009). Of course, quite often an episode of spontaneous thought will be *initiated* in a deliberate manner (see later discussion in this chapter and Seli et al., 2016), but as it unfolds it turns into unintentional mental behavior. Both low-level attention and uncontrolled, automatic thinking will frequently count as an intelligent and adaptive type of inner behavior. Nevertheless, as long as it is taking place, we seem to lack the ability to terminate or suspend it—we are fully immersed in an inner narrative and cannot deliberately "snap out of it." This highlights that perhaps the most relevant and hitherto neglected phenomenological constraint for a theory of mental autonomy is that, subjectively, we do not notice this fact. Therefore, on the functional level of analysis, my positive proposal is that mind-wandering is the graded loss of the ability for veto control on the mental level, which can be described

as a graded loss of mental autonomy and epistemic self-control.

Epistemology of Mental Self-Knowledge

Mind-wandering is interesting for philosophers because it has important implications for theories of self-knowledge. First, every philosophical account of *conscious* self-knowledge now needs to do justice to the discovery that it is a highly discontinuous process, and that this discontinuity is only weakly reflected on the level of conscious experience itself. Second, unnoticed rationality deficits and self-deception from cognitive corruption are possible at any point in time (see Metzinger, 2013b, Example 4; Windt, 2015, p. 479). Clearly, we can have a specific epistemic ability, but we can also temporarily lose our knowledge of possessing this ability. In mind-wandering, the relevant ability is our potential for cognitive self-control, most importantly the very basic and fundamental capacity for what I have called "mental veto control." If this ability is not explicitly represented in our phenomenal self-model, then we—as a whole person—are not able to exert it. We suffer from an epistemic deficit, an absence of representation that is not represented *as* an absence—and the ensuing lack of conscious self-knowledge has well-documented causal consequences.

Recall the notion of having an internal model of "horizontal mental causation." For philosophers this means that one mental event causes another mental event. Closely related to this is the idea of "vertical mental causation," which typically means that a mental event could cause a *physical* event—say, a bodily movement—in a top-down fashion. Many contemporary philosophers think that something like this is not possible (we sometimes call this "the causal closure of the physical," assuming that every physical event that has a cause has a physical cause; see, e.g., Kim, 1993, 2000). But if we take our own phenomenology seriously, we discover that the human brain models mind–body interactions very differently, giving rise to Cartesian intuitions (Metzinger, 2003a, Section 6.4.1). However, there is a third possibility: *intramental* vertical causation, and this term may be another example of a potentially useful and heuristically fecund conceptual instrument for the mind-wandering community. Intramental vertical causation would be the case where one mental event causally influences another mental event, but not—as in the case of horizontal causation within the domain of mental events, as discussed earlier—in terms of continuing a chain of such events, but in terms of terminating such a chain, by top-down control. Let us ignore the philosophical metaphysics of the mind–body problem for now, and just look at the necessary functional architecture in our minds. My point is that in order to know about our ability for mental veto control (i.e., our capacity to terminate or suspend an ongoing train of thought or other mental process), we would first need an inner *model* of the possibility of top-down intramental causation, of one mental event terminating or modulating a chain of events on a lower level. A speculative empirical hypothesis would say that exactly this model disappears in our brains after the onset of a mind-wandering episode.

What makes this phenomenon interesting is that it does not seem to bother us very much, to the point that many of us initially doubt the empirical data on the frequency of attentional lapses and spontaneous, task-unrelated thought. There seems to be a widespread form of "introspective neglect," resembling a form of anosognosia or anosodiaphoria, related to the frequent losses of cognitive self-control characterizing our inner life. Obviously, "widespread" does not mean that *all* instances of task-unrelated thought involve introspective neglect—we know that there is intentional "tuning out" as well as "zoning out," as discussed earlier, that up to 41% of reported mind-wandering can be engaged with intention (Seli et al., 2016, p. 606), and that in certain memory, learning, and problem-solving contexts, reduced cognitive control can even provide a benefit (Amer et al., 2016, p. 907). That said, the phenomenon of mind-wandering is also clearly related to denial, confabulation, and self-deception. I once gave a talk about mind-wandering to a group of truly excellent philosophers, pointing out the frequent, brief discontinuities in our mental model of ourselves as epistemic agents, and one participant interestingly remarked, "I think only ordinary people have this. As philosophers, we just don't have this because we are intellectual athletes!" I think the truth of the matter may be just the opposite: high-performing intellectuals are particularly unaware of their own spontaneous, task-unrelated thoughts. The introspective experience and the corresponding verbal reports of one's own mind-wandering seem to be strongly distorted by overconfidence bias, illusions of superiority, and the introspection illusion (in which we falsely assume direct insight into the origins of our mental states, while treating others' introspections as unreliable). It is probably also influenced (and not only for philosophers of mind) by confirmation bias related to

one's own theoretical preconceptions and culturally entrenched notions of "autonomous subjectivity," by self-serving bias, and possibly by frequent illusions of control on the mental level. This interestingly relates the field of spontaneous thought to other burgeoning and increasingly active areas of research like self-deception (Pliushch & Metzinger, 2015). One positive empirical prediction resulting from this discussion is that at least all of the biases I have listed as examples in the preceding should be considerably weakened in long-term practitioners of mindfulness meditation (Hölzel et al., 2011).

What Exactly Is a "Unit of Identification"?

On the level of content, every onset and every ending of an episode of mind-wandering are characterized by an unexpected shift or sudden switch in the phenomenal "unit of identification" (UI). Here is an example. Let us say that at first you identify with the conscious content of an internal model of the self as currently standing at a red traffic light, waiting for it to turn green. Then an internal simulation of yourself as buying tofu and bananas pops up, as you "remember" that you need to buy tofu and bananas. Now you identify with the protagonist of *this* inner narrative, with the virtual self that constitutes the center of an automatic inner action simulation. Phenomenologically, and for a short moment only, you literally "become someone else." For a brief moment you "zone out" completely, and this constitutes an involuntary and unexpected shift in the UI. Then perceptual coupling may quickly be restored and you re-identify with the "driver," a model of the self as an attentional agent, quickly checking if the lights have turned green. This is the end of your mind-wandering episode. Phenomenologically, the driver is real again, and the shopper is only virtual—the shopper is now *not* the UI any more, but just the retrospective content of a sudden memory leading to a decision and an action plan. Now you may decide to "tune out" again, perhaps to see if an active inner simulation of yourself as buying tofu and bananas "makes other things come to mind." In initiating this, you are an autonomous mental agent. However, in the very moment you "remember" that you also wanted to buy almond butter and raisins, the UI switches again and you quickly "zone out" for a fast update, an enriched mental simulation of the shopper and its now extended task list. This is the beginning of mind-wandering episode number two, and it is functionally characterized by another brief loss of mental autonomy—another bout of "involuntary mental time travel" (Song et al., 2012). This second episode may take less than a second to unfold, and as the light suddenly turns green you "snap back" into the driver model, hastily shifting gears. The "snapping back" is the shifting of the UI, and it is the end of your second mind-wandering episode. There have been two episodes and four switches in the UI.

Mind-wandering is interesting for philosophers because it has great potential for illuminating a deeper understanding of phenomenal self-consciousness and the supra-bodily mechanisms of phenomenal self-identification (Blanke & Metzinger, 2009). Furthermore, if the model I have sketched in the preceding is correct, then progress in empirical research into mind-wandering and the computational modeling of neuroscientific data decisively depend on a better understanding of what exactly a UI is.

Let us say that for every self-conscious system *S* there exists a *phenomenal unit of identification* (UI), such that

- *S* possesses a single, conscious model of reality;
- the UI is a part of this model;
- at any given point in time *t*, the UI can be characterized by a specific and determinate representational content *C*;
- such that *C* constitutes the system's phenomenal self-model (PSM; Metzinger, 2003) at *t*.

If we assume a "predictive processing" model of human brain activity (Clark, 2016; Friston, 2010; Hohwy, 2013), then, for all human beings, *C* is always counterfactual content. The UI ultimately represents the best hypothesis the system has about its own global state. For human beings, *C* is dynamic and highly variable, and it does not have to coincide with the physical body as represented (for an example, see de Ridder, van Laere, Dupont, Menovsky, & van de Heyning, 2007). There exists a minimal UI, which likely is constituted by pure spatiotemporal self-location (Blanke & Metzinger, 2009; Metzinger, 2013a, 2013b; Windt, 2010); and there is also a maximal UI, likely constituted by the most general phenomenal property available to *S* at any point *t*, namely, the integrated nature of phenomenality per se (Metzinger, 2013a, 2013b, 2016). *C* is phenomenally transparent. Internally, *S* models the representational content constituting the UI as neither counterfactual nor veridical, but simply

real. Phenomenally experienced realness is empirical Bayes-optimality; it is an expression of successful prediction error minimization, high model evidence, and counterfactual richness (e.g., invariance under counterfactual manipulation). The UI is the transparent partition of the PSM.

Self-consciousness and the possession of a UI are what make verbal self-reports possible. For some *S*, if *S* has functionally adequate linguistic abilities, it can indirectly refer to itself by referring to *C*, and so generating autophenomenological reports of the type "I *am* this!" Mastery of the first-person pronoun "I" consists in successful linguistic self-reference via the UI. It is a form of displaced reference, because it only directly refers to *C* without *S* being able to experience this fact consciously at *t*. The possession of a UI is the central causally enabling factor for all forms of intelligent behavior, bodily or mental, which presuppose the ability for self-reference. The possession of a UI is conceptually necessary for self-consciousness because self-consciousness *is* phenomenally represented identification, based on counterfactual content, via transparency. Biological systems sustain organismic integrity by preserving the integrity of their UI, constantly trying to minimize PSM-related uncertainty. Thus confabulation, delusion, and functionally adequate forms of self-deception are attempts to sustain the integrity and stability of the UI across time, under exceptionally high degrees of uncertainty.

My last positive proposal for developing a novel conceptual framework is the following: mind-wandering and "spontaneous task-unrelated thought" can be conceived of as an unintentional form of mental behavior, centrally involving involuntary and initially unnoticed shifts in the UI. It is presently unknown whether such shifts serve a biological purpose in all or only in some cases, let alone if there is one general function or specific neurodynamical signature under which all instances of UI-switching can be subsumed. But the general principle would be that distinct episodes of mind-wandering, whether separated by a period of meta-awareness and a regaining of M-Autonomy or not, are always "bracketed" by UI-shifts. Isolating the neural correlates and the dynamic functional mechanisms constituting such "brackets" would constitute an important step forward in describing the temporal boundaries and conceptually individuating single occurrences of our research target. If this is correct, then the more fundamental conceptual insight is that phenomenal self-consciousness is a highly discontinuous process.

Notes

1. Some sections of this chapter strongly draw on Metzinger (2013a and 2015). I want to thank Kalina Christoff and Kieran Fox for helpful comments on an earlier version of this chapter, and Lucy Mayne for equally helpful comments plus excellent editorial help with the English version of this text.
2. The term "spontaneity" plays a role in a number of classical philosophical theories of mind, perhaps most prominently in the theory of Immanuel Kant. Unfortunately, this point would lead beyond the scope of the present contribution. Let me point, however, to an interesting link connecting Kant to our currently best mathematical models of brain function: Presupposing a predictive-processing framework, mind-wandering might also be seen as the expression of a very deep form of "neurocomputational creativity" inherent in the very generative model of reality, which our brains continuously create and update by minimizing free energy (Friston, 2010). Continuous free-energy minimization would then be the creative mechanism that implements what Kant had in mind, when he spoke of "spontaneity." As Robert Hanna writes about spontaneity in Kant, "A cognitive faculty is spontaneous in that whenever it is externally stimulated by raw unstructured sensory data as inputs, it then automatically organizes or 'synthesizes' those data in an unprecedented way relative to those inputs, thereby yielding novel structured cognitions as outputs (B1–2, A50/B74, B132, B152). So cognitive spontaneity is a *structural creativity* of the mind with respect to its representations. [. . .] Kant also uses the term 'spontaneity' in a somewhat different sense in a metaphysical context, to refer to a mental cause that can sufficiently determine an effect in time while also lacking any temporally prior sufficient cause of itself (A445/B473). Call this *practical* spontaneity. What is shared between the two senses of spontaneity, practical and cognitive, is the unprecedented, creative character of the mind's operations" (Hanna, 2016, Section 1.1) Free energy minimization would then be the transcendental condition of possibility for both knowledge and action (see Metzinger & Wiese, 2017).
3. Metzinger (2013a) was the first empirically informed sketch of an explicit, positive model of mind-wandering from the philosopher's camp. A substantial and careful criticism of this model can be found in Irving (2016).
4. "Transparency" is a property of conscious representations, namely, that they are not experienced *as* representations. Therefore, the subject of experience has the feeling of being in direct and immediate contact with their content. Transparent conscious representations create the phenomenology of naïve realism. An opaque phenomenal representation is one that is experienced *as* a representation, for example in pseudo-hallucinations or lucid dreams. Importantly, a transparent self-model creates the phenomenology of identification (Metzinger 2003a, 2008). There exists a graded spectrum between transparency and opacity, determining the variable phenomenology of "mind-independence" or "realness." Unconscious representations are neither transparent nor opaque. See Metzinger (2003b) for a concise introduction.
5. This section strongly draws on Metzinger (2015).
6. Functional properties are abstract properties referring to the *causal role* of a state (the set of its causal relations to input, output, and other internal states), without implying anything about the properties of its physical realization. Just like states described in a Turing machine table or computer software, they are multi-realizable. Since M-autonomy is a functional property, it could in principle also be implemented in a machine.

References

Allen, C., & Bekoff, M. (1999). *Species of mind: The philosophy and biology of cognitive ethology*. Cambridge, MA: MIT Press.

Amer, T., Campbell, K. L., & Hasher, L. (2016). Cognitive control as a double-edged sword. *Trends in Cognitive Sciences*, *20*(12), 905–915. doi: 10.1016/j.tics.2016.10.002.

Antrobus, J. S. (1968). Information theory and stimulus-independent thought. *British Journal of Psychology*, *59*(4), 423–430.

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Science*, *13*(1), 7–13. doi: 10.1016/j.tics.2008.10.003

Bortolotti, L. (2015a). The epistemic innocence of motivated delusions. *Consciousness and Cognition*, *33*, 490–499.

Bortolotti, L. (2015b). Epistemic benefits of elaborated and systematized delusions in schizophrenia. *British Journal for the Philosophy of Science, 67*(3), 879–900. http://dx.doi.org/10.1093/bjps/axv024.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Cohen, A.-L. (2013). Attentional decoupling while pursuing intentions: A form of mind wandering? *Frontiers in Psychology, 4*, 1–9. doi: 10.3389/fpsyg.2013.00693

Davidson, D. (1988/2001). *Essays on actions and events: Philosophical essays*. Oxford: Oxford University Press.

de Ridder, D., van Laere, K., Dupont, P., Menovsky, T., & van de Heyning, P. (2007). Visualizing out-of-body experience in the brain. *New England Journal of Medicine*, *357*(18), 1829–1833.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge: Cambridge University Press.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Forster, S., Nunez Elizalde, A. O., Castle, E., & Bishop, S. J. (2015). Unraveling the anxious mind: Anxiety, worry, and frontal engagement in sustained attention versus off-task processing. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(3), 609–618. doi: 10.1093/cercor/bht248.

Fox, K. C. R., Nijeboer, S., Solomonova, E., Domhoff, G. W., & Christoff, K. (2013). Dreaming as mind wandering: evidence from functional neuroimaging and first-person content reports. *Frontiers in Human Neuroscience*, *7*, 1–18. doi: 10.3389/fnhum.2013.00412

Giambra, L. M. (1989). Task-unrelated thought frequency as a function of age: A laboratory study. *Psychology and Aging*, *4*(2), 136.

Hanna, R. (2016). Kant's theory of judgment. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.https://plato.stanford.edu/archives/win2016/entries/kant-judgment/.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hölzel, B. K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., & Ott, U. (2011). How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. *Perspectives on Psychological Science*, *6*(6), 537–559.

Irving, Z. C. (2016). Mind-wandering is unguided attention: Accounting for the "purposeful" wanderer. *Philosophical Studies*, *173*(2), 547–571.

Kim, J. (1993). *Supervenience and mind: Selected philosophical essays*. Cambridge: Cambridge University Press.

Kim, J. (2000). *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.

Klinger, E. (2013). Goal commitments and the content of thoughts and dreams: Basic principles. *Frontiers in Psychology*, *4*, 1–17. doi: 10.3389/fpsyg.2013.00415

Kühle, L. (2015). Insight: What is it, exactly? In T. K. Metzinger & J. M. Windt (Eds.), *Open MIND*. Frankfurt am Main: MIND Group.

Marler, P., & Ristau, C. A. (2013). *Cognitive ethology: Essays in honor of Donald R. Griffin*. New York and London: Psychology Press.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196–204. doi: 10.1037/a0014104.

Metzinger, T. (2003a). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Metzinger, T. (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*(4), 353–393. doi: 10.1023/B:PHEN.0000007366.42918.eb

Metzinger, T. (2004). Why are identity disorders interesting for philosophers? In T. Schramme, & J. Thome (Eds.), *Philosophy and Psychiatry* (pp. 311–325). Berlin: de Gruyter.

Metzinger, T. (2006). Conscious volition and mental representation: Toward a more fine-grained analysis. In N. Sebanz & W. Prinz (Eds.), *Disorders of volition*. Cambridge, MA: MIT Press.

Metzinger, T. (2008). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. *Progress in Brain Research*, *168*, 215–278.

Metzinger, T. (2009). Why are out-of-body experiences interesting for philosophers? The theoretical relevance of OBE research. *Cortex*, *45*(2), 256–258. doi: 10.1016/j.cortex.2008.09.004

Metzinger, T. (2013a). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, *4*, 1–19.

Metzinger, T. (2013b). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, *4*, 1–17. doi: 10.3389/fpsyg.2013.00746.

Metzinger, T. (2014). First-order embodiment, second-order embodiment, third-order embodiment. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition*. New York: Routledge.

Metzinger, T. (2015). M-Autonomy. *Journal of Consciousness Studies*, *22*(11–12), 270–302.

Metzinger, T. (2017). The problem of mental action. In T. Metzinger and W. Wiese (Eds.), *Philosophy and predictive processing* (pp. 1–26). Frankfurt am Main: MIND Group.

Metzinger, T. (forthcoming). Why is virtual reality interesting for philosophers? *Frontiers in Robotics and AI (Research Topic: The Impact of Virtual and Augmented Reality on Individuals and Society).*

Metzinger, T., & Windt, J. M. (2007). The philosophy of dreaming and self-consciousness: What happens to the experiential subject during the dream state? In D. Barrett & P. McNamara (Eds.), *The new science of dreaming*, Vol 3: *Cultural and theoretical perspectives* (pp. 193–247). Westport, CT: Praeger.

O'Brien, L., & Soteriou, M. (Eds.) (2009). *Mental actions*. Oxford; New York: Oxford University Press.

Pliushch, I., & Metzinger, T. (2015). Self-deception and the dolphin model of cognition. In R. J. Gennaro (Ed.), *Disturbed consciousness: New essays on psychopathology and theories of consciousness* (pp. 167–208). Cambridge, MA: MIT Press.

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences*, *20*(8), 605–617. doi: 10.1016/j.tics.2016.05.010.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, *66*, 487–518.

Song, X., Wang, X., & Krueger, F. (2012). Mind wandering in Chinese daily lives: An experience sampling study. *PLoS ONE*, *7*(9), e44423. doi: 10.1371/journal.pone.0044423

von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(01), 1–16.

Voss, U., & Hobson, A. (2015). What is the state-of-the-art on lucid dreaming? In: T. K. Metzinger & J. M. Windt (Eds/), *Open MIND,* 1–20. Frankfurt am Main: MIND Group. http://open-mind.net/papers/what-is-the-state-of-the-art-on-lucid-dreaming-recent-advances-and-questions-for-future-research.

Wilson, G., & Shpall, S. (2016). Action. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2016/entries/action/.

Windt, J. M. (2015). *Dreaming: A conceptual framework for philosophy of mind and empirical research*. Cambridge, MA: MIT Press.