Tackling the Hard Problem of Consciousness

What It is that Makes Us Phenomenal Subjects of Experience,
From a Rational-Causalist Point of View

Diploma Thesis
phil. nat. Faculty
University of Berne, Switzerland

presented by
Guido Gloor Modjib

to Prof. Dr. Gerd Graßhoff
Department of Philosophy

January 5, 2010

Abstract

What is generally known as "the hard problem of consciousness" is a very fascinating, but also a very disputed matter. It is the one thing we do not seem to be able to understand about our brains: How does a clump of molecules, a mere mass of gray matter, a network of brain cells, produce consciousness?

There are four important parts to this question, all of which warrant a closer look: How does the brain produce phenomenality? What are qualia? The most mystifying part: How does subjectivity emerge? Finally, what is intentionality? With all these in place, how does it all fit together?

During my search for answers to these questions, I encountered two philosophers with interesting proposals: Daniel C. Dennett and Thomas Metzinger.

D. Dennett's works *Consciousness Explained* (1991) and *The Intentional Stance* (1987) break not only with Cartesian dualism, but also with Husserlian phenomenology and its descendants. Metzinger published *Being No One* in 2003 as his first major work in English. He tries to reinterpret phenomenological views in a radically new way, and ends up even dismissing (at least from ontology in a stricter sense) the very thing that seems to be at the core of our subjectivity: the self.

Both Dennett and Metzinger have a somewhat novel approach to the questions at hand, meaning they look at them in an interdisciplinary manner and try to verify their findings in the light of recent psychological and neuroscientific research.

When analyzing subjectivity, which to me seems to be the most miraculous achievement of consciousness, I looked beyond philosophical works and found Marvin Minsky, a pioneer of artificial intelligence, with his book *The Emotion Machine* from 2006. Coming from a computer science background, he has a more pragmatic approach than the aforementioned philosophers.

In this thesis, after looking at those philosopher's positions on the four parts of the hard problem in-depth and attempting to merge their positions on the four parts, I present a synopsis that attempts to unify their theories and categorizes the processes and prerequirements which a system needs to satisfy in order to be able to experience subjective consciousness. In doing so, the thesis splits up the aforementioned four apparently time-proven concepts into a total of seven new concepts where one clearly forms the basis for the next, and by that means untangles overlapping semantic fields.

Acknowledgements

This thesis would not have been possible without plentiful support from many people, both at the University of Berne and in my private life.

First and foremost, I want to thank my wife, Arzo Modjib Gloor, for being such a wonderful person and believing in me all this time. I also want to thank the rest of my family for their support.

At University, I want to thank Prof. Dr. Gerd. Graßhoff for his very constructive criticism. Furthermore, this thesis would not have been possible without the aid of Prof. Dr. Eduard Marbach. I also want to thank Dr. Guido Löhrer and Dr. Helmut Linneweber-Lammerskitten for their support over the years.

My workplace has been very supportive both regarding this thesis and in general. I want to thank Paul Moser, Beat Löffel and Tanja Rottermann for their assistance and for enduring my impossible work hours.

I had very helpful (although unfortunately usually too short) discussions with many friends, in particular: Stefan Aeschbacher, Andrea Borsato, Simon Bünzli, Patrizia Hasler, Mark Hinnen, Andreas Hunziker, and Thomas Ott.

Finally, I would like to say a particularly profound "thank you" to Daniela Reist, proofreader par excellence, and to Simon Bünzli and Stefan Aeschbacher, who cross-read the thesis, found plenty of errors and mistakes, and inspired whole additional chapters.

Contents

Part I
Introduction

Before we can get to the heart of the matter of what consciousness is all about, we need to make sure that we are all on equal footing. Often, philosophers use the same terms with diverging meanings. Even if I take great care not to talk about Wittgensteinian *Scheinprobleme*, it is important to avoid misunderstandings due to linguistic inconsistencies. I attempt to avoid such misunderstandings by defining major terms that are important for the discussion when they first appear, and clarifying my useage of more wide-reaching concepts.

Let us start our journey by finding out what that hard problem is.

1 The hard problem

What is generally known as "the hard problem of consciousness" is a very fascinating, but also a very disputed matter. It is the one thing we do not seem to be able to understand about our brains: How does a clump of molecules, a mere mass of gray matter, a network of brain cells, produce consciousness?

There are four important parts to this question, all of which warrant a closer look: How does the brain produce phenomenality? What are qualia? The most mystifying part: How does subjectivity emerge? Finally, what is intentionality? With all these in place, how does it all fit together?

During my search for answers to these questions, I encountered two philosophers with interesting proposals: Daniel C. Dennett and Thomas Metzinger.

D. Dennett's works *Consciousness Explained* (1991) and *The Intentional Stance* (1987) break not only with Cartesian dualism, but also with Husserlian phenomenology and its descendants. Metzinger published *Being No One* in 2003 as his first major work in English.[1] He tries to reinterpret phenomenological views in a radically new way, and ends up even dismissing (at least from ontology in a stricter sense) the very thing that seems to be at the core of our subjectivity: the self.

Both Dennett and Metzinger have a somewhat novel approach to the questions at hand, meaning they look at them in an interdisciplinary manner and try to verify their findings in the light of recent psychological and neuroscientific research.

When looking at subjectivity, which to me seems to be the most miraculous achievement of consciousness, I looked beyond philosophical works and found Marvin Minsky, a pioneer of artificial intelligence, with his book *The Emotion Machine* from 2006. Coming from a computer science background, he has a more pragmatic approach than the aforementioned philosophers, and will assist us in our strive towards a better overview.

2 The goals of this thesis

I want to focus on how Dennett and Metzinger (and, for subjectivity, Minsky) try to answer the four partial questions I mentioned, and present my arguments on what I perceive to be shortcomings of their theories. I will also attempt to combine all of their points of view into one model, where possible.

Finally, in chapter 7, I form a new model, based on the great research they provide, that might give a more solid foundation for further research.

[1] While I am in no position to criticize Metzinger's mastery of the English language myself, he faced a lot of criticism for his nomenclature. Mainly, it was said that his inventing his own names for novel concepts or novel views on established concepts made it harder to understand what he really means. Nevertheless I will often use his nomenclature and semantics, and attempt to show up what he means by some particularly peculiar expressions. See chapter 3.3.1.

Part II
The four questions

When facing a really challenging problem, it is often best to break it up into smaller, hopefully less difficult problems that can be solved individually – if that is even possible at all. The problem known as 'the hard problem of consciousness,' the question of how it can possibly be that our brain, being a mere clump of cells, can produce a kind of consciousness that develops things like feelings, thoughts, plans and fantasies, is one such difficult problem, and it *is* possible to break it up. However, in doing so we have to embrace and unify concepts that have been subject of philosophical research for decades, subjects where every philosopher has his own views on how they are to be understood.

As I mentioned before, there are four major aspects to it. These four aspects are in some ways arbitrary – I have not found another list that satisfied me, and so I compiled this from the sources I found. However, I feel that they are able to encompass all the important parts that make up the hard problem, while at the same time being fine-grained enough to warrant that we learn something by analyzing them a bit more in-depth. I was inspired by the subjects traditionally chosen by philosophy, and ordered them in apparent increasing complexity. One part is meant to lay the foundation for the next.

Of course, the four subjects cannot be looked at merely in isolation. For the time being, I do not yet intend to challenge the ways in which other philosophers understand one particular concept or another, but rather attempt to understand what different readings there are, what is the core of those concepts and how they are linked together. I postpone looking at the bigger picture to chapter 7, where we will also abandon the ordering of these four parts again and look at the hard problem from a different perspective. The resulting new levels will be placed in slightly different order compared to this part here, too, because some of the intuitions[2] that led me to the order we will see first will have to be revised.

It is important to keep in mind that what seems to be definitions in the following list of concepts really is merely an attempt at summarizing entire philosophical debates on meanings, and thus do not contain any references. Please refer to the respective chapters for my attempts at clearing up what the terms actually mean.

The four provisional concepts I am interested in and the resulting partial questions for now are:

- *Phenomenality:* The special way things are experienced by agents like us. Of course, for a first-person perspective to exist, we need subjectivity, so resolving that apparent circularity is an important part of our look at this subject. Phenomenality, however, also is what enables all the other bits and pieces we need in order to understand how consciousness happens.[3] What is phenomenality, and what entities or processes produce it?

- *Qualia:* The particular non-subdivisable entities of atomic quality we perceive when we experience things.[4] This has been a passion of many philosophers since Lewis and Nagel; there is a pecularity about our experience in that it seems to consist of atomic smallest bits of experience which cannot be subdivided any further – the particular way something seems to be irreducibly red or loud to us, for example. What is so special about these smallest

[2] As a matter of fact, it was more than intuitions, and the genesis of this thesis saw multiple times that I reordered them. The trouble is one that we will only be able to resolve in chapter 7: The four questions depend on each other. There are parts of phenomenality that require subjectivity, while qualia are the basis of phenomenality but also depend on phenomenal experience. Intentionality is, for some, what enables phenomenal experience, but for others, what phenomenal experience forms. With this in mind, I will nevertheless attempt to get a clear picture on what is unique and specific about these (for now, four) parts, and only later try to split them up and build one structure, where no interdependences between the resulting parts are necessary anymore.

[3] I am not merely interested in how it works, but also in how it is that it got to work like that, and what the roots, causes and effects of consciousness are.

[4] This is a particularly heated point of debate, I think Metzinger's qualification that this thesis touches in chapter 4.3.3 is enlightening concerning why philosophers tend to disagree on whether such atomic entities exist or not.

entities of quality? Are they, can they really be atomic and irreducible? How are qualia conceptually definable?

- *Subjectivity:* The way we see everything from a first-person perspective. We experience both ourselves and our fellow humans as agents, and we believe that our fellow human beings experience themselves as persons as well. How is such a first-person perspective generated, and what else is included when someone speaks of "myself"?

- *Intentionality:* The world is given to us, and we relate to it. In introspection, our inner workings are given to us, as seen from a point of view which stands in a relation to these inner workings. How and why is that so? How come that we feel we are in direct metaphysical contact with (for example) an apple tree in the garden by means of some kind of "intentional arrow" or "mental focus"?

I will dedicate a major chapter to each of these four parts, defining what exactly is puzzling about them, what different readings there were, and then looking at possible explanations for their origins and constitution. In doing so, we will also look at philosophers who worked with these concepts and attempt to find out what they really are, at least according to those philosophers' definitions.

When looking for a concise definition of the four big questions of what phenomenality, qualia, subjectivity and intentionality really are, many philosophers made proposals as to what they could be. I want to look at two of the most recent proposals, those of Dennett and Metzinger.

Generally, Metzinger digs deeper and develops traditional ideas in entirely new ways, while Dennett presents important counterpoints to these traditional views and has a more radically reductionist point of view.

Additionally, I will also include the proposals of Marvin Minsky (a true pioneer of artificial intelligence research) in the chapter about subjectivity. Minsky is influenced partially by Dennett, but is a computer scientist and artificial intelligence researcher with a more pragmatic and solutions-oriented background. Thus he brings an interesting and fresh point of view into the debate.

However, all three of them have to stand up compared to more traditional views, thus I decided to include these in my summary as well:

- For *phenomenality*, Edmund Husserl's phenomenology can well be seen as the origin of the idea, and has followers to this very day.

- For *qualia*, we will look at Clarence Irwing Lewis' introduction of the concept. Also, Thomas Nagel might not have named qualia as such, but his work certainly was an important multiplier of the idea that something is special about subjective experience, so we will briefly look at how he introduces the idea as well.

- The traditional view in regards to the origin of *subjectivity* has to be the dualism of René Descartes, which is implicit in many theories to this very day.

- Concerning *intentionality*, Franz Brentano greatly influenced Husserl by fleshing out the traditional idea of "immanent objectivity" in our relationship to the world.

Furthermore, I will add a short consolidatory chapter to all four sections, attempting to make sense of which ideas are the most promising for going forward from here.

After this lengthy introduction, allow me to start with the first major term of traditional phenomenology: phenomenality.

3 Phenomenality

There seems to be something special about the way we see the world and ourselves, something that is maybe ineffable, maybe inexplicable or mystical, but definitely hard to understand. There is something left if we explain intentionality and subjectivity, and subtract them from consciousness

as a whole. I call what is left "phenomenality". Dennett and Searle include more than Metzinger in the notion of "intentionality," resulting in some parts of what I call "phenomenality" being included in "intentionality" – I will stay true to Metzinger's definition of intentionality here, as we will se in chapter 6.

Phenomenality, and I think most philosophers agree on that, is the special nature in which we (and probably other conscious agents) see things, independent of what else there is to consciousness.

I would argue that phenomenality in this sense is the most basic thing that is special about consciousness, and as such it *is* tempting to just attribute it as a special property to intentionality, or to say that what is special about phenomenality is nothing but that experience consists of qualia.[5]

I will try to go another way here, showing what is special about phenomenality itself. Agreeing with Metzinger, many bits and pieces of the following chapters will refer to some event being phenomenal in nature, and here we will find out what I mean by that.

3.1 Husserl: Systematic reflection

Edmund Husserl prompted a radical paradigmatic shift with his works.[6] He meant to do two things: Approach older concepts on what is mental in novel ways, and make subjective experience available for objective study through systematic reflection – standing in Descartes' tradition and taking facts gained through introspection at face value.

Husserl put great emphasis on intentionality as the prime carrier of phenomenality. Unlike Martin Heidegger, who later put a greater emphasis on ontological consciousness independent of intentionality (and offering an explanation for the subconscious that intuitively might seem to be better, although as we will see it is not necessarily more correct), Husserl put the focus on how we might be able to describe our own phenomenal experience, and how events are perceived as being immediately given to our consciousness. Importantly, events are given immediately indeed, but occasionally only partially and inadequately so – Husserl agrees that it is perfectly possible to misrepresent things.

The immediate directedness is given through the three parts of any intentional act as follows:[7]

- Immanent content, which is raw perceptive data.[8]
- Intentional content, which is sense, the way the intentional object is intended (memory, fantasy, expectance, wish, etc.).
- Intentional object, which is the thing the intentional act is directed at (in the case of perception, the actual object).

What really makes out phenomenality for Husserl is that all these parts have to be given for an event in question, otherwise it is not phenomenal. Particularly intentional content is something that intuitively is only possible for systems that have true, intrinsic intentionality.

The main criticism that Husserl had and has to face was that his theory was too subjectivist. However, he never meant to create a subjectivist theory – but I will save this discussion for chapter 3.4.2.

[5] Since qualia are (for now) defined as the smallest possible atomic parts of experience, this seems entirely plausible. More on this in chapter 4.

[6] Husserl's works were mostly unpublished, at least during his lifetime. His work was published posthum starting 1950 as *Husserliana*.

[7] I already used this list in [Gloor Modjib(2008)] – many thanks to Prof. E. Marbach for the great lectures and seminars, and the fruitful discussions that helped me better understand Husserl's approach. I hope I do Husserl's work justice in this thesis.

[8] Raw perceptive data in this sense is everything that reaches the system from the outside, just the way it is first processed by the system's sensory organs, but also already processed by the system's object formative and integrative capabilities and thus perceived as 'an entity now out there in the world.' This is the lowest level of data that can be accessed by introspection in cases of phenomenal perception. The implication of this definition is that the same object, perceived in to different situations, will produce the same intentional object and potentially the same intentional content (if the situations are very similar), but not the same immanent content. See [Zahavi(2003)], pages 24-26.

3.2 Dennet: "Heterophenomenology"

Dennett's heterophenomenology is phenomenology as seen from a third-person perspective. He defines it like this:

> I have defended the hypothesis that there is a straightforward, conservative extension of objective science that handsomely covers the ground – *all* the ground – of human consciousness, doing justice to all the data without ever having to abandon the rules and constraints of the experimental method that have worked so well in the rest of science. This third-person methodology, dubbed heterophenomenology (phenomenology of *another* not oneself), is, I have claimed, the sound way to take the *first* person point of view as seriously as it can be taken.[9]

Essentially, Dennett defines heterophenomenology mostly by what it is not: It is not subjectivist, but intersubjective. It is not first-person pilot studies, but third-person scientific research. He claims that everything that would be phenomenology is intersubjectivizable, thus symbolic, and thus not ineffable. There is no interesting subsymbolic mental content, or at least, none that we could do proper science with.

He does put his position into perspective:

> I suspect that when we claim to be just using our powers of inner *observation*, we are always actually engaging in a sort of impromptu *theorizing* – and we are remarkably gullible theorizers, precisely because there is so little to "observe" and so much to pontificate about without fear of contradiction. [...] Am I saying we have absolutely no privileged access to our conscious experience? No, but I am saying that we tend to think we are much more immune to error than we are.[10]

Ultimately, it comes down to the following: In heterophenomenology, phenomenological reports of individual agents are not to be taken at face value, but as theorist's fictions. Only after collecting and cataloguing them, looking for similarities and differences between different agent's fictions, can a researcher start attempting to explain the existence of these heterophenomenological reports. According to Dennett, that is all phenomenological reports are good for: They are but collections of data and pointers for later research into the true nature of how consciousness happens:

> We organize our data regarding these phenomena into theorist's fictions, "intentional objects" in heterophenomenological worlds. Then the question of whether items thus portrayed exist as real objects, events, and states in the brain – or in the soul, for that matter – is an empirical matter to investigate. If suitable real candidates are uncovered, we can identify them as the long-sought referents on the subject's terms; if not, we will have to explain why it seems to subjects that these items exist.[11]

Notably, for Dennett, heterophenomenology (unlike his notion of traditional phenomenology) is not a research topic, but a methodology. Dennett does not seem to enter the discussion as to what it really is that makes an agent's experience phenomenal – as we will see even better in chapter 4.2.

3.3 Metzinger: "Multilevel constraints"

Metzinger proposes a catalogue of conditions that he calls "multilevel constraints" in order to define what makes the phenomenal experience of consciousness possible and necessary if they apply to any given system.

Any system that satisfies those conditions necessarily has to be conscious – and the brain is just one such system, even though it is the only one we know to date. He does not explicitly

[9] Citation from the online version of [Dennett(2003)], emphasis his.
[10] Citation from [Dennett(1991)], page 68, emphasis his.
[11] Citation from [Dennett(1991)], page 98.

legitimate calling them "constraints." This nomenclature is consistent with his theory however once we realize that they are not primarily about what a system has to be able to do, but importantly also about what boundaries the system's capabilities need to have, in order for the system to be truly conscious. Particularly the transparency constraint that we will meet in chapter 3.3.11 might well have been a prime motivation for that.

The constraints are "multilevel" in that Metzinger attempts to capture all the important levels of description he considers to be important for consciousness research for each of them.[12] These include the phenomenological level, the representationalist level, the informational-computational level, the functional level and finally the physical-neurobiological level. I will further explore these levels and Metzinger's use of the term "constraint" in chapter 3.4.1.

I cannot discuss all those levels for each constraint, but I will try to explain each constraint individually nevertheless, glossing over the different levels and starting each constraint's discussion with a quote from Metzinger's book,[13] and close with a very short synopsis of how a system can satisfy the constraint in question.

I will not argue for the individual constraints here, as that would mostly be merely attempting to duplicate Metzinger's fine work. We will encounter some of the constraints again later in this thesis however, in other contexts.

Important in the context of answering the question what phenomenality is however is the chapter 3.3.16, where we will see how Metzinger uses these constraints in order to differentiate between different levels of phenomenal experience, and how he shows that phenomenality is not an all-or-nothing phenomenon.

First, however, I want to criticize Metzinger's choice of names, and then introduce three other concepts that Metzinger improves our understanding of: Introspection, how consciousness can not be anything but a process, and Metzinger's concept of local supervenience.

3.3.1 Metzinger's choice of names

I adopted Metzinger's numbering and choice of names for all of his constraints. This does at first sight make it unnecessarily hard to understand what exactly he means by some of those names, and my list of Metzinger's constraints in this chapter can also be seen as an attempt to translate the concepts he nominates into a more common philosophical vocabulary.

Various critics have criticized both Metzinger's book[14] and my thesis on the basis of unknown nomenclature and words that do not appear to be English. Metzinger being German does explain, but not necessarily justify his strange choice of constraint names. However, I believe that when introducing novel concepts, it makes sense to name these with novel words as well – and, that once they are named, it makes sense to keep those names.

Furthermore, I thought that the linguistic turn stopped our desparate clinging to words and nomenclature, and opened our minds for appreciation of concepts however they are named. In this spirit, I kept Metzinger's nomenclature for his constraints.

3.3.2 Introspection

Metzinger makes it clear that there is not merely trivial introspection. Rather, there are varying degrees of introspective access to mental states – which results in phenomenally represented information.[15] While I will only use the term "introspection" in a very loose and undefined matter, compared to Metzinger, I still think it is important that we have a short look at his more thorough investigation.

Metzinger uses ordinals to differentiate these different kinds of introspection. They are:[16]

[12] See [Metzinger(2003)], page 110.

[13] In fact, I will stay so close to his argumentation overall that I will occasionally even merely paraphrase him.

[14] The book they criticized was [Metzinger(2003)], where Metzinger introduces the multilevel constraints presented in this chapter.

[15] See [Metzinger(2003)], page 31.

[16] See [Metzinger(2003)], pages 36f. This list is very close to Metzinger's.

Introspection$_1$: "External attention" – representing an internal system state, but referring to a part of the world. This "corresponds to the folk-psychological notion of attention."[17]

Introspection$_2$: "Consciously experienced cognitive reference" – the experience of attending to an object in our environment, while forming a (new or already known) mental concept of it.[18]

Introspection$_3$: "Inward attention" – representing an internal state that also refers to an internal state, a part of the experienced self-model.[19]

Introspection$_4$: "Consciously experienced cognitive self-reference" – this generates self-knowledge and includes all situations in which we think about ourselves as ourselves.[20]

While introspection$_3$ is important for making information phenomenally subjective, introspection$_4$ is probably the most interesting kind of introspection, "the phenomenon of *cognitive self-reference* as exhibited in reflexive self-consciousness."[21] Arguably, it is closest to our intuitive notion of "introspection."

However, as I already mentioned, I do not differentiate between these different kinds of introspection in my thesis. It is merely important to note that there are indeed different levels of introspective access to our mind's own internal workings.

3.3.3 Consciousness as a process

Because the brain is merely an information processing machine,[22] mental states cannot be states in a traditional, rigid sense. Rather, they have to be reformed and reworked continuously through that process, in order to be experienced as transtemporally constant.

With regards to the first step, representation, the argument is somewhat simple:

> The concept of "mental representation" can be analyzed as a three-place relationship between representanda and representata with regard to an individual system: Representation is a process which achieves the internal depiction of a representandum by generating an internal state, which functions as a representatum [...]. The representandum is the *object* of representation. The representatum is the concrete internal *state* carrying information related to this object. Representation is a *process* by which the system as a whole generates this state.[23]

The object of representation can obviously be an internal state or a counterfactual as well.[24] Furthermore, not all entities in the world necessarily have a representation – As we will see in chapter 4.3, there are presentata just like representata. Nevertheless, the most fundamental issue so far is that the nature of those mental states themselves, the representatum, is as dynamic as the nature of the process of the mental representation that leads to them.

Mental states are merely intermediate results of the continuous remodeling process in which our brains are constantly engaged.

[17] Citation from [Metzinger(2003)], page 36. Introspection$_1$ is a subsymbolic metarepresentation of the world model. It is subsymbolic because we cannot consciously access it (see chapter 3.3.17), it is a metarepresentation because it is a conscious access to a representation, and it is the world model because it contains everything we perceive as being part of our world (notably, including the self-model; see chapter 3.3.5).

[18] Introspection$_2$ is the conceptual (symbolic) metarepresentation of the world model.

[19] See chapter 5.3 for how Metzinger fleshes out that self-model. Introspection$_3$ is a subsymbolic metarepresentation of the self-model. We will have a look at Metzinger's definition of that self-model in chapter 5.3.

[20] Introspection$_3$ is a conceptual (symbolic) metarepresentation of the self-model.

[21] Citation from [Metzinger(2003)], page 37.

[22] I argue for this point of view in chapter 6.4, and it is important to note that both Dennett (see chapter 6.2.3) and Metzinger (without that implicit premise of his, this chapter would look entirely different) agree.

[23] Citation from [Metzinger(2003)], page 20.

[24] See [Metzinger(2003)], pages 43f.

> Human brains function in a similar way [to a flight simulator]. From internally represented information and utilizing continuous input supplied by the sensory organs they construct an internal model of external reality. This global model is a *real-time model*; it is being updated at such a great speed and with such reliability that in general we are not able to experience it *as* a model anymore.[25]

Mental states are not the most fundamental entities that are experienced as ontologically given, however; rather, phenomenal presentational content is. Chapter 4.3.5 elaborates on this circumstance in closer detail. For now, it is important to realize that this presentational content is what drives the processes which form mental states in the first place:

> Presentational content will always be an important element of any such explanation [of a full-blown untranscendable reality model], because it is precisely this kind of mental content that generates the phenomenal experience of presence, of the world as well as of the self situated in this world.[26]

3.3.4 Local supervenience

Local supervenience in humans, the way Metzinger uses it, is defined by the content supervening on only some brain centers and not the entire brain, while being at least partially independent from the environment. The first time Metzinger mentions the concept is as follows:

> First, for phenomenal content the "principle of local supervenience" holds: *phenomenal* content is determined by internal and contemporaneous properties of the conscious system, for example, by properties of its brain. For *intentional* content (i.e., *representational* content as more traditionally conceived) this does not have to be true: *What* and *if* it actually represents may change with what actually exists in the environment. At the same time the phenomenal content, how things subjectively *feel* to you, may stay invariant, as does your brain state.[27]

Metzinger later refines the concept *en passant*:

> According to the principle of local supervenience, for every kind of phenomenal content in humans there will be at least one minimally sufficient neural correlate.[28]

Mental states and phenomenal states, the things that supervene on brain states, are merely abstract conceptual content. Furthermore, this abstract conceptual content is always part of a process.[29] It is merely data that is being sent from one one functional center in the brain to another, modified by the various algorithms that are wired in the brain.

Seen on the physical layer, mental states are 'realized' in nothing but electrical currents running between neurons and clusters of neurons, modified and controlled by the neural network layout that these neurons have.

Phenomenal states are in a causal two-way relationship with the brain states they supervene on: The processes that contain the abstract conceptual content influence the brain states they supervene on, because the phenomenal states are realized in these brain states. The brain states on the other hand influence the abstract conceptual content, because the processes are executed by means of changing physical states in the brain.[30]

With these important foundations in place, let us now turn towards Metzinger's multilevel constraints (using his nomenclature for the constraint's names), and try to understand them. In chapter 3.3.16 we will then see which ones among them really are necessary for consciousness to what degree.

[25] Citation from [Metzinger(2003)], page 555.
[26] Citation from [Metzinger(2003)], page 98.
[27] Citation from [Metzinger(2003)], page 13.
[28] Citation from [Metzinger(2003)], page 170.
[29] See chapter 3.3.3.
[30] Essentially, this is a kind of conceptual dualism. See chapter 8.2.

3.3.5 Constraint 1: "Global availability"

> Phenomenally represented information is precisely that subset of currently active information in the system of which it is true that it is globally available for deliberately guided attention, cognitive reference, and control of action [...].[31]

Metzinger is talking about global availability of information for conscious processes in a phenomenally experiencing system. The information is *globally* available because every part of consciousness has access to them. According to Metzinger, there are three basic kinds of global availability:[32]

- Global availability for deliberately guided attention: In order to inspect things more in-depth, we can guide our attention towards certain aspects of ourselves or our environment.
- Global availability for cognitive reference: We can think about things, we can form mental concepts of them, we can put them into relations with past experiences or plan actions involving them.
- Global availability for control of action: We can catch incoming balls, push buttons when stimuli occur, and also can act according to plans we have formed beforehand (which are then globally available as well).

In addition to the above, Metzinger also mentions speech control, autobiographical memory, phenomenal cognition and others[33] – all of which are merely subcategories of the three more basic kinds however.

Every such kind of global availability is likely to have different neural correlates, but their whole is experienced as embedded into one complete world model and available to processes of consciousness anyway. The world model is all there is for a conscious mind, and its boundaries are the boundaries of the agent's reality.

Metzinger notes that important points about global availability are the following: flexibility, selectivity of content, and a certain degree of autonomy.[34]

The thus specified global availability enables the large number of specialized modules of our brain to interact with those phenomenal states that are globally available. Vice versa, it provides an interface to many modules for new such phenomenal states that enter the system's world model. This allows the system to react flexibly and quickly to new threats and challenges from its environment.

Occasionally, some sensory contents might not be available for cognitive reference and concept formation (for example, we cannot form a mental concept of "exactly this colour"[35]). Occasionally, they are not even available for deliberately guided attention in early processing stages (like when we pull away a hand from a flame "instinctively").[36] Contents can be transparent[37] or opaque.

However, they are all available, while they are experienced or remembered or imagined, to our entire consciousness, to all the other processes running in the brain – or, at least, that is how we experience it phenomenally.

The "global availability" constraint thus means, simplified: A system that satisfies the "global availability" constraint needs to have access to all the information available in any part of the system from all its active processes.

[31] Citation from [Metzinger(2003)], pages 117f.
[32] See [Metzinger(2003)], page 124.
[33] See [Metzinger(2003)], page 118.
[34] See [Metzinger(2003)], page 119.
[35] See chapter 4.3.2.
[36] I will discuss different kinds qualia, with different degrees of global availability of sensory contents, in chapter 4.3.3.
[37] See chapter 3.3.11 for what Metzinger means with "transparency."

3.3.6 Constraint 2: "Activation within a window of presence"

> The experience of presence coming with our phenomenal model of reality may be the central aspect [...]: It is, as it were, the temporal immediacy of existence *as such*. If we subtract the global characteristic of presence from the phenomenal world-model, then we simply subtract its existence. [...] Only persons possessing a subjective Now are *present* beings, for themselves and others.[38]

The importance of the phenomenal, subjective Now is that it is a requirement for a concept of time (which is "constituted by a series of important achievements"[39]), and also a prerequisite for constraint 5, dynamicity. A system that does not have a concept of Now is not in a present, and consequently cannot produce a phenomenal experience with a concept of past or future presents, nor is it able to produce a phenomenal experience of a concept of coherence at all.

Furthermore, the concept of Now is needed for the experience of presentation; only if we can experience something as being presented right now we can experience it as being presented at all. Metzinger thus also calls this the "presentationality constraint."

> If events are not only represented as being in temporal succession but are integrated into temporal figures [...], then a present emerges, because these events are now internally connected. They are not isolated atoms anymore, because they form a context for each other. [...] something like object formation takes place in which isolated events are integrated into a Now. One can describe the emergence of this Now as a process of segmentation that separates a vivid temporal object from a temporal background that is only weakly structured. This can, for instance, happen if we do not experience a musical motive as a sequence of isolated sound events, but as a holistic temporal figure. A psychological moment is not an extensionless point, but for beings like us it possesses a culturally invariant duration of, maximally, three seconds.[40]

For human beings, the concept of Now thus has a duration of maximally three seconds which is culturally invariant and thus must have biological causes. This proposal of Metzinger fits in nicely with neurophysiological research for example by E. Pöppel,[41] which was summarized as follows by V. Evans:

> Pöppel (1994) has argued that two kinds of perceptual moment can be distinguished. The first, PRIMORDIAL EVENTS, which last for a fraction of a second, serve in effect as a 'linking activity,' to integrate or bind spatially distributed information in the brain between and within different modalities. This facilitates the integration of spatially-distributed sensory information as primordial events, e.g., the perception of an object in which visual input, auditory input and information from other modalities are integrated into a coherent percept. The second kind, the perceptual moment with an outer range of 2-3 seconds, serves to link these primordial events into a coherent unity, which, he argues forms the basis of our concept of the present.
>
> According to Pöppel, perceptual moments in the 2-3 second range involve what he terms TEMPORAL BINDING (as opposed to the binding spatially-distributed activities). He proposes that it is the perceptual moment of approximately 2-3 seconds to which the concept of the present (our experience of now) can be traced.[42]

Metzinger thus talks about a very specific meaning of Now when he introduces this three seconds rule – three seconds appear to be the longest possible timespan that we can possibly still experience as being part of the same phenomenal moment.

[38] Citation from [Metzinger(2003)], page 126, emphasis his – I will employ his capital Now to denote the concept of this window of presence.
[39] Citation from [Metzinger(2003)], page 126.
[40] Citation from [Metzinger(2003)], page 127.
[41] See [Pöppel(1994)].
[42] Citation from [Evans(2003)], page 26.

As Metzinger also points out, a complete physical description of the universe would not have to include any notion of Now – the concept merely creates "temporal internality" for organisms, which is a "highly successful and functionally *adequate* fiction"[43] that allows us to understand causality and the subjectively experienced flow of time by situating us not only in a world, but also in a present.[44] Or, as Einstein put it:

> People like us, who believe in physics, know that the distinction between past, present, and future is only a stubbornly persistent illusion.[45]

The "activation within a window of presence" or "presentationality" constraint thus means, simplified: A system that satisfies the "activation within a window of presence" constraint needs to experience current events and phenomenal impressions as happening, and it needs to incorporate these into a window of presence (the phenomenal Now) as happening together right now.

3.3.7 Constraint 3: "Integration into a coherent global state"

> If and only if a person is conscious, a world exists for her, and if and only if she is conscious can she make the fact of actually living *in* a world available for herself, cognitively and as an agent. Consciousness [...] makes situatedness globally available to an agent.[46]

This is the generalized case of constraint 1, global availability. If there are facets of our world model that are globally available, these facets have to be perceived as something global. The entirety of all those facets that are perceived as something global is what we call the coherent global state – it is perceived as being the same, all throughout all modalities.

This means that these facets must be perceived as being parts of one global whole in the sense that they all are globally available – Because they all are available to all brain centers at the same time, their entirety is what is necessarily perceived as a global whole. This perceived global whole is what constitutes the world, or rather, the world model.[47]

This constraint is absolutely fundamental: Being situated in a coherent representation of a world allows us to see the world naively-realistically as "my single world" from a first-person perspective. Since consciousness cannot be anything but a process,[48] we (and other conscious systems) must be constantly remodeling that world model.

It is important to see that "what is at issue is not knowledge, but the structure of experience"[49], and that is more or less what Descartes describes with the concept of indivisibility, the unity of consciousness. We aren't aware that it is only a model, a representation.[50] Yet we experience the world as a lived real whole, with the building blocks not being elements, but parts with a multitude of part-whole relations of that reality. And, the world we live in is completely indistinguishable from one moment to the next and thus experienced as being always the same.

---

[43] Citation from [Metzinger(2003)], page 128, emphasis his.

[44] See [Metzinger(2003)], page 129.

[45] Citation from Albert Einstein's letter to the family of his lifelong friend Michele Besso, after learning of his death, as quoted in [Kowalsky(2003)].

[46] Citation from [Metzinger(2003)], page 131, emphasis his.

[47] The same reservations that Minsky introduces in [Minsky(2006)], as we will see in chapter 5.4.2, seem to apply to world models just like models of the self: It is perfectly reasonable to assume that agents even in non-pathological cases experience not just one, but several world models, and are able to switch between them as needed (for example, a rational and a supernatural world model). This does not invalidate Metzinger's constraint 3 however, as in non-pathological cases, those world sub-models will still be perceived as being part of one coherent whole world model. Nevertheless, this subject will be revisited in chapter 5.5.5.

[48] See chapter 3.3.3.

[49] Citation from [Metzinger(2003)], page 132.

[50] See more on the issue of our unawareness of this fact in chapter 3.3.11 about transparency.

Having just one single world model is a pretty efficient way of reducing chaos from incoming sensory data,[51] reducing ambiguity, with the side effect of also reducing data and thus computational load. This is supported by empirical evidence: If two sources of contradictory information are made available to a human through different sensory channels,[52] only one world model emerges. Planning thus becomes possible against the background of what Metzinger calls the "world zero", the perceived (as unquestionably immediately) real world.

The "integration into a coherent global state" constraint thus means, simplified: A system that satisfies the "integration into a coherent global state" constraint needs to perceive of its surroundings as belonging to one global coherent whole, a "world zero" that is given to the system as unquestionably real.

3.3.8 Constraint 4: "Convolved holism"

> Conscious experience itself can be described as a phenomenon possessing a hierarchical structure, for instance, by being composed of representational, functional, and neurobiological entities assignable to a hierarchy of levels of organization.[53]

Convolved holism as a concept is not all too complicated. Simply put, it speaks of the world as being convolved (as in 'having parts within parts') and a holism (as in 'being an entire complete whole'). Everything is part of the big world model, yet within this, there are substructures; it is divided into concepts, objects, regions and a multitude of other, overlapping groupings.

These substructures can be multimodal, including vision, hearing, time, even social contexts. Metzinger calls them "levels of phenomenal granularity,"[54] and explains their mereological, hierarchical nesting like this: "On lower levels of phenomenal granularity different aspects may be bound into different low-level wholes [...], but ultimately all of them are parts of one and the same global whole."[55]

For convolved holism to be possible at all, it is necessary that we have one all-encompassing world model, and we experience all those parts and wholes as being in the phenomenal Now. Thus constraints 2 and 3 are prerequisites for convolved holism – one could even say that convolved holism is a natural extension of the coherent global state. On the other hand, it is perfectly thinkable for a system to have a whole world model that is not partitioned at all, not even into the self-world distinction which is so important for us humans, leading to "a noncentered, maximally simple model of reality."[56] Consequently, it makes sense to have "convolved holism" in a unique constraint and not working it into one of the earlier ones.

The same set of part-whole relations of course applies to causal contexts in our environment, which makes convolved holism a functionally very adequate solution for the requirement for systems to be able to react quickly to small shifts in their perception of their environment, as well as for shifts in those system's perceptions of themselves – so in other words, introspection is essentially nothing but a special kind of experienced convolved holism.

The "convolved holism" constraint thus means, simplified: A system that satisfies the "convolved holism" constraint needs to be able to subpartition the world it is in (conceptually, or in terms of time and place) into smaller wholes, which in turn can be further subpartitioned. Such partitions must still be experienced as a whole, but with potential further parts.

[51] 'Sensory data' here denotes all the data which is directly produced by sensory input of any kind, the data produced by the nerve endings in sensory organs. The distinction between sensory data and other kinds of data is not enormously important however, particularly once the "offline activation" constraint is introduced in chapter 3.3.12. For lower-level processes in the brain, it is all just data anyway (see chapter 8.2). However, what remains important is that sensory data is what drives the "world zero," it is the only kind of data that is (in non-pathological cases) perceived as unquestionably and immediately given.

[52] See [Metzinger(2003)], page 136, where Metzinger refers to [Leopold and Logothetis(1999)].

[53] Citation from [Metzinger(2003)], page 143.

[54] Citation from [Metzinger(2003)], page 144.

[55] Citation from [Metzinger(2003)], page 145.

[56] Citation from [Metzinger(2003)], page 147.

3.3.9 Constraint 5: "Dynamicity"

> Our conscious life emerges from integrated psychological moments, which, however, are themselves integrated into the flow of subjective time.[57]

Just like convolved holism is a natural extension of the coherent global state, dynamicity is a natural extension of the presentationality constraint, constraint 2. If we have a phenomenal, temporal Now, we can easily perceive of more Nows in both past and future. This must not necessarily be the case for a system, but if it is, it allows for phenomenal concepts of duration and change, and generally perception of the flow of time. The concept is not easy to grasp, and Metzinger has to resort to analogies:

> It is not as if you see the clouds drifting through a window, the window of the Now. There is no window frame. It is not as if the Now would be an island emerging in a river, in the continuous flow of consciously experienced events, as it were – in a strange way the island is a *part* of the river itself.[58]

The subjective experience of time is dependent on attention and representational content, but the core issue with time perception seems to be duration, permanence, the transtemporal identity of objects for the system. Everything else we experience is always perceived as being a property of such experientally presented, dynamical objects.

It is not easy to form a concept of the perception of dynamic time beyond this. Steps leading to a lower-level representational explanation of this concept would include finding how events are individuated, then finding how patterns and sequences of such events are formed. Already those two steps are controversial however, and Metzinger surrenders when faced with the aforementioned transtemporal identity problem.[59]

The "dynamicity" constraint thus means, simplified: A system that satisfies the "dynamicity" constraint needs to perceive of time as something that flows from the future to the past, and of the phenomenal Now, the window of presence, embedded in it. It also needs to be able to model transtemporal identity of objects within this perception of time.

3.3.10 Constraint 6: "Perspectivalness"

> In order to meet this constraint, one needs a detailed and empirically plausible theory of how a system can internally represent itself *for* itself, and of how the mysterious phenomenon today called "first-person perspective" by philosophers can emerge in a naturally evolved information-processing system. Subjectivity, viewed as a phenomenon located on the level of phenomenal experience, can only be understood if we find comprehensive theoretical answers to the following two questions. First, what is a consciously experienced, phenomenal *self*? Second, what is a consciously experienced phenomenal *first-person perspective*?[60]

Metzinger points out that this is among the most interesting phenomena of phenomenology, and I tend to agree. Perspectivalness, the specific nature of how our perception is always from a conscious and self-aware first-person point of view, was among the chief things that made me read Metzinger's work.

Earlier, I had been disappointed concerning what exactly subjectivity is while reading Dennett. He merely glosses over the functional aspect of this interesting point (as we will see in chapter 5.2), and does not offer a real solution to the problems it poses.

Metzinger makes out several levels of subjectivity, the total of which makes up perspectivalness:

[57] Citation from [Metzinger(2003)], page 151.

[58] Citation from [Metzinger(2003)], page 153, emphasis his.

[59] See [Metzinger(2003)], page 154, where Metzinger writes: "The core issue, for which I have no proposals to make, clearly seems to consist in [...] internally representing the *permanence* of already active phenomenal wholes."

[60] Citation from [Metzinger(2003)], page 157, emphasis his.

1. The proto self-model that Metzinger does not make entirely explicit, which provides the basis for the next step. "It is a subjectively immediate and fundamental form of nonconceptual self-knowledge preceding any higher forms of *cognitive* self-consciousness [...] constituted preattentively, and automatically on this most fundamental level."[61] The next step necessarily requires a more basic automatic and fundamental self-consciousness, as is explained in chapter 5.5.3.

2. The phenomenology of being someone, of consciously experienced selfhood, is the first explicit level of subjectivity – it is potentially both cognitive and conceptual. This is what Metzinger's theory on the phenomenal self-model (PSM, see chapter 5.3 and in particular chapter 5.3.2) is all about. There is an epistemic asymmetry that only appears on this level of representational organization: In order to fully explain consciousness, we also need to explain why it is epistemically irreducible[62] – a point where Dennett would probably disagree, since he does not believe in irreducibility in any form.

3. The second explicit level of subjectivity is the phenomenal property of perspectivalness itself, a structural feature of phenomenal space as a whole. It is "a model of the system as *acting and experiencing*"[63] intentionally, which will be discussed in chapter 6.3 – the phenomenal model of the intentionality relation (PMIR).

4. On the third explicit level of subjectivity, social cognition (the untranscendable "we") is only possible if there is a phenomenal first-person perspective (the untranscendable "me"). Chapter 6.4.3 elaborates in detail on social cognition.

Functionally, there is an embodiment constraint in all conscious organisms we know to date: They are experientially centered in that they only have one physical body and limited reach. Consequently, representing this small region of space (being a region of maximal stability and invariance) as self "enormously enriches and differentiates the functional profile of an information-processing system, by enabling it to generate an entirely new class of actions – actions directed toward *itself*."[64]

Interestingly, perspectivalness does appear among the other constraints here, but in fact partially follows from some of the other constraints by necessity.[65] While I can see the justification for placing it among the other constraints, I believe that Metzinger could have stated this more clearly in his book.

Metzinger lists 11 constraints, yet there are actually merely 10 plus the aforementioned prereflexive proto self-model .The PSM and PMIR follow from them and form the perspectivalness constraint.

[61] Citation from [Metzinger(2003)], page 158, emphasis his.

[62] Metzinger makes this epistemic irreducibility explicit in [Metzinger(2003)], page 159:

> Even if one is convinced that phenomenal content can be ontologically reduced to some set of functional brain properties or other, we still need an answer to the question as to why it obviously remains *epistemically* irreducible. What kind of knowledge *is* perspectival, first-person knowledge?

The miracle is solved in [Metzinger(2003)], page 575, after the PMIR (phenomenal model of the intentionality relation, see chapter 6.3) is introduced:

> Phenomenal content is epistemically irreducible, because – in standard situations – it is integrated into a global model of reality *structured* by a PMIR. The special and hitherto somewhat mysterious fact that the phenomenal character of conscious states seems to constitute an irreducible first-person form of content can be reduced to the fact that this character is typically represented *under a PMIR*. And *this* way of gaining knowledge about your own mental state certainly is irreducible to, say, any scientific procedure producing knowledge about its neurofunctional correlate. It is *another* way of gaining knowledge – one that existed long before philosophy and science came into being.

[63] Citation from [Metzinger(2003)], page 159, emphasis his.

[64] Citation from [Metzinger(2003)], page 161, emphasis his.

[65] Deducing this is what his chapters about the PSM and the PMIR, and consequently chapters 5.3 and 6.3 in this thesis, are about.

The "perspectivalness" constraint thus means, simplified: A system that satisfies the "perspectivalness" constraint needs to phenomenally perceive of itself as a self from a first-person perspective. In addition, the system has to be able to model a self and an intentionality relation reaching to arbitrary object components in an asymmetrical relationship to the self.

3.3.11 Constraint 7: "Transparency"

> Transparency in this sense is a property of active mental representations already satisfying the minimally sufficient constraints for conscious experience to occur. For instance, phenomenally transparent representations are always activated within a virtual window of presence and integrated into a unified global model of the world. The second defining characteristic postulates that what makes them transparent is the *attentional unavailability of earlier processing* stages for introspection.[66]
>
> For any phenomenal state, the degree of phenomenal transparency is inversely proportional to the introspective degree of attentional availability of earlier processing stages.[67]

Metzinger's definition of transparency represents a novel approach. According to the standard definition, vehicle properties of certain mental states are not available for introspection whereas content properties are.[68] However, as Metzinger points out, there clearly are cases in which what formerly were vehicle properties (not physical vehicle properties, but for example thoughts or emotions) can become opaque and are available for introspection as well. There are different degrees of transparency according to how much of those vehicle properties is currently available for introspection.

Furthermore, vehicle and content are not as easily distinguishable as it seems, but are rather different aspects of the same ongoing process – so attempting to split them apart bears "subtle residues of Cartesian dualism"[69], which makes definitions of transparency using the terms of vehicle and content less desirable. As Metzinger writes, "for every kind of phenomenal content in humans there will be at least one minimally sufficient neural corellate."[70]

Metzinger goes on to describe common potential misunderstandings of the concept:[71]

1. Transparency is a phenomenological concept, not an epistemological notion. So it is possible to be wrong about one's own mental contents; In fact, that happens rather often in pathological cases. It's however not possible to be aware of all of one's own processes when thinking about one's own consciousness, exactly because the core processes are hidden from introspection.

2. Phenomenal transparency is subsymbolic and used as a concept in philosophical neurophenomenology, not as a part of formal semantics. As such it can exist in nonlinguistic creatures and is not a property of context nor referential in an ontological sense. Any conscious being can have (and necessarily has, because transparency will be shown to be a prerequisite for phenomenality in chapter 3.3.16) transparent mental states.

[66] Citation from [Metzinger(2003)], page 165, emphasis his.
[67] Citation from [Metzinger(2003)], page 165.
[68] Of course, what exactly constitutes these vehicle and content properties is a matter of definition as well. Trivially, in this context, vehicle properties are everything physical, while content properties are everything mental.
[69] Citation from [Metzinger(2003)], page 166.
[70] Citation from [Metzinger(2003)], page 170. Also see chapter 3.3.4 for local supervenience.
[71] See [Metzinger(2003)], pages 166ff. Metzinger, in addition to these two, also talks about how phenomenal transparency is not equivalent to the kinds of transparency implemented in transparent technical systems like proxy servers or email servers on the internet. Those systems have "untranscendable" (I'm using quotes here because there is no subject that could mean to transcend them) internal mechanisms where "user information may well be internally changed and reprocessed in many different ways, but is always retransformed into the original format before reaching the output stage *without causal interaction with the user*." (Citation from [Metzinger(2003)], page 168, emphasis his.) These kinds of technical transparency are however not phenomenal and thus – at least until this changes – not interesting for the discussion. Personally, I am not convinced that this is as common a misunderstanding as the other two.

Transparency, for a system that employs it, seems to imply the existence of the entities represented. This leads to the conclusion that a naive realism concerning these entities is intuitively very attractive for us humans by necessity – after all, phenomenal realism is the way we experience the world. We do not see what we are looking through, the medium that the experience takes place in is indetectable for introspective access.

Metzinger calls transparency "a special form of darkness."[72] The opposite of transparent states are opaque ones – they are those where darkness is made explicit, where we represent that something is merely a representation,[73] for example in lucid dreams or obvious hallucinations, or when we become aware of emotions driving our thoughts, or even when we plan or imagine or remember.

This constraint can be summed up like this:

- In standard, transparent cases of experience, we know while knowing that we know.
- In opaque cases of experience, we know while knowing that we could be wrong.

The "transparency" constraint thus means, simplified: A system that satisfies the "transparency" constraint needs to be unable to introspectively access all the processes that build its consciousness. Its surroundings, the "world zero" must be directly given. Transparency can be gradual, some (opaque) processes can be available, or made available, to the system via introspection.

3.3.12 Constraint 8: "Offline activation"

> Phenomenal *simulations* [...] are generated in a way that is largely independent of sensory input. Higher-order, that is, genuinely *cognitive* variants of conscious contents in particular, can enter in that way into complex simulations: they are generated by such simulations.[74]

A system who satisfies this constraint is able to run mental simulations of alternate realities in the widest sense. It is able to use its processing capabilities, including those responsible for "motor-to-sensory transformations in terms of bodily actions and their bodily consequences,"[75] independent of its sensory organs.

This enables both memory and future planning ("internal representation of goal states"[76]) and, generally, counterfactuality (possibility versus reality) and internal experiments. The implications of this are far-reaching. This constraint does enrich the mental and phenomenal capabilities of any system if it is satisfied, and it is the prime prerequisite for mental agency and personhood (but not necessarily phenomenality):

> Mental agents are systems deliberately generating phenomenally opaque states within themselves, systems able to initiate and control ordered chains of mental representations and for whom *this* very fact is cognitively available. Mental agents are systems experiencing themselves as the thinkers of their own thoughts. They can form the notion of a "rational individual", which in turn is the historical root of the concept of a *person*.[77]

Interestingly, and as expected under naturalist assumptions, many of the same physical and neural structures seem to be used for both online and offline activation of mental contents.

[72] Citation from [Metzinger(2003)], page 169.
[73] Because this makes opaque phonomenal states representations of representations under the aspect of awareness, opaque phenomenal states are *meta*representations by necessity.
[74] Citation from [Metzinger(2003)], page 179, emphasis his.
[75] Citation from [Metzinger(2003)], page 183.
[76] Citation from [Metzinger(2003)], page 181.
[77] Citation from [Metzinger(2003)], page 180, emphasis his.

The "offline activation" constraint thus means, simplified: A system that satisfies the "offline activation" constraint needs to be able to form models and representations independant from external inputs.

3.3.13 Constraint 9: "Representation of intensities"

> What [...] qualia have in common is that they vary along a continuous dimension of intensity. This variation on the level of simple content is a characteristic and a salient feature of consciousness itself. [...] It is only satisfied in the domain of simple and directly stimulus-correlated conscious content.[78]

Metzinger explicitly excludes metaphorical uses of intensity relations and only includes those associated with the experience of qualia.[79] For color perception, the intensity dimension is brightness, for sound, it is loudness – but it always seems to be the most fundamental phenomenal dimension associated with a certain perception, and there is no simple sensory content that does *not* possess an intensity parameter. Metzinger calls it the sensory content's "analogue representation."[80]

In some cases, it is even possible for us to experience intensity without associated dimensions. For color, that would be a human experiencing pure brightness without hue and saturation – Metzinger refers to Ganzfeld experiments,[81] where subjects experience a colorless, formless visual experience after looking at a uniform field of color (said Ganzfeld) for some minutes.

What is not possible, however, is that there is no analogue representation of a certain stimulus, due to biological reasons: Stimuli are perceived through physical sensors that detect certain modal qualities along with their intensities. Not detecting something (for example, no light in a dark room) merely has the associated sensors detect an intensity of nearly zero (below the lowest detectable threshold).

Since intensity directly infers signal strength, it is highly adaptive for an organism to make those detected intensities globally available, too, and some qualities (like pain) are able to fixate the system's attention on certain specific aspects of its world model:

> To again give a concrete example, the biological function of consciously experienced *pain* can convincingly be interpreted as "attention fixation" – it locks the organism's attention onto whatever part of its own body it is that has been damaged.[82]

The "representation of intensities" constraint thus means, simplified: A system that satisfies the "representation of intensities" constraint needs to represent an intensity dimension of phenomenally experienced entities. For each such represented entity, this intensity dimension must be the prime defining characteristic.

3.3.14 Constraint 10: "The homogeneity of simple content"

> Just like the intensity constraint, the homogeneity constraint now to be introduced is only satisfied in the domain of phenomenal presentata. [...]
>
> [...] the phenomenological predicates that refer to homogeneous presentational content as if they were referring to a cognitively available *property* seem to introduce a further simple property that apparently cannot be reductively explained. It is the internal,

[78] Citation from [Metzinger(2003)], page 184.

[79] See chapter 4.3 for how Metzinger analyzes and classifies different kinds of qualia.

[80] Citation from [Metzinger(2003)], page 185.

[81] See [Metzinger(2003)], pages 101f, where Metzinger refers to [Hochberg et al.(1951)Hochberg, Triebel, and Seaman].

[82] Citation from [Metzinger(2003)], page 187. Also interesting in this context, as tangential as they are, are the elaborations of Minsky in [Minsky(2006)], page 66, on how pain leads to suffering:

> Any pain will activate the goal *"Get rid of that pain"* – and achieving this will also make that goal go away. However, if that pain is intense and persistent enough, this will arouse yet other resources that tend to suppress your other goals – and if this grows into a large-scale "cascade," there won't be much left of the rest of your mind.

> structureless *density* of simple phenomenal color experiences and the like that has traditionally supported antireductive theoretical intuitions.[83]

This constraint is also called "Ultrasmoothness". What Metzinger refers to by this constraint is what Sellars called the grain problem[84] – the ultimately uniform experience of things like color in a certain area, an absence of internal structure. Sellars writes:

> Putting it crudely, colour expanses in the manifest world consist of regions which are themselves colour expanses, and these consist in their turn of regions which are colour expanses, and so on; whereas the states of a group of neurons, though it has regions which are also states of groups of neurons, has ultimate regions which are *not* states of groups of neurons but rather states of single neurons.[85]

However, as Metzinger points out, the entire area in question appears as directly given, it can well be called atomic itself, since it appears not to be further divisible. As soon as we start thinking about subregions, the concept of those subregions starts forming (they are relative to the curently employed representational architecture), but before that, there just is the entirely transparent concept of the simple, ultrasmooth, representationally atomic, one-color surface.

As Metzinger says, "these features appear as directly given and offer themselves to an interpretation as intrinsic, irreducible, first-order properties."[86] Consequently, there is not really a grain problem, because the experience of such surfaces itself is grainless and does *not* have all those subregions that would need to have neural analogons.

This partially revolutionizes the concept of causal roles of sensory input. Due to the fact that we need to represent entities as wholes because otherwise we would necessarily represent their parts as well, in order for any perception (or more precise, any representational state) to have a causal role at all, it has to be homogenous. Homogeneity is "the expression of success of the integrational processes in the brain,"[87] processes that lead to world that is perceived as being coherent in the first place. Note the parallels to constraint 4, convolved holism: A homogenous perceptual object is always perceived as a whole with potential, arbitrary parts that it can, but need not be, split into.

The "homogeneity of simple content" or "ultrasmoothness" constraint thus means, simplified: A system that satisfies the "homogeneity of simple content" constraint needs to be able to represent areas or regions of phenomenal input data (also in a metaphorical sense, for other modalities than vision, and also in counterfactuals) as homogenous, atomic and without further subdivisions.

3.3.15 Constraint 11: "Adaptivity"

> If we want to understand how conscious experience, a phenomenal self, and a first-person perspective could be *acquired* in the course of millions of years of biological evolution, we must assume that our target phenomenon possesses a true teleofunctionalist description. Adaptivity – at least at first sight – is entirely a third-person, objective constraint.[88]

A common reaction when it comes to non-natural intelligences[89] is: "But *none* of these things is ever going to have genuine *emotions*!"[90]

[83] Citation from [Metzinger(2003)], page 189f.
[84] See [Metzinger(2003)], page 189, where Metzinger refers to [Sellars(1963)] and others.
[85] Citation from [Sellars(1963)], page 26, as cited in [Metzinger(2003)], page 190, emphasis as in [Metzinger(2003)].
[86] Citation from [Metzinger(2003)], page 192
[87] Citation from [Metzinger(2003)], page 195.
[88] Citation from [Metzinger(2003)], page 198, emphasis his.
[89] This can be artificial or postbiotic, or in fact any kind of intelligence that is not purely biological.
[90] Citation from [Metzinger(2003)], page 199, emphasis his.

The underlying issue is that artificial systems do not require any teleofunctionalist properties; their goals and ideals (or rather, the corresponding representative mental states) do not necessarily have to make sense. Only evolution ensures that the development of goal states which are detrimental to an organism's reproduction do not enter the common gene pool of that species. In other words, goal states in real intelligence do generally make sense.[91] Emotions, then, are the expression of those goal states in human beings. Being attracted to a member of the opposite sex (as a trivial example) is beneficial for reproduction.[92]

This, by the way, should in no way trivialize the what-is-it-likeness of emotions.[93] I merely mean to exemplify their functional role for the survival of a species, as normative value functions.

Similar arguments apply to consciousness and phenomenal states in general, although the benefits always have to be weighed up (and brutally and remorselessly *are* weighed up in evolutionary processes) against their downsides, like the larger brain that makes human childbirth more stressful (and potentially lethal for both mother and child) than that of other animals. If large brains thus are to prevail, they need to have other benefits that weigh up those shortcomings, or they will not stand up to evolutionary pressure. And of course, there is such a benefit: Consciousness increases the odds of survival of a particular species, because it increases flexibility and adaptivity.

> Consciousness, first, is an instrument to generate successful behavior; like the nervous system itself it is a device that evolved for motor control and sensorimotor integration. Different forms of phenomenal content are answers to different problems which organisms were confronted with in the course of their evolution. Color vision solves another class of problems than the conscious experience of one's own emotion, because it makes another kind of information available for flexible control of action.[94]

The adaptivity constraint is further discussed in chapters 3.4.5 (where I question the necessity of the constraint) and 3.4.6 (where I propose a weaker adaptivity constraint instead).

The "adaptivity" constraint thus means, simplified: A system that satisfies the "adaptivity" constraint needs to have a proper history, it needs to have grown and adapted to primary evolutionary pressure, and it needs to exhibit behaviour that is beneficial for the evolutionary niche it fits into.

3.3.16 Levels of consciousness

I duplicated the entire catalog of Metzinger's constraints because I believe they have merit as one of the first steps towards truly understanding consciousness, and, to that end, phenomenality. Metzinger does not attempt to diminish the difficulty of the hard problem, but admits what I think is important in this context:

We need a new interdisciplinary approach that includes neurophysiology, psychology, computer science, mathematics and of course philosophy, if we are to understand the complex ways in which a mere physical mass of neurons produce the emerging behaviour[95] of consciousness and phenom-

[91] This is true at least most of the time and if seen in a bigger context. Of course, the rapid secondary and tertiary evolution (through the replication of memes and temes) that we are now going through (see [Dawkins(1976)] and [Blackmore(2008)]) makes some of the goal states that were a result of genetic evolution occasionally outdated. Furthermore, things that are not detrimental (as opposed to necessarily beneficial) to a species' survival can also make it into the gene pool.

[92] I want to distance myself from naturalistic fallacies and homophobic readings of this example. Being attracted to the opposite sex being beneficial for reproduction does not make being attracted to your own sex any less probable or desirable from a moral or ethical point of view. Genetic dispositions to that end will however necessarily always remain in the minority, as individuals that exhibit them are less likely to reproduce.

[93] See chapter 4.1.

[94] Citation from [Metzinger(2003)], pages 200f.

[95] Lewes defines emergence in [Lewes(1875)], p. 412: "Every resultant is either a sum or a difference of the co-operant forces; their sum, when their directions are the same – their difference, when their directions are contrary. Further, every resultant is clearly traceable in its components, because these are homogeneous and commensurable. It is otherwise with emergents, when, instead of adding measurable motion to measurable motion, or things of one kind to other individuals of their kind, there is a co-operation of things of unlike kinds. The emergent is unlike its components insofar as these are incommensurable, and it cannot be reduced to their sum or their difference."

enality. And there is no need to be afraid of solutions that appear to be counter-intuitive, quite the contrary – our intuitions are partially formed by our current models of the world, models that can well be wrong, and thus it only makes sense to question them.

It is clear that while us human beings apparently do satisfy all those 11 constraints, it is not necessary to do so in order to be conscious. There are not just differentiations between "conscious" and "not conscious", but various levels in between. Metzinger determines which of these constraints have to be satisfied for a system to be called "conscious," and how far that specific kind of consciousness goes:[96]

- *Minimal Consciousness:* If a system is to be conscious at all, it at least has to satisfy the constraints of presentationality (2), globality (3) and transparency (7 – at least in a limited form, we do not require opaque content here yet). This is equivalent to "the presence of a world"[97] – which does not require subjectivity, differentiated representation of causality, space or time, and certainly no planning. Such a system would be "frozen in an eternal Now, and the world appearing to this organism would be devoid of all internal structure."[98]

- *Differentiated Consciousness:* This adds convolved holism (4) and dynamicity (5), enabling the system to perceive complex situations as such, and adding temporal structure.

- *Subjective Consciousness:* Adding perspectivalness (6) centers the space of experience on an active self-representation and thus adds a consciously experienced first-person perspective to the phenomenal space. The kind of consciousness that many animals enjoy seems to be of this kind.

- *Cognitive, Subjective Consciousness:* Here, we add offline activation (8), include both transparent and opaque content (7 again, but this time unlimited) and thus enable "an explicit phenomenal representation of past and future, of possible worlds, and possible selves"[99] – and decouple all these representations from current external input. Such systems could also represent themselves as representational systems, would be thinkers of thoughts, and would also be able to escape naive realism at least in thought experiments.

Note that this list does not include some constraints. Global availability (1) – although since this is a special case of the globality constraint, it is probably implied already in minimal consciousness. Representation of intensities (9) is arguably not always important, although it can be necessary if the system is to satisfy the adaptivity constraint.

Adaptivity (11) itself,[100] which is also not included, might be necessary for us to even recognize a system as being conscious at all. Nevertheless, this does not mean that we would necessarily have to include it as a prerequisite for consciousness as well. It is only our cognition that is affected by the problem, not necessarily the concept of consciousness itself: We can only identify a conscious system as such if it satisfies the adaptivity constraint.[101]

Homogeneity of simple content (10) is interesting, because it is, for systems like us humans, a prerequisite for proper concept formation and thus necessary at least for episodic memory and forward planning that cognitive, subjective consciousness introduces. It might even be necessary for convolved holism; how can a mind form a concept of wholes if it is not able to conceive of its contents as homogenous? I propose that this constraint should be added as a requirement for differentiated consciousness already.

[96] This list is very similar to the one in [Metzinger(2003)], pages 204ff.
[97] Citation from [Metzinger(2003)], page 204.
[98] Citation from [Metzinger(2003)], page 204.
[99] Citation from [Metzinger(2003)], page 205.
[100] Or at least the weaker adaptivity that I suggest in chapter 3.4.6.
[101] I guess that Metzinger would not agree with this conclusion, considering the effort he put into stressing the importance of the adaptivity constraint. He would probably see it as a prerequisite for any consciousness whatsoever. See chapters 3.3.15 and 3.4.5.

3.3.17 Phenomenality is ineffable

What really makes phenomenal experience special is its ineffability[102]:

> The insight of such fine-grained information evading perceptual memory and cognitive reference [...] allows us to do justice to the fact that a very large portion of phenomenal experience, as a matter of fact, is *ineffable*, in a straightforward and conceptually convincing manner. [...] The beauty of sensory experience is further revealed: there are things in life which can only be experienced *now* and by *you*. In its subtleness, its enormous wealth in highly specific, high-dimensional information [...], it is at the same time limited by being hidden from the interpersonal world of linguistic communication.[103]

Ineffability then comes naturally with the combination of at least perspectivalness and transparency. Other constraints (like global availability, activation within a window of presence, or convolved holism) make parts of phenomenal experience even more ineffable.

Phenomenality, just like consciousness, is not an all-or-nothing attribute of systems. It seems to me[104] that the extent to which a system's states (or rather, processes) are experienced as phenomenal by the system, then, is analogous to the extent of consciousness it exhibits, and thus analogous to the number and nature of multilevel constraints it satisfies.

3.4 Consolidation

Considering how much space I gave each of the three different points of view regarding phenomenality, the reader may have realized that I am quite partial to Metzinger's views. I think that he offers the first proper explanation of consciousness in general (despite Dennett in [Dennett(1991)] promising to do so 12 years earlier), and phenomenality in particular.

Metzinger is modest about it, stating that the list is preliminary and "deliberately formulated in a manner that allows it to be continuously enriched and updated by new empirical discoveries."[105] Nevertheless, his explanations seem sound for the time being, and I consider them to be a good start.

3.4.1 Metzinger's nomenclature: "Multilevel", "constraints"

At first sight, it seems odd that Metzinger chooses the name "multilevel constraints" for his definitions of what could be described as preconditions that are necessary for a system to exhibit consciousness. In order for this name to make sense, we have to look at it this way: An otherwise completely unconstrained data processing system of any kind that satisfies a sufficient number of these constraints[106] will exhibit consciousness. The things Metzinger describes are apparently not so much minimal preconditions, but really restrictions that can be applied to any system that is able to perceive, process and output data.

As we saw in chapter 3.3, the constraints are "multilevel" in that Metzinger attempts to capture all the important levels of description he considers to be important for research concerning the nature of consciousness for each of them. He distinguishes the following levels:[107]

- The *phenomenological* level of description. Insight into this level is gained introspectively and then analyzed epistemically, it describes how a system experiences the constraint phenomenologically.

[102] I will use the term "ineffability" in the following meaning: Something that is ineffable evades conceptualization, it is unspeakable not because it would be forbidden by some dogma to speak about it but because it is literally impossible to form a concept of an ineffable entity. Because it is not possible to form a concept of such things, it is also not possible to communicate or remember them. This seems to be the way Metzinger uses the term as well.

[103] Citation from [Metzinger(2003)], pages 94f.

[104] I am cautious here because while Metzinger does explicitly talk about levels of consciousness with the definitions (see chapter 3.3.16), he does not explicitly draw the analogy I am suggesting here.

[105] Citation from [Metzinger(2003)], page 117.

[106] See chapter 3.3.16 for the different levels of consciousness different combinations of constraints will produce.

[107] See [Metzinger(2003)], page 110; I paraphrase some of his descriptions.

- The *representationalist* level of description, describing the intentional content – the relationship between form and content for phenomenal representata. Insight into this level is gained introspectively as well.
- The *informational-computational* level of description, which concerns the computational function and teleological goal of the constraint, what function the constraint serves for the system as a whole. This and the further levels can only be analyzed by external means, not introspection by the system itself.
- The *functional* level of description, concerning purely causal properties independent of physical realization. An analysis of this level necessarily raises questions about the possibility of functional correlates of consciousness independent of realization.
- The *physical-neurobiological* level, investigating direct neural corellates of the constraint in usually human brains. This is a largely unexplored subject for many of Metzinger's constraints.[108]

Still, of course the multilevel constraints could also be called "minimal (multilevel) preconditions". This holds true for all constraints except constraint 11, the adaptivity constraint, which is special in more than one way anyway however. I explore this thought further in chapter 3.4.5.

3.4.2 What Husserl might answer to Dennett

in chapter 3.2, it was stated that Dennett's biggest gripe with traditional phenomenology is that it is not necessarily intersubjective. He makes this quite explicit:

> Lone-wolf autophenomenology, in which the subject and experimenter are one and the same person, is a foul, not because you can't do it, but because it isn't science until you turn your self-administered pilot studies into heterophenomenological experiments.[109]

However, I am not sure if Dennett's position is really that different from Husserl's with respect to what is science and what is not. Maybe Heidegger's continuation of Husserl's theories, putting significantly more focus on what he considers to be a fact, that phenomenality is a subjective phenomenon, would fall victim to the above argumentation of Dennett's. Husserl, however, may even agree with Dennett:

> It is absolutely clear; scientific conclusions regarding phenomena are not feasible after phenomenologic reduction, in particular if we conceptually define those phenomena as absolutely distinct and unique. Only if we enter the empirical-psychological sphere, if we look at the phenomena as experiences of an experiencing self that is in an interrelation with nature, can we conceptualize [what scientific conclusions we can reach] in a similar way to how psychologists do it in experimental practice.[110]

Thus, individual phenomenal experiences only gain a scientific meaning if they are also looked at in the context that the subject experiencing them was in. Furthermore, it is important to recognize that empirical research includes not only individual, distinct and unique such phenomenal experiences, but rather an entire class of them, preferrably from multiple subjects or at least from the same subject reproducibly experienced.

Husserl also does not presuppose infallibility of phenomenal experience:

[108] Also see chapter 8.2, where I explain why the theory presented in this thesis essentially amounts to a conceptual dualism.

[109] Citation from the online version of [Dennett(2003)].

[110] Citation from [Husserl(1984)], page 224, my translation. Original text: "Es ist danach völlig klar, wissenschaftliche Feststellungen in bezug auf die Phänomene sind nach der phänomenologischen Reduktion nicht zu machen, *notabene* wenn wir diese Phänomene als absolute Einzelheiten und Einmaligkeiten fixieren und begrifflich bestimmen wollen. Nur wenn wir in die empirisch psychologische Sphäre gehen, wenn wir die Phänomene als Erlebnisse eines erlebenden Ich, das im Zusammenhang einer Natur steht, betrachten, können wir eine solche Fixierung vollziehen in der Art, wie jeder Psychologe es im experimentellen Verfahren tut."

> What the referenced object actually is is not topic of our research; it may exist or not, and we may doubt the meaning of its existence, evidentially it is part of the essence of experience that it experiences something, an object, and I can now ask what [the experience] takes the object to be.[111]

In light of these passages from Husserl, I am uncertain as to what precisely it is that so radically distinguishes heterophenomenology from phenomenology. Both Dennett and Husserl agree that we have some (more or less limited) privileged access to our mental states. They further agree that proper scientific research can only be performed from a third-person, intersubjective point of view. Only if we stop taking phenomenological accounts at face value, and start forming theories that attempt to explain what it is that makes agents enjoy the particular phenomenal experiences that they do.

Considering that neither of them actually offers more insight into what phenomenal experience itself is however, I lay this issue aside, and for the remainder of this thesis focus on Metzinger's more thorough theory of the genesis of phenomenality. After all, he actually did what both Dennett and Husserl proposed: He tried to model how agents manage to make autophenomenological reports[112] in the first place.

3.4.3 Another look at "global availability" and "convolved holism"

When Metzinger talks about the constraints "global availability" and "convolved holism,"[113] he for the most part focusses on the phenomenal side of them. Minsky also suggests that things certainly *seem to be* globally available. Minsky quotes Newman:

> [In the Global Workspace theory] The theater becomes a workspace to which the entire audience of "experts" has potential access, both to "look at" other inputs and contribute their own. [...] Individual modules can pay as much or as little attention as suits them, based upon their particular expertise and proclivities. At any one moment, some may be dozing in their seats, others busy on stage [...] [but] each can potentially contribute to the direction the play takes. In this sense the global workspace resembles more a deliberative body than an audience.[114]

Minsky however goes on to criticize just the complete global availability that Metzinger suggests:

> However, this raises several questions about the extent to which different resources can speak the same language, and [I] argue that different resources will need to use multiple levels of representations and different short-term memory systems to keep track of various kinds of contexts. Besides, if each specialist could broadcast signals to all the rest, the workspace would become so noisy that the system would need to develop ways to restrict the amount of communication.[115]

Global availability should also not be overrated:

> How could such machines [such as individually unique human brains] work reliably, in spite of so much variety [in terms of environmental and self-modeling]? To explain this, quite a few thinkers have argued that our brains must be based on "holistic" principles, according to which every fragment of process or knowledge is "globally distributed" (in some unknown way) so that the system's behavior would still be the same in spite of the loss of some of its parts.

[111] Citation from [Husserl(1984)], page 231, my translation. Original text: "Allerdings das Sein des Gegenstands lassen wir dahingestellt; aber mag er sein oder nicht sein, und mögen wir zunächst über den Sinn dieses Seins noch so sehr im Zweifel sein, evident ist es zum Wesen der Wahrnehmung gehörig, daß sie etwas wahrnimmt, einen Gegenstand, und ich kann nun fragen, als was nimmt sie den Gegenstand für wahr."

[112] We looked at Dennett's introduction of (auto)phenomenological reports and their role as mere basis for proper scientific research in chapter 3.2.

[113] See chapters 3.3.5 and 3.3.8 respectively.

[114] Citation from [Newman(1995)] as found in [Minsky(2006)], pages 125f.

[115] Citation from [Minsky(2006)], page 126.

> However, [I] suggest that we do not need any such magical tricks – because we have so many different ways to accomplish any type of job. Also, it makes sense to suppose that many parts of our brains evolved as ways to correct (or to suppress) the effects of defects in other parts.[116]

So, while it certainly may *seem* as if we had everything globally available, what in fact happens is that sensory inputs are processed in parallel by a wide variety of processes, and many brain centers (which Minsky calls "critics" and "selectors") have access to parts of these processes and can potentially process them further.

However, this kind of "global availability" is neither magical nor entirely global. Rather, which brain centers have access to what processes is stored in the dynamical hardware[117] of the brain. I do not mean to claim that domain borders between different processes cannot be crossed – quite the opposite, that possibility is indeed (according to Metzinger) one of the prime advantages of our kind of consciousness. Yet, these domain borders can only be crossed in certain ways that are predetermined in the same dynamical hardware that implements the processes themselves.

This does not stop us from phenomenally experiencing the byproducts and intermediate and end results of our brain's information processing as utterly and totally holistic and global.

3.4.4 The trouble with "perspectivalness"

Metzinger's "perspectivalness" constraint[118] seems to be quite straightforward. The constraint describes a first-person point of view which starts out with a proto self-model which is nonconceptual, happens before cognitive processing, and is automatic and requires no effort whatsoever. The complete constraint then consists of a nonconceptual prereflexive proto self-consciousness, a PSM[119] and a PMIR.[120] Given a prereflexive proto self-model, we can conclude that a complete self-model has to arise from the system.

However, such a proto self-consciousness is anything but trivial, and not that straightforward after all. What exactly is the entity that we are prereflexively, nonconceptually, aware of? It cannot be a self-model, as that would mean that it would require itself in an infinite regress. There is some basic form of perspectivalness which is given for all of today's known conscious systems; They are all spatially confined to a relatively small area, and perceive their entire surroundings via a single viewpoint. Is there more to perspectivalness than that? Can that be enough to give rise to a complete PSM later on?

Finding this prereflexive proto self-model will require empirical experiments with postbiotic or artificial systems to truly advance in this matter, as today very little about it is known. I analyze this further in chapter 5.5.3, after discussing Metzinger's PSM.

3.4.5 "Adaptivity" questioned

Metzinger heeds another warning regarding his "adaptivity" constraint[121]: He says that in order to be truly conscious, a postbiotic or artificial system also has to fully satisfy this constraint. He goes even further and quotes Davidson's "The Swampman."[122] In the thought experiment, Davidson is standing near a tree in a swamp during a thunderstorm. When lightning strikes the tree, Davidson is disintegrated, and by pure coincidence the tree is turned into an exact physical replica of Davidson (including brain with current neural state etc.).[123]

[116] Citation from [Minsky(2006)], page 345.

[117] I use "dynamical hardware" here due to a distinct lack of a better term when it comes to process descriptions implemented in neurons – that are both rigid in the short term and dynamic in the long term.

[118] See chapter 3.3.10.

[119] Phenomenal self-model, see chapter 5.3.

[120] Phenomenal model of the intentionality relation, see chapter 6.3.

[121] See chapter 3.3.15.

[122] See [Metzinger(2003)], page 206.

[123] This does bear some similarities to the paradox of the ship of Theseus, but also striking differences. Two things are very different: The timespan over which the transformation happens, and the location in which it happens. The ship of Theseus is transformed over a longer period of time, while the swampman is transformed instantaneously

The aim of this thought experiment is showing that an exact replica of a human being might move, think, talk, and argue just like the original, it might even have the exact same kind of phenomenality. Yet it lacks the original Davidson's intentional content. Metzinger writes:

> [...] for instance, it has many false memories about its own history be they as conscious as they may. The active phenomenal representations in Swampman's brain would [...] *not* satisfy the adaptivity constraint, because these states would have the wrong kind of history. They did not originate from a process of millions of years of evolution. [...] It would [...] still be consciousness in a weaker sense, because it does not satisfy the adaptivity constraint [...].[124]

It is interesting how this is argument runs counter to Dennett's dismissal of original intentionality, which is explored in chapter 6.2.1, but also seems to contradict other parts of Metzinger's work. I do not see how the requirement for some kind of original intentionality here is compatible with Metzinger claiming that all that intentionality really is, is a phenomenal model .[125]

Due to these inconsistencies, I do not see the argument for an "adaptivity" constraint as having much merit.

Furthermore, if it was correct,[126] every subsequent generation of humans would be less conscious than the former: After all, much of what we act and live by is not genetically hardcoded, but rather learned and perceived; it also has "the wrong kind of history" built-in, only attached by means of being passed down through the generations. All the things we take as signs of proper intelligence beyond basic empathy (including being able to play chess, talking with each other in a common language, reading and writing) do not have this kind of proper, system-internal history. They were introduced from the outside, from our parents and teachers, our environment, our experiences. And that is the very thing Metzinger tells us is what is wrong with a hypothetical postbiotic intelligence – even if it were built from actual organic tissue, and even if it had gone through an evolutionary process of its own:

> [...] it would still follow that these postbiotic phenomenal systems would only be conscious in a slightly weaker sense than human beings, because human beings were necessary to trigger the second-level evolutionary process from which these beings were able to develop their own phenomenal dynamics. From a human perspective, just like Swampman, they might not possess the right kind of history to count as maximally conscious agents.[127]

Obviously, humans are necessary to trigger the second-level evolutionary process that eventually leads to childbirth as well,[128] just like the further second-level evolutionary processes that are in place when we teach and learn in schools and universities.

3.4.6 A weaker adaptivity constraint

When Metzinger formulated his adaptivity constraint, he seems to have been driven by mainly one of his background assumptions, one that he explicitly states at the very beginning of his book: Teleofunctionalism. He introduces his commitment like this:

---

and spontaneously. The location of the new ship of Theseus is the same as the old one was, while the swampman stands in a different spot than Davidson did. However, there are striking resemblances as well, because in both thought experiments, an exact physical replication is created while the original is destroyed, and the question is what traits that replication shares with the original. It all comes down to what transtemporal aspects the traits in question have of course, and that is rarely well-defined.

[124] Citation from [Metzinger(2003)], page 206.

[125] The phenomenal model in question is the PMIR (or phenomenal model of the intentionality relation), which will be explained in chapter 6.3.

[126] I owe this point to my good friend Simon Bünzli.

[127] Citation from [Metzinger(2003)], page 207.

[128] I like the take of the movie "Robots" from 2005 on this. Herb Copperbottom: "Making the babies is the best part!"

> [...] representata have been specified by an additional *teleological* criterion: an internal state X represents a part of the world Y *for* a system S. This means that the respective physical state within the system only possesses its representational content in the context of the history, the goals, and the behavioural possibilities of this particular system. This context, for instance, can be of social or evolutionary nature. [...] It is for this reason that we can always look at mental states with representational content as instruments or weapons. If one analyzes active mental representata as internal tools, which are currently used by certain systems in order to achieve certain goals, then one has become a teleofunctionalist or a teleorepresentationalist.[129]

Teleology may not be the best approach for everything (and in fact, when Sehon used it in [Sehon(2005)] along with the term "supervenience" Metzinger also uses,[130] he constructed a theory that lacks scientific foundation and appears to be of mostly speculative nature, as all theories involving metaphysical and non-observable entities necessarily do).

But when Metzinger uses it, it translates to the assumption that the things our brains do[131] must be good for something – because if they were not, they would not have been a fitness criterion for evolution, and consequently would not have survived the last few million years.

This is consistent with scientific findings.[132] I do not agree with Metzinger's assumption that million years of evolution are a necessary prerequisite for something being a fitness criterion however, nor do I agree with the assumption that merely because a fit system has been duplicated (as Davidson's Swampman was, see chapter 3.4.4) it suddenly loses some kind of metaphysical fitness *pre*condition. However, this is exactly what Metzinger's reading of the adaptivity constraint presupposes: That there is some secret ingredient to proper intentionality after all, a ghost in the machine reminiscent of dualist theories despite all our efforts to banish it.[133]

I do not deny the importance of evolution for today's kind of intelligence. It certainly seems to have been essential, both as a driving force and a shaping constraint. However, even from a purely teleofunctionalist point of view, 'proper history' cannot be a criterion for the degree of intentionality a system possesses.

'Fitness for purpose,' however, can be, even if we are talking about intuitively non-replicable things like emotions (which, essentially, are nothing but 'logic of survival' indicators) or qualia – and as such, I would like to propose a weaker form of Metzinger's adaptivity constraint. While an organism may pursue the goals of its ancestors, while this may be important for all of today's biological sentient systems (including humans), it is not necessary for other, as of yet unknown, kinds of proper consciousness.

My suggestion is to take Metzinger's normative approach to what good representata[134] are and slightly alter it. His normative definition of good representata is:

> Phenomenal representata are *good* representata if, and only if they successfully and reliably depict those causal properties of the interaction domain of an organism that were important for reproductive success.[135]

The change I propose is simple:

[129] Citation from [Metzinger(2003)], page 26, emphasis his.

[130] See chapter 3.3.4.

[131] As a matter of fact, he is not only talking about our brains, but all organic entities that were subject to a genesis in evolution.

[132] Even if we have to ignore for now that teleological explanations are not necessary for every bit of functionality, as some might only have been not detrimental to survival, as opposed to being strictly beneficial. A well-known example is the appendix – it does have a function and must once have been necessary for survival, but no longer is, but nevertheless it also is not detrimental to our survival.

[133] This is particularly odd because Metzinger is quite good at finding "subtle residues of Cartesian dualism" elsewhere, see chapter 3.3.11.

[134] Of course, the very notion of representata being "good" implies them being good *for something*, unless we want Platonian ideas entering the discussion. The argument is prone to becoming circular if "good" is defined as 'beneficial for survival.' Metzinger does not explore this further, but 'fitness for purpose' seems to be a fitting description of his intentions at this point.

[135] Citation from [Metzinger(2003)], page 203, emphasis his.

> Phenomenal representata are *good* representata if, and only if they successfully and reliably depict those causal properties of the interaction domain of an organism that *are* important for *successful survival of the genotype.*

The two changes thus are the following:

- "Are" instead of "were", because while the history of a system's genesis might be interesting for evolutionary or historiographical purposes, it is not required for adaptivity.
- "Survival" instead of "reproduction" because a postbiotic or artificial system does not necessarily need to be able to reproduce, particularly if it is not plagued by death of individuals and phenotypes like we humans are. Of course, when applied to biologic organisms, "survival of the genotype" presupposes reproduction for the individual organism.

3.4.7 Rational-causalist phenomenology reconsidered

In [Gloor Modjib(2008)], I suggested going beyond Husserl's phenomenology and altering two fundamental phenomenological premises in order to make phenomenology compatible with a non-dualist world view. These premises were:

- "Mental objects are immaterial": as soon as we give up the notion that mental objects are per se not understandable, we can start to analyze them; Which is exactly what I attempt to do in this thesis.
- "Mental objects can be directed at objects in the world": if we assume that mental objects are merely directed at mental models of things in the world,[136] we can causally analyze their interdependence and form models of this directedness, because the entities that are now directed at each other are of the same kind.

I do not think I even deviated that much from Husserl's ideas with those suggestions.[137] Husserl realized that there is a cognitive closure, that we cannot perceive transparent properties of mental events (using Metzinger's notion of "transparency", as per constraint 7). Husserl merely did not attempt to look behind what is introspectively accessible, because he thought there would be no point in attempting to explain something when we do not even know what the explanandum itself is. Neither did he attempt to form an ontological theory, but rather an epistemologically plausible one.

I believe that neurological and psychological research has brought us to a point where that explanandum becomes – at least a little bit – less obscure.

[136] These models are in turn created by things in the world that 'enter' the 'mental space' via physically analyzable pathways, for example light hitting light-detecting nerve cells in the eye.

[137] Thanks again to Prof. E. Marbach for insightful discussions regarding this subject.

4 Qualia

Qualia are generally said to be the specific "what-is-it-like" structure of experience. Opinions on what exactly this structure is, and how to best research it, differ widely. What is common to most notions of qualia is that they are supposed to be among the most basic available bits of epistemic knowledge concerning conscousness, and that they seem to be indivisible from a first-person point of view.

Examples are our experience of distinct shades of red and green in a picture we look at, or of the specific character of the sound of a piano, or the particular kind of taste of a glass of good wine, or the enjoyment we derive from a particular manner in which sunlight is cast through treetops in a forest. Dennett makes an example of the plethora of qualia we can experience and enjoy. He later deconstructs this example, but I find it to be a rather good description of why we find qualia so interesting:

> Green-golden sunlight was streaming in the window that early spring day, and the thousands of branches and twigs of the maple tree in the yard were still clearly visible through a mist of green buds, forming an elegant pattern of wonderful intricacy. The windowpane is made of old glass, and has a scarcely detectable wrinkle line in it, and as I rocked back and forth [in my rocking chair], this imperfection in the glass caused a wave of synchronized wiggles to march back and forth across the delta of branches, a regular motion superimposed with remarkable vividness on the more chaotic shimmer of the twigs and branches in the breeze.
>
> [...] The enjoyment I felt in the combination of sunny light, sunny Vivaldi violins, rippling branches – plus the pleasure I took in just thinking about it all – how could *all that* be just something physical happening in my brain?[138]

What is precisely the crux of the matter – how could it? Wait and see.

4.1 Lewis: Introducing qualia

Clarence Irving Lewis introduced qualia as maximally simple forms of sensory content like this:

> In any presentation, this content is either a specific quale (such as the immediacy of redness or loudness) or something analyzable into a complex of such. The presentation as an event is, of course, unique, but the qualia which make it up are not. They are recognizable from one to another experience. [...]
>
> What any concept denotes – or any adjective such as "red" or "round" – is something more complex than an identifiable sense-quale. In particular, the object of the concept must always have a time-span which extends beyond the specious present; this is essential to the cognitive significance of concepts. The qualia of sense as something given do not, in the nature of the case, have such temporal spread. Moreover, such qualia, though repeatable in experience and intrinsically recognizable, have no names. They are fundamentally different from the "universals" of logic and of traditional problems concerning these.[139]

Qualia for Lewis are ineffable[140] and predate concepts and knowledge. He points out that "qualia are subjective; they have no names in ordinary discourse but are indicated by some circumlocution such as 'looks like.'"[141]

Qualia are recognizable and thus also categorizable – but they are not determinate concepts just yet. There is however a correlation between qualia as categorizable perceptual content and determinate concepts. That correlation is both potentially different from one situation to the next

[138] Citation from [Dennett(1991)], pages 26 and 406, emphasis his.
[139] Citation from [Lewis(1929)], pages 60f.
[140] See [Lewis(1929)], pages 123f.
[141] Citation from [Lewis(1929)], page 124.

(when different aspects of a concept are in focus) and different from one person to another, but qualia are reliably graspable in such determinate concepts.

Epistemically, it is not possible to be mistaken about qualia; we always know which qualia we experience. However, it is possible that the concepts to which the experienced qualia relate are wrong, and (while part of our world model), not part of the external world. As Metzinger points out,[142] Lewis qualia are available for attentional, behavioural and cognitive processing.

Lewis work was hugely influental. Thomas Nagel later coined the term of "what-it-is-like-ness," which further exemplified what is special about simple sensory content. For instance, taking his most prominent example, we face serious problems when we attempt to imagine what it is like to be a bat:

> In so far as I can imagine this (which is not very far), it tells me only what it would be like for me to behave as a bat behaves. But that is not the question. I want to know what it is like for a bat to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications.[143]

It is worth noting that Nagel was not anti-physicalist, nor anti-reductionist, even though this quote seems to imply this.[144]

We will confront Lewis' notion of qualia with Ruffman's and Metzinger's in chapter 4.3.3. Nagel's observations seem to hold true for all of them. We cannot imagine what it is like to be a bat – all we can imagine is what it is like to be a (particular) human being who pretends to be a bat.

4.2 Dennett: No qualia whatsoever

Dennett adopts a quite radical stance regarding qualia: For him, they (at least in their traditional reading) do not exist.[145] There is nothing irreducibly, inexplicably special about supposedly atomic phenomenal qualities of subjective experience. He proposes a fresh start, one without the history of the discussion about qualia and their special nature:

> Even though philosophers have discovered the paradoxes inherent in this closed circle of ideas [regarding qualia] – that's why the literature on qualia exists – they haven't had a *whole alternative vision* to leap to, and so, trusting their still-strong intuitions, they get dragged back into the paradoxical prison. That's why the literature on qualia gets more and more convoluted, instead of resolving itself in agreement. But now we've put in place such an alternative vision, the Multiple Drafts model.[146]

Let us examine this Multiple Drafts model in detail, and establish why Dennett believes that it can replace qualia.

---

[142] See [Metzinger(2003)], page 84.

[143] Citation from [Nagel(1974)], page 3.

[144] Nagel merely pointed out that "at the present time the status of physicalism is similar to that which the hypothesis that matter is energy would have had if uttered by a pre-Socratic philosopher." (Citation from [Nagel(1974)], page 6.) Consequently, physicalism appears magic to us, because we cannot imagine how moving away from the specificially first-person way of experiencing things can get us any closer to how that specificially first-person way of experiencing things happens. This must be explainable on a functional level, as well as by neurophysiology (which, as Nagel himself admitted, faces no such obstacles). Nevertheless, we are faced with some obstacles right now, because in our current position, we have to rely on empathy for emulation and proper appreciation of other minds and have no way of easily and objectively formalizing them.

[145] See [Dennett(1988)].

[146] Citation from [Dennett(1991)], page 370.

4.2.1 Orwellian vs. Stalinesque revisions

It is a fact that sometimes, we misrepresent past happenings. This is not only true for things which happened days or months ago, but sometimes we even seem to have experienced them differently to how they actually and observably happened, right after the fact, or even seemingly in the act of perceiving them. One of the many examples Dennett puts forward is how subjects seem to experience taps within 1 to 3 seconds, in only three distinct places (wrist, elbow and upper arm) on one of their arms, as if a small animal was steadily crawling up that arm:

> The astonishing effect is that the taps seem to the subjects to travel in regular sequence over equidistant points up the arm – as if a little animal were hopping along the arm. Now, at first one feels like asking *how did the brain know* that after the five taps on the wrist, there were going to be some taps near the elbow? The subjects experience the "departure" of the taps from the wrist beginning with the second tap, yet in catch trials in which the later elbow taps are never delivered, subjects feel all five wrist taps at the wrist in the expected manner.[147]

Dennett tries to find out how this can happen. He suggests two seemingly mutually exclusive alternatives:

- *Orwellian revisions:* This explanation has parallels to the Ministry of Truth in Orwell's novel 1984[148] in that it rewrites history all the time: This theory suggests that processes in the brain rewrite memories after the experience itself has happened, altering not the experience itself, but our memories of it. So, if we have wrong memories of happenings in the past, we had correct memories once, but they changed over time.
- *Stalinesque revisions:* Like the Stalin government misrepresented facts through the censored media to its populace, complete with false evidence and elaborate productions, this explanation changes things as they happen: We experience worldly contents that have been altered by lower-level brain processes already. So, if we have wrong memories of happenings in the past, it is not because the memories are misrepresentations of past experience, but because that past experience already was factually wrong.

For both these alternatives, "in the past" can well be only seconds ago – after all, short-term memory may be a very special kind of memory with a very special structure, but it consists of mental content[149] just like long-term memory.

4.2.2 Multiple drafts

The fact that both Orwellian revisions seem to be clearly correct in some situations,[150] while Stalinesque revisions seem to be obviously correct in others,[151] it appears that the answer must be a compromise that incorporates aspects of both.

---

[147] Citation from [Dennett(1991)], page 143.

[148] See [Orwell(1949)].

[149] Mental content appears to be formed by ever-changing physical representations where the current form of the vehicle is actually part of and forms the content. Metzinger agrees with Dennett on this. For example, in [Metzinger(2003)], page 114, when Metzinger is talking about how an intentionality relation is internally modeled:

> Intended cognition now means that a system actively – corresponding to its own needs and epistemic goals – changes the physical basis on which the representational content of its current mental state supervenes.

This alteration of the physical basis goes even further, because the brain actively remodels neuron connections in processes like learning or forming memories, and those very connections are what forms the mental content in the first place.

[150] For example, if Stalinesque revisions were correct, there would need to be a delay from the onset of a bit of experience to when we perceive it, so experiments with reaction time all clearly contradict the Orwellian theory.

[151] In a psychological experiment where a red and a green dot next to each other were alternately flashing, subjects experienced a dot that moved back and forth, changing color inbetween. Whenever they were probed, there was no experience of flashing dots, only moving ones, at least if they were flashing with a certain minimum speed.

Furthermore, there is no Cartesian theater,[152] Dennett merely finds a "spatiatemporal smearing of the observer's point of view in the brain."[153] This suggests that there is no such point that would allow us to decide whether a revision has been made *before* or *after* some perception has become conscious.

Dennett's model of multiple drafts builds on those two observations and finds a quite elegant solution to the problems they pose to dualist or epiphenomenal views:

> According to the Multiple Drafts model, all varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous "editorial revision."[154]

Dennett continues to explore some consequences that seem oddly inconsistent at first sight, but are perfectly rational from a multiple drafts point of view. Looking at these processes as a continuous stream from sensory input to eventual dissolving (or storage in long-term memory), he points out:

> Probing this stream at different places and times produces different effects, precipates different narratives from the subject. If one delays the probe too long [...], the result is apt to be no narrative left at all – or else a narrative that has been digested or "rationally reconstructed" until it has no integrity. If one probes "too early," one might gather data on how early a particular discrimination is achieved by the brain, but at the cost of diverting what would otherwise have been the normal progression of the multiple stream. Most important, the Multiple Drafts model avoids the tempting mistake of supposing that there must be a single narrative [...] that is canonical – that is the *actual* stream of consciousness of the subject.[155]

The vantage of this theory is that it has empirical legitimation,[156] and that timings for example for processing of visual inputs can be found with experiments. Several different kinds of judgements that a subject makes can be distinguished, among them "the mere onset of stimulus, then location, then shape, later color (in a different pathway), later still (apparent) motion, and eventually object recognition."[157] And all those judgements do not necessarily come together somewhere in one locus for a final review (although they can, in memory, where they can then be overwritten or incorporated in other contents in a continuous editorial process). Rather, they are rather directly available for further processing and, potentially, actions.[158]

Dennett goes further and even offers some additional suggestions:

- There is no "optimal time of probing,"[159] since while there can be a version that is stored in long-term memory, earlier probings (when more parallel processes are still running) usually

[152] The Cartesian theater is Dennett's proposal for the traditional dualist answer to the following problem as formulated in [Dennett(1991)], page 107:

> It seems that if we could say *exactly* where, we could say exactly when the experience happened. And vice versa: If we could say exactly when it happened, we could say where in the brain conscious experience was located.

According to Dennett's interpretation of Descartes (see [Descartes(1642)]), there is a locus of understanding in traditional dualist theories, like a theater where everything is presented to an internal audience, the place where all the data collected by other mechanisms in the brain comes together and *makes sense*. This would make the understanding of experience a lot easier.

However, the idea of the Cartesian theater has multiple problems concerning interpretation of things that are presented within the theater, it would require substance dualism (see chapter 5.5.1), and it would fail to account for the different experimental findings concerning when change in memories happens (see chapter 4.2.1).

[153] Citation from [Dennett(1991)], page 126.

[154] Citation from [Dennett(1991)], page 111.

[155] Citation from [Dennett(1991)], page 113.

[156] Dennett makes it his mission throughout [Dennett(1991)] in particular to list empirical evidence incompatible with less flexible and more traditional theories, all of which are compatible with Dennett's view.

[157] Citation from [Dennett(1991)], page 134.

[158] Metzinger would distinguish between different kinds of availability here, as we saw in chapter 3.3.5. Dennett's focus lies elsewhere of course, but it is important to see where their two theories can benefit from each other – this seems to be one such place.

[159] See [Dennett(1991)], page 136.

yield more information about the specific nature of the running processes, and some mental content is at that time phenomenally perceived as being in the center of attention.

- Subjective time is similar to a narrative, in which earlier events can be incorporated after later events ("Bill arrived at the party after Sally, but Jane came earlier than both of them"[160]); there is thus the possibility of backwards projection of perceived events in time.[161]

Being able to spin a narrative is what "a something it is like something to be"[162] is all about for Dennett. A section of the world becomes an observer as soon as it starts composing a narrative, or rather, a "skein of contents [which] is only rather like a narrative because of its multiplicity; at any point in time there are multiple drafts of narrative fragments at various stages of editing in various places in the brain."[163] And that, according to Dennett, is exactly what our brain does, and thus it is able to phenomenally experience its surroundings in what it believes to be qualia.

4.2.3 Disqualifying qualia

Dennett starts with the observations that led him to his multiple drafts theory, and adds some further layers, before he attempts to disqualify the concept of qualia.

An important point of Dennett's theory regarding qualia is that feature detectors which make us experience something as "blue" or "bitter" co-evolved with the features they detect. Color vision evolved because there were features distinguishable by color. More precisely, it evolved because color happens to be a concept that is able to interpret random side features of chemical compositions of certain aspects of our ancient predecessor's surroundings, and able to do that in a way that is beneficial for an individual's survival – in that it makes the individual able to draw distinctions that happen to correlate with (what is now known as) color differences.

So color vision evolved because there were features distinguishable by color, and vice versa, starting small and growing from there - yet the fact that a butterfly's wings have the same blue hue as cobalt is mere coincidence. "Why is the sky blue? Because apples are red and grapes are purple, not the other way around."[164]

From this follows that a particular quality is always bound to detectors for that particular quality. As Dennett writes:

> The only *readily available* way of saying just what shape property M is is just to point to the M-detector and say that M is the shape property property detected by this thing here. [...] What property does Otto judge something to have when he judges it to be pink? The property he calls pink. And what property is that? It's hard to say, but this should not embarass us, because we can say why it's hard to say. The best we can do, practically, when asked what surface properties we detect with color vision, is to say, uninformatively, that we detect the properties we detect.[165]

This does not explain the particular way things feel for us. The common term describing what

[160] Citation from [Dennett(1991)], page 137.

[161] A particularly striking example and one of my favourites among the experiments Dennett brings forth to support his theory is one by Libet, which Dennett introduces in [Dennett(1991)], pages 154ff (of which experiment alledgedly a reproduction was never attempted) – see [Libet(1981)], where Libet also argues against opposition he faced from Churchland.

In this experiment performed during a neural operation with awake subject and open brain, Libet stimulated the left and right hand of a patient, and also stimulated the observedly corresponding neural areas (that sit on the opposite side of the brain - left for the right hand and vice versa) directly. As Dennett writes in [Dennett(1991)], page 155: "Most strikingly, Libet reported instances in which a patient's left *cortex* was stimulated *before* his left *hand* was stimulated, which one would tend to think would surely give rise to two felt tingles: first right hand (cortically induced) and then left hand. In fact, however, the subjective report was reversed: 'first left, then right.'"

[162] See [Dennett(1991)], page 137.

[163] Citation from [Dennett(1991)], page 135.

[164] Citation from [Dennett(1991)], page 378.

[165] Citation from [Dennett(1991)], pages 382f.

makes this feeling special seems to be the expression that we 'enjoy' something[166] – this nomenclature stresses the difference between neuroanatomy and experience, between information and qualia. Qualia (which Dennett does not support) cannot be divided into mechanical parts, while mere information can be split up without losing anything. Dennett believes that Qualia do not exist insofar as that there are no such irreducible basic entities of experience.

To prove that these differences between the individual experiences of different people are explicable in their entirety, Dennett proposes another thought experiment. He has people want to experience Bach as if they were his contemporaries, without being tainted by more recent musical developments or other advances in the meantime. He claims that:

> If we want, we can carefully list the differences between our dispositions and knowledge and [the people who did actually hear Bach originally], and by comparing the lists, come to appreciate, in whatever detail we want, the differences between what it was like to be them listening to Bach, and what it is like to be us.[167]

Noteworthy is furthermore that inverted qualia do not seem to exist: People who wear goggles that turn the world upside down (always, for extended periods of time) adapt to that after a while, and are able to ski downhill or drive bicycles through city traffic.[168] Asked whether they have adapted to their experiential world being upside down, or whether they are mentally turning it right side up again, subjects answer that this is not the right question.

As Dennett points out, this is perfectly in line with the multiple drafts model in that multiple of the parallel processing algorithms happening in the brain will form different views on that question. Some of these algorithms (like shape or color detection mechanisms) are completely independent of the orientation of the environment, while others (like ducking the right way when something is thrown at you) will probably take much longer to adapt than regular orientation and navigation.[169] Parts of the brain will necessarily disagree with each other[170] whether it's adaptation or mental content modification.

Essentially, Dennett's position amounts to the following conclusion: It is, at least in theory, possible to describe any experience in its entirety, and description is all there is to the experience. As such, qualia are just one (fully reducible) abstraction of that description, and are superfluous in that they are not an entity that is a necessary part of our kit for investigating consciousness.

4.3 Metzinger: Phenomenal presentational content

Metzinger, like Dennett, does not believe that the classical concept of qualia withstands scientific scrutiny. He says that "qualia, in terms of an analytically strict definition – as the simplest form of conscious experience in the sense of first-order phenomenal experiences – do not exist."[171]

For Metzinger, the traditional notion of qualia is that they are are maximally simple (atomic) subjective universals, recognizable from one experiential episode to the next via subjective identity criteria, and forming the intrinsic core of all subjective states. Judgements about qualia cannot be false, because they are defined through the subjective determination that judges them.[172]

These background assumptions provide the basis for my detailed exploration of his argumentation.

---

[166] For example, Dennett uses this nomenclature in [Dennett(1987)], page 290, to illustrate what defenders of original (as opposed to derived) intentionality as a concept might say: "Any computer program, any robot we might design and build, no matter how strong the illusion we may create that it has become an genuine agent, could never be a truly autonomous thinker with the same sort of original intentionality we enjoy."

[167] Citation from [Dennett(1991)], page 388. Note the wording, I will argue in chapter 4.4.1 that appreciation is more than merely description.

[168] See [Dennett(1991)], page 393.

[169] See [Dennett(1991)], page 397. Dennett writes ibd.: "The more complete [their adaptation] was, the more the subjects dismiss the question as improper or unanswerable."

[170] This is merely a figure of speech, what actually happens is that conflicting theories are formed in different parts of the brain that implement different algorithms, and those conflicting theories are at least partially accessible to introspection.

[171] Citation from [Metzinger(2003)], page 64.

[172] See [Metzinger(2003)], pages 66ff.

4.3.1 Qualia are inefficient

From a computational point of view, it makes sense that not all phenomenal experiences have such qualia (in terms of maximally simple mental content) that they automatically produce:

> It would be uneconomical to take over the enormous wealth of direct sensory input into mental storage media beyond short-term memory: A reduction of sensory data flow obviously was a necessary precondition (for systems operating with limited internal resources) for the development of genuinely cognitive achievements. [...] Computational load has to be minimized as much as possible. Therefore, online control has to be confined to those situations in which it is strictly indispensable.[173]

This is of course a teleological observation which does not necessarily mean that qualia do not exist themselves. However, this is precisely what Metzinger claims, as we will see.

4.3.2 Most simple forms of content do not exist

Metzinger wants to show that there is no such thing as 'always implied most simple forms of content' when we talk about mental events. To that end, he attempts to find out how we could go about individuating these most simple forms in the first place – and already this stage of the project fails.[174] D. Raffman shows in [Raffman(1995)] that we can certainly distinguish two particular, barely noticeably different shades of red when faced with them, they are available for attention. However, we cannot identify or recognize them transtemporally.[175] Metzinger thus concludes:

> What Raffman has shown is the existence of a shallow level in subjective experience that is so subtle and fine-grained that – although we can *attend* to informational content presented on this level – it is neither available for memory nor for cognitive access in general. Outside of the phenomenal "Now" there is no type of subjective access to this level of content. However, we are, nevertheless, confronted with a disambiguated and maximally determinate form of phenomenal content.[176]

This leads to some obvious conclusions. Maximally simple content as such does not exist, because we are not able to even form a concept of it:

> So the problem precisely does not consist in that the very special content of these states, as experienced from a first-person perspective, cannot find a suitable expression in a certain natural language. It is not the unavailability of external color predicates. The problem consists in the fact of beings with our psychological structure and in most perceptual contexts not being able to recognize this content *at all*.[177]

However, unlike Dennett, Metzinger still calls his followup concept (or rather, his differentiated followup concepts) "qualia," and gives it special phenomenal properties that are ineffable or transparent. He also admits that Lewis qualia for some certain entities (like a pure red) exist[178] – but

[173] Citation from [Metzinger(2003)], page 79.

[174] It is interesting how here, a failure of the project he undertakes actually is a success for Metzinger.

[175] Diana Raffman finds that "our ability to judge whether two or more stimuli are the same or different in some perceptual aspect (pitch or color, say) far surpasses our ability to type-identify them. [...] The point is clear: we are much better at discriminating perceptual values [...] than we are at identifying or recognizing them." Citation from [Raffman(1995)] as quoted in [Metzinger(2003)], pages 69f.

[176] Citation from [Metzinger(2003)], page 70.

[177] Citation from [Metzinger(2003)], page 72.

[178] According to Metzinger, we can easily form a concept of a pure color. It will of course include a range of colors that are all classified as "pure red", even though the strict definition would exclude some of that that we are not able to distinguish without external references. See [Metzinger(2003)], page 73, where he writes:

> One form [of sensory content] is *categorizable* sensory content, as, for instance, represented by pure phenomenal colors like yellow, green, red and blue; the second form is subcategorical sensory content, as formed by all *other* color nuances. The beauty and the relevance of this second form lie in that it is so subtle, so volatile as it were, that it evades cognitive access in principle. It is nonconceptual content.

they do not exist for all kinds of experience, and thus they can not possibly be the basis of all phenomenal experience.

4.3.3 Lewis qualia, Raffman qualia, Metzinger qualia

Metzinger considers the discussion about qualia to date, and notices that there are some completely different readings of the term that are often used interchangeably. He offers the following distinction, according to the number of constraints they satisfy in descending order:[179]

- Lewis qualia: Globally available for attention, mental concept formation, and different types of motor behaviour such as speech production and pointing movements.
- Raffman qualia: Attentionally available and available for motor behaviour in discrimination tasks, but not available for cognition.
- Metzinger qualia: Attentionally available, but ineffable and not accessible to cognition, as well as not available for motor output.

Obviously, there is a decline in as how real events are perceived from one to the next of these kinds of qualia – Raffman qualia are perceived as being less real than Lewis qualia, and Metzinger qualia are perceived as being even less real than Raffman qualia. Nevertheless, there are examples for all three kinds; two of them we have seen in chapter 4.3.2, where a pure red would be a Lewis quale, while a a non-pure color would be a Raffman quale. Examples of Metzinger qualia (that would be immediately destroyed if the subject were to report them, since that would require making them available for motor control) are "highly abstract forms of consciously experienced mental content, as they sometimes appear in the minds of mathematicians and philosophers."[180]

There are more logically possible alternatives, of which only some make sense for us human beings (such as the instinctive movement when we are catching a ball, which is available for motor output, but not necessarily for cognition or attention – notably, this is not a function of the brain, but of the autonomic nervous system). The combinations mentioned here however are, according to Metzinger, "identifying the phenomenologically most interesting terms."[181]

4.3.4 Qualia are reducible

For Metzinger, qualia (understood as noncategorizable, but attentionally available forms of sensory content) are reducible in principle. He suggests tackling the issue of reducing them to physicalist terms by finding the physical structures upon which sensory content locally supervenes,[182] and by functionally analyzing the causal role of these structures.[183]

However, he does not banish qualia from the ontological landscape entirely: "On the contrary, this type of simple, ineffable content does exist and there exist higher-order, functionally more rich forms of simple phenomenal content – for instance, categorizable perceptual content."[184] Yet they are not the most simple form of mental content – rather, *presentational content* is.

4.3.5 Phenomenal presentational content

Presentational content is nonconceptual, cognitively unavailable, homogeneous, perceived as a fundamental aspect of reality, and fully transparent.[185] Metzinger finds that these characteristics

[179] See [Metzinger(2003)], page 74 for Lewis and Raffman qualia, and page 75 for Metzinger qualia.

[180] Citation from [Metzinger(2003)], page 76. A friend of mine recently updated his Facebook status with "Andrea Borsato, im Rausch des Denkens, fühlt sich so, als hätte er die Finger in eine philosophische Steckdose gesteckt." ("Andrea Borsato in the flow of thought, feels as if he had plugged the finger into a philosophical power outlet.") This is the kind of feeling Metzinger must mean.

[181] Citation from [Metzinger(2003)], page 90. The three combinations not mentioned (counting "only for motor behaviour" as stated) do indeed make a lot less sense – for example, how can something be available for cognition and attention, but not motor behaviour?

[182] See chapter 3.3.4.

[183] See [Metzinger(2003)], page 85.

[184] Citation from [Metzinger(2003)], page 86.

[185] See [Metzinger(2003)], pages 86f and 95.

of presentational content finally properly conceptualize what it means that something (namely, said presentational content) is *ineffable*:

> For the first time it allows us to do justice to the fact that a very large portion of phenomenal experience, as a matter of fact, is *ineffable*, in a straightforward and conceptually convincing manner. There is no mystery involved in the limitation of perceptual memory. But the beauty of sensory experience is further revealed: there are things in life which can only be experienced *now* and by *you*. In its subtleness, its enormous wealth in highly specific, high-dimensional information, and in the fine structure exhibited by the temporal dynamics characterizing it, it is at the same time limited by being hidden from the interpersonal world of linguistic communication.[186]

Presentational content also locally supervenes on internal physical and functional properties, like any other form of mental content. Metzinger offers presentational content as a followup concept for qualia, or rather, a concept which would explain what seems to be so special about qualia. At the same time, "presentational content never occurs alone; it is always incorporated into a higher-order whole"[187] in terms of the "convolved holism" constraint from chapter 3.3.8.

Phenomenal presentational content then is experienced from a first-person point of view, but it is important to recognize that this can not mean that it is presented *to* an entity (say, a homunculus). Moreover, it is not simply property exemplification – because that would already imply that those properties could be compared to each other, requiring transtemporal and logical identity criteria.

Metzinger goes beyond traditional models and suggests that "our own consciousness is by far too subtle and too 'liquid' to be, on a theoretical level, modeled according to linguistic and public representational systems."[188] Instead, he offers this view:

> Starting from elementary discriminatory achievements we can construct "quality spaces" or "sensory orders" of which it is true that the number of qualitative encodings available to a system within a specific sensory modality is given by the *dimensionality* of this space, and that any particular activation of that form of content which I have called "presentational" constitutes a *point* within this space, which itself is defined by an equivalence class with regard to the property of global indiscriminability, whereas the subjective experience of *recognizable* qualitative content of phenomenal representation is equivalent to a *region* or a *volume* in such a space.[189]

This is a rather mathematical description. However, the idea that qualia as such do not exist, but rather every recognizable quale corresponds to a (non-divisible by introspection) volume in this "quality space," makes sense and also takes into account that there certainly are numerous wider or narrower interpretations of, for example, redness. This "quality space" then is continuously regenerated in a dynamical process, and no actual entities are generated – rather, ceteris paribus: The same input always drives the same presentational processes (which are then experienced as always the same objects being perceived in the same ways).[190] As such, presentational content is always temporal content, too, in that it presents what happens *right now*.

Presentational content itself, the way Metzinger defines it, has interesting properties. Phenomenal states generated in this kind of process are particularized by three important implications:[191]

1. To the system, they have to appear as fundamental aspects of reality (generating a reference system, the "world zero"), because they are available for guided attention, but cannot be further subdivided.

[186] Citation from [Metzinger(2003)], page 95, emphasis his.
[187] Citation from [Metzinger(2003)], page 88.
[188] Citation from [Metzinger(2003)], page 93.
[189] Citation from [Metzinger(2003)], page 93, emphasis his.
[190] Ceteris paribus because the brain itself can of course be restructured by external or internal events, that can then also change those equivalence classes. Somebody who spends his life tuning pianos will have more (acquired) conceptual categories for "the way a piano sounds" than somebody who never listens to one.
[191] Adopted from [Metzinger(2003)], page 95.

2. They are fully transparent.

3. They enable us to take a step towards the functional and neuroscientific investigation of the physical underpinnings of sensory experience.

For Metzinger, my approach (and in particular the order of the chapters) in this thesis is the wrong way around. It is not phenomenality that enables the perception of qualia, but rather, presentational content (which can be experienced as qualia) is the foundation of phenomenality. This is one of the major changes I make to the order of partial concepts in the summary in chapter 7.

4.4 Consolidation

Qualia were a somewhat diffuse topic when they were first introduced, and the issue is still quite complex. Evidence and thought experiments contradict the idea that qualia are atomic metaphysical entities that experience is made of. Dennett felt this was sufficient to totally abandon the term, while Metzinger on the other hand redefined 'qualia' and gave them a less metaphysical meaning.

Let us look closer at how exactly their views differ. Both philosopher's concepts share some similarities:

- Both concepts of qualia are in principle reducible.
- Both concepts of qualia include differentiated ways in which a particular experience can be classified into categories.

However, there are important differences between their theories. The most striking ones are:

- Dennett abandons the term "qualia" altogether, while Metzinger gives the term "qualia" an ontological legimitation, namely, that of a specific categorizable and transtemporally comparable volume in quality space. Still, Metzinger attempts to analyze them and says that that there is nothing atomical about them; rather, they are merely paradigmatic definitions of arbitrary abstraction.
- Dennett's concept of what traditional qualia are is (in addition to their reducibility in principle) also practically reducible to purely syntactical terms, in that that phenomenological descriptions can always be intersubjectively specified and conceptualized. Metzinger, on the other hand, in addition to Dennett's intersubjectivizable entities, postulates that there is transparent presentational content, subsymbolic and non-categorizable, which can not be intersubjectively specified or conceptualized.

Metzinger allows an intersubjective view at qualia, and even presentational content: These entities can be intersubjectively individuated by means of the neural correlates they supervene on,[192] and their functional role. However, the specific what-is-it-like-ness[193] is only available from the first person point of view.

Dennett goes a lot further and states that all such experiential content is not only intersubjectivizable, but understandable in its entirety by any other conscious system, even if this other system has completely different dispositions and experiential contents.

4.4.1 The difference between appreciation and description

Essentially, Dennett postulates that 'appreciation' (see chapter 4.2.3) is the same as understanding a description, meaning that a description is able to comprise all aspects of an experiential content, and that reading such a description enables a conscious system to 'fully relive' (which seems to be the proper way to read 'appreciate') those contents.

[192] See chapter 3.3.4.
[193] See chapter 4.1.

Dennett disagrees with Nagel[194] and says that Nagel's claims (that human beings would only be able to know what it is like to be a human who *pretends* to be a bat) have no foundation. Dennett believes that it would be perfectly possible for us to know what it is like to be a bat. This implies two things: That the description can completely capture the experience, and also that fully reliving the experience (appreciating it) based on such a complete description is possible. You will recognize the following citation from chapter 4.2.3:

> If we want, we can carefully list the differences between our dispositions and knowledge and [the people who did actually hear Bach originally], and by comparing the lists, come to appreciate, in whatever detail we want, the differences between what it was like to be them listening to Bach, and what it is like to be us.[195]

The facts do not treat Dennett's theory well. There is mental content which is available for attention, but not for concept formation; Metzinger and Raffman qualia from chapter 4.3.3 are. This mental content itself cannot be made intersubjectively available – the best we can ever hope to achieve is to map which are the neural correlates it supervenes on.[196] Mapping the functional roles would be necessary as well, but even more complex. I doubt that this is ever possible, unless we fully understand *all* aspects of the particular, individual brain in question. It is a phenotype (maybe even an extended phenotype[197]) we are talking about, not a genotype. Thus even completely understanding the human brain in general (and not the specific brain that had the experience in-depth) would not be enough.[198]

Moreover, if we were ever to obtain such a complete description of all functional roles and neural correlates of a certain experiential content, the only way to 'relive'[199] it would be to have exactly the *same* neural correlates activated in the *same* way with exactly the *same* functional roles. Ultimately, this means that a brain which is exactly the same has to be the one reliving the content, as otherwise, at least some of the (many and complex) functional roles will be different.[200]

This amounts to the fact that we can not fully appreciate tales about our childhood and, arguably, some childhood or otherly faint memories themselves, even if we do not have to be reminded of them by third parties. Consider William James' example:

> We hear from our parents various anecdotes about our infant years, but we do not appropriate them as we do our own memories. Those breaches of decorum awaken no blush, those bright sayings no self-complacency. That child is a foreign creature with which our present self is no more identified in feeling than it is with some stranger's living child today. [...] It is the same with certain of our dimly recollected experiences. We hardly know whether to appropriate them or to disown them as fancies, or things read or heard and not lived through.[201]

[194] See chapter 4.1, and [Nagel(1974)].

[195] Citation from [Dennett(1991)], page 388.

[196] Note that due to the fact that the brain rarely does only one thing at once, even this task will be hard – yet not impossible, probably best achievable through differential analysis of a series of tests.

[197] See [Dawkins(1982)].

[198] This argumentation is assuming that we are talking about the kinds of qualia that *specific* humans have. Since the particular kinds of memories and associations any event triggers is different from human to human, this appears to make sense.

[199] Really appreciating and thus reliving a description would need to be an exact repetition of all the conceptual and nonconceptual mental contents, including a sense of self-identification.

[200] I made an example of this in [Gloor(2007)], page 24, referring to Searle using a hamburger in [Searle(1980)]:

> There's other things too that aren't even intersubjective (or, even less intersubjective than the attributes already mentioned). For instance, one person might like hamburgers while another one doesn't, for one person the first intuition with hamburgers is Krusty Burgers while another one thinks of Burger Queen first, one attributes "tastes good" and another one "makes fat" as a first thought – human memory systems are very different in that aspect, just like any other, and every person has a different mental image when thinking about the same things.

[201] Citation from [James(1890)], as cited in [Minsky(2006)], page 308.

4.4.2 Dennett's "multiple drafts" and qualia

Dennett's "multiple drafts" model from chapter 4.2.2 seems to be an adequate description of what happens in the brain, and can and should act as a replacement for the concept of Cartesian theaters.[202] It takes the massive distributed parallelism in our brains into account, and computational models based on it (or rather, models based on the same thoughts and paradigms) seem to prove the case for Dennett's theory. A fine example is Watt's research into auditory pathways,[203] where intermediate representations are reliably constructed and multiple versions of the data are available for further processing (which versions, in his case, include visualization and probably logging).[204]

However, Dennett also presents his model of multiple drafts as an alternative to the traditional theories of qualia, not only as a stand-alone theory. Let us look at how exactly it is not only a fresh approach to how the brain works, but also an alternative that can explain the same things as qualia did. The question whether the theory of multiple drafts necessarily implies that there are no (non-eliminable) qualia seems particularly interesting.

In order to make the transition from qualia to the theory of multiple drafts, Dennett first notes that one of the prime motivations behind the concept of qualia was that the entities that qualia must be would only be "in the eye and brain of the beholder,"[205] not out there in the world ready for us to observe. However, if they are in our brains, there is no inner "figment"[206] that they are made of either; Dennett categorically denies that entities like "occurrent pink" would need to be included in popular science. Unlike Sellars, but like Metzinger,[207] Dennett is not opposed to the idea that us making judgements about a color is all there is to color vision, but he is also convinced that color vision (or rather, our capabilities of making judgements about a color influenced by the external world) co-evolved with colored entities in the world.

This implies that the color is out there somehow after all. It is there, ready to be presented to us, and our detectors, specificially tailored to detecting them, produce representations within our brains. These detectors are specificially built to detect color, and Dennett introduces the concept of such purpose-built detectors with an analogy:

> A particularly useful example is provided by the famous case of Julius and Ethel Rosenberg, who [...] improvised a clever password system: a cardboard Jell-O box was torn in two, and the pieces were taken to two individuals who had to be very careful about identifying each other. Each ragged piece became a practically foolproof and unique "detector" of its mate: at a later encounter each party could produce his piece, and if the pieces lined up perfectly, all would be well. [...] tearing the cardboard produces an edge of such informational complexity that it would be virtually impossible to reproduce by deliberate construction. [...] The particular jagged edge of one piece becomes a practically unique pattern-recognition device for its mate; it is an apparatus or transducer for detecting the shape property M, where M is uniquely instantiated by its mate.[208]

Obviously, if entities (or even entity properties) are out there ready to be observed, even though

[202] I described Dennett's concept of the Cartesian theater in a footnote in chapter 4.2.2.

[203] See [Watts(1993)], and in particular [Watts(2009)], as referred to in [Kurzweil(2001)].

[204] An interesting difference is that in this case, the observing the process does not alter it, which is a big difference to observing our own internal processes through introspection. Dennett points this out in [Dennett(1991)], page 407, referring to the thought experiment I cited in chapter 4: "Had we probed earlier – had the author picked up a tape recorder while he sat rocking, and produced the text there and then – it would surely have been quite different. Not only richer in detail, and messier, but also, of course, reshaped and redirected by the author's own reactions to the very process of creating the text – listening to the actual sounds of his own words instead of musing silently."

[205] Adopted from [Dennett(1991)], page 370.

[206] Dennett uses this term to denote the figmentary paint that the cartesian theater is made of, in [Dennett(1991)], pages 370f:

> But now, if there is no inner *figment* that could be colored in some special, subjective, in-the-mind, phenomenal sense, colors seem to disappear altogether! *Something* has to be the colors we know and love, the colors we mix and match. Where oh where can they be?

[207] See chapter 3.3.14 and [Sellars(1963)], as referred to in [Dennett(1991)], page 372.

[208] Citation from [Dennett(1991)], page 376.

we might not be able to clearly define them by their physical description alone, there must be detectors that reliably find them. And how these detectors are built at their core is precisely what the multiple drafts theory is about. The theory explains perfectly well how it makes sense that qualitative properties are out there in the world, ready for us to perceive, and how that perception can work without "figment" or occurent qualities.

What Dennett does not adress is the ineffability of qualia, the distinct way they look to us. He does not aim for appreciation, but merely description.[209] More importantly, he also does not adress the lower-level components of the process of perception or attention that are, as Metzinger points out, sometimes non-categorizable and not available for concept formation.

Dennett's theory of multiple drafts thus is an answer to how M-detectors can detect M-properties, and how they can make those M-properties globally available in a step by step procedure[210] which finds 'easy' properties first and 'hard' properties later.[211] It is, however, not a theory that really makes qualia obsolete.

4.4.3 Qualia: Still necessary?

According to Metzinger, qualia are reducible in principle, once we agree that phenomenal presentational content is an ontological possibility and most probably what makes us experience qualitative properties in our brains.

However, they still are a necessary discriminatory step. We have a metaphorical quality space in which phenomenal presentational content occurs, and we are able to phenomenally conceptualize volumes within that quality space transtemporally.[212] These volumes are what qualia depict; they correspond to categories we are able to form within that entire quality space.

It may be possible to distinguish between first-order and second-order qualities.[213] Regardless of whether they are primary or secondary however, there are ontologically legitimate qualia of "redness" into which all kinds of color perception that are red (within distinct, certainly individual boundaries) fall.

To conclude, qualia are at the same time reducible and necessary – a rather odd place to be in. They form the categorical borders of our conscious experience, but do not do the true ineffable nature of phenomenal experience justice, nor are they the truly atomic building blocks of experience. They are a necessary level of abstraction that holds information not contained anywhere else.[214]

And, being categorizable[215] unlike many forms of direct sensory perception, qualia are what our memories are made of.

---

[209] See chapter 4.4.1.

[210] Even though the detecting procedure has multiple steps, these are processed in parallel, except when one process has to wait for the results of another.

[211] Note how 'easy' and 'hard' do not have to correspond to the *perceived* complexity of the processing steps involved, but rather with how optimized the system in question is for solving them. Things like detecting that the shape in front of us is a face which is looking to the right and belongs to your grandmother is a very complex and complicated thing, while playing chess is comparably easy, from a computational point of view. Just because something seems easy to us does not make it easy for another system – it is just that our bodies have adapted to the social requirement of such face detection mechanisms for millions of years, while the ability to play chess has formed in the last couple of thousand years only.

[212] See chapter 4.3.5 and the citation from [Metzinger(2003)], page 93.

[213] Color, for example, appears to be a secondary quality of features such as shape and motion – see [Locke(1690)], as referred to in [Dennett(1991)], page 371.

[214] See chapter 8.2.

[215] At least Lewis qualia are categorizable, whereas Metzinger and Raffman qualia are precisely the kinds of sensory perception that are non-categorizable – see chapter 4.3.3.

5 Subjectivity

There is more to consciousness than merely having phenomenal experiences through qualia. We are able to put these things into context with something that seems to be a constant over time, a personal unity: the self. This is a fairly amazing trait, which has some very interesting characteristics.

For example, let us assume that you think about something you did as a four years old child – either recalling through your own memories, or through reminders from your parents or other relatives. In such a situation, we might not be able to emphatically put ourselves into the shoes of that four years old child anymore; the world we lived in back then seems nearly as alien to us now as the world of Nagel's bats. We have different value systems now, different and many more experiences we can relate to, different theories regarding both the world around us and ourselves. But most philosophers agree that this four years old child *was* you, however many differences there might be.

Another example is intentionality (see chapter 6). When thinking about directedness, there necessarily has to be a point where this arrow of directedness originates from. This point is always the same, an untranscendable "me" that stays constant throughout all our experiences.

This seems to be the just moment to incorporate parts of Minsky's work into our deliberations: He observes that often, we have not just one, but multiple self models. Normative ones (what do I want to be like?), descriptive ones (how do I usually act in this situation?), of both ourselves and others. Often, the models we have of others are (or may appear) more accurate than the models these others have of themselves. How can that be? Should I not be the final authority when judging what it is that makes me myself? Do not I alone have some kind of privileged access that should enable me to form a more accurate, a more proper 'self' than others, who merely form mental (empathical) models of me?

Interestingly, there are many deviate forms of subjectivity. One of these is a recurring theme among Metzinger, Dennett, and many other philosophers of mind: The multiple personality disorder. Certainly, the best opportunity to examine the 'self' is when it is already dissected, when its inner workings are more prone to lay bare because the protective layer around it has been removed and broken. This is metaphorically spoken, but frequently true in a stricter sense, as it seems that traumatic childhood experiences are often what causes such multiple personality disorders in the first place.

Admittedly, the secret of subjectivity was one of the prime forces which prompted me to write this thesis. As I noted in [Gloor Modjib(2008)], I had some difficulties understanding what constitutes a self. I still consider it to be one of the most fascinating aspects of the philosophy of mind. However, the theories of Dennett, Minsky and particularly Metzinger seem to be concise argumentations for a modern and scientific model of what this 'self' could be.

5.1 Descartes: Dualism

Descartes essentially postulates two distinct entities that make up a thinking thing (like, for example, a human being):[216]

- The *res cogitans*, a matter that is thinking and not extended
- The *res extensa*, a matter that is extended and not thinking

These entities are distinct from each other because Descartes can perceive them clearly as being distinct (which would not be allowed by God if it were not true), and God can form such distinct entities that are nonetheless linked. Furthermore, ideas of external entities are only formed within the *res cogitans*, and they are truthful representations of actual external entities, because God is not a deceiver. Both these (in this thesis, very bluntly put) argumentations rely on the truthfulness and benevolence of God, of which further proof is attempted earlier in Descartes' work.

[216] See [Descartes(1642)].

However, as soon as we include omnipotence in a theory, we can just as well stop arguing about rational reasons and causes, so I would prefer a theoretical framework that does not rely on metaphysics.

There are more modern dualist argumentations[217] that do not require supernatural entities (and thus better satisfy Occam's Razor[218]). But even if we assume a rational explanation and proper scientific methodology, substance dualism, which is the kind of dualism at the focus of this discussion,[219] essentially is the postulation of two kinds of essences: A physical, nonthinking essence (which builds the world, the body and the brain) and a nonphysical, thinking essence (which builds the mind). Thus, it is the separation of body and mind, and it did not really start with Descartes – it dates back to Zarathushtra, Plato and Aristotle.

Substance dualism is subject of further discussion in chapter 5.5.1.

5.1.1 Finding the subject in dualism

The somewhat pragmatic (and probably for that very reason intuitively plausible) approach of Descartes, when he proposes differentiating between a thinking matter that is not extended and an extended matter that does not think, makes it very easy to point at what creates subjectivity, which at this point becomes an only very loosely defined term.

This is because the very distinguishing criterion for the *res cogitans* is that it is able to produce this subjectivity. As such, subjectivity is built into the *res cogitans*; this is a logical necessity because a subject is a prerequisite for mental contents of any form that are experienced from a first-person point of view.

Descartes, being deeply religious,[220] roots this separation of entities between the *res cogitans* and the *res extensa* in the soul and, in the end, in the divine spirit. Strictly looking at what constitutes subjectivity itself however does not reveal anything but dogmas – it is not further analyzable, atomic to scrutiny, as the only form of scrutiny available in a purely non-physical context is introspection.

Thus, since we are neither able to analytically (physically) analyze a purely non-physical entity, nor able to introspectively (mentally) dig deep enough to expose the metaphysical processes that would expose subjectivity itself, the following is quite all there is to say on this matter:

Subjectivity is that which defines the *res cogitans*.

5.2 Dennett: Center of narrative gravity

Dennett points out the two contradictory stances opinions about selves can take:[221]

- Obviously, there are selves in the world: We exist! The question presupposes its own answer.
- Obviously, there are no selves in the world: There are no distinct entities that control our bodies and think our thoughts, neither in our brains nor beyond[222] our brains.

[217] Those more modern theories include, among others: epiphenomenalism (first introduced by La Mettrie), psychophysic parallelism (as per readings of Leibniz and Malebranche), teleological supervenience (most prominently discussed by Sehon), occasionalism (as taught in the Ash'ari schools of early Islamic philosophy).

[218] Attributed to William of Ockham, Occam's Razor is the postulation that "entities must not be multiplied beyond necessity," a guideline that states that often, simpler theories with less dogmating settings are better than more complicated ones, if both kinds are able to explain the same results.

[219] The other major kind of dualism is conceptual dualism – the view that things are not strictly reducible to more abstract levels of description (for example, going from mental states to neurons) without losing valuable information. The views expressed in this paper could be seen as a form of such conceptual dualism. See chapter 8.2.

[220] He was, of course, a man of science and had a very investigative spirit, something that is arguably rarely seen in today's rather traditionalist clergy of most religions; insofar he was religious in a sense that today would not be typical.

[221] See [Dennett(1991)], page 413.

[222] Since these distinct entities would necessarily be metaphysical, they would be "beyond" in a metaphysical sense as well.

These positions are contradictory, but both intuitively plausible, thus a reasonable conclusion seems to be that they both are partially true.[223] Dennett has a proposal that is a mix of the two, indeed. This proposal, however, is not as intuitive[224] as the two just presented premises, however.

5.2.1 On evolution and absolutism

Dennett introduces his views on how selves benefitted our ancestors during the evolution of what was to become human beings.

Selves must have been created through evolutionary processes: The first single-cell organisms will most probably not have had selves, yet today we do have them. Therefore, organisms with selves obviously evolved from organisms without selves.[225] For us and similar organisms, then, the self is a tool, a means that allows us to distinguish between what is "me" and what is "the rest of the world"[226] – porously and with less clear boundaries than we would imagine, as there are synergies and symbiotic relations with some microorganisms in our stomach, for example.

Dennett talks about the extended phenotype,[227] which may include things like a spider's web, or a beaver's dams, or a hermit crab's shell, even a human's clothes (or for some of us, their cars), where 'our own' territories are included. The way we human beings include a self in that extended phenotype is fascinating:[228]

> Each normal individual of this species [homo sapiens] makes a *self*. Out of its brain it spins a web of words and deeds, and, like the other creatures, it doesn't have to know what it's doing; it just does it. This web protects it [...] and provides it a livelihood [...] and advances its prospects for sex [...].[229]

The illusion that a soul is necessary for a self is no more warranted than the notion that termite colonies have souls.[230] The seemingly highly organized collaborative work this colony produces is nothing but the product of "a million of semi-independent little agents, each itself an automaton, doing its thing,"[231] as most scientists today agree. Dennett shows an interesting parallel:

> So wonderful is the organization of a termite colony that it seemed to some observers that each termite colony had to have a soul [...]. So wonderful is the organization of a human self that to many observers it has seemed that each human being had a soul, too: a benevolent Dictator ruling from Headquarters.[232]

[223] Dennett uses a similar line of argumentation in chapter 4.2.1, when he suggests a middle ground between Orwellian and Stalinesque revisions.

[224] Intuition is a strange concept, because in this context it relies on imagination; it is a disposition to consider certain imagined facts as true. This makes intuition very brittle, because failure of imagination is not an insight into necessity – what Dennett calls the "philosopher's syndrome" in [Dennett(1991)], page 401. It is not the case that something is necessarily untrue merely because we cannot imagine it to be true. We are built a certain way, with certain dispositions that helped us survive long enough to be where we are now. Those dispositions were (and, most of them, are) adequate for survival, and not necessarily insights into reality.

[225] I am assuming that the common theories on evolution, including Darwin's revolutionary work, make sense. Susan Blackmore made a very apt recapitulation of [Darwin(1859)] in [Blackmore(2008)]:

> If you have species that vary, and if there is a struggle for life such that nearly all of these creatures die, and if the very few that survive pass on to their offspring whatever it was that helped them survive, then those offspring *must* be better adapted to the circumstances in which all this happened than their parents were.

In addition to this, my reading of evolution also does not want to rely on the supernatural, as relying on that is a position of faith and not of science.

[226] See [Dennett(1991)], pages 174f and 414.

[227] See [Dawkins(1982)] for the origins of the concept, and [Dennett(1991)], pages 415 and 417, for Dennett's adaptation. It is worth noting that there are also cultural influences that shape the extended phenotype, in addition to genetic ones.

[228] It is interesting that Dennett only talks of humans here, although other animals like for example octopi are highly probable to have selves as well (for anecdotal evidence, see [Telegraph.co.uk(2008)]).

[229] Citation from [Dennett(1991)], page 416, emphasis his.

[230] See [Marais(1937)] for an actual proposal of such a termite colony's soul.

[231] Citation from [Dennett(1991)], page 416.

[232] Citation from [Dennett(1991)], page 416.

Dennett regards the self in human beings essentially to be a "center of narrative gravity for a narrative-spinning human body,"[233] an enormous simplification and abstraction of the complexity of the action that happens in the system by means of the multiple drafts model. It makes referrals and ownership appropriations easy and arguably even possible.

Dennett also points out that a self can not plausibly be an all-or-nothing phenomenon, and that absolutism[234] is misguided:

> Since selves and minds and even consciousness itself are biological products (not elements to be found in the periodic table of chemistry), we should expect that the transitions between them and the phenomena that are not them should be gradual, contentious, gerrymandered. This doesn't mean that everything is always in transition, always gradual; transitions that look gradual from close up usually look like abrupt punctuations between plateaus of equilibrium from a more distant vantage point.[235]

Dennett continues:

> But many people who are quite comfortable taking this pragmatic approach to night and day, living and nonliving, mammal and premammal, get anxious when invited to adopt the same attitude toward having a self and not having a self.[236]

5.2.2 What a self is

Ultimately, Dennett's proposition as to what a self exactly is is not that different from Metzinger's that we'll see in chapter 5.3 – although he explores the concept much less in-depth. Dennett has this to say concerning the part of the hard problem we are currently looking at:

> A self, according to my theory, is [...] an abstraction defined by the myriads of attributions and interpretations (including self-attributions and self-interpretations) that have composed the biography of the living body whose Center of Narrative Gravity it is. As such, it plays a singularly important role in the ongoing cognitive economy of that living body, because, of all the tings in the environment an active body must make mental models of, none is more crucial than the model the agent has of itself.[237]

However, that is all he says on the subject of what a self actually is. He explores further why it makes sense to have a self, and how magnificient a fiction a self is from a functional point of view (and maybe from a teleological point of view as well, although he does not use that word). He also explores the moral implications of a self that is taken to be a fiction only, concerning responsibility and agency. Since this moral tangent is not main subject of this thesis, I will not further pursue that issue.

5.3 Metzinger: Being no one

Metzinger's convincing argument goes deeper and further than those of other thinkers. On the surface, it is quite similar to Dennett's: The self is merely a model we create. The following section shows Metzinger's reasoning.

From an evolutionary point of view, it makes sense to regard the self as merely a model. As Metzinger points out, such a self model can both be a tool and a weapon.[238] Awareness of the existence and continuity of a self as opposed to the world is what enables self-directed action in a stricter sense, in two ways:

[233] Citation from [Dennett(1991)], page 418. Again, Dennett only talks of humans, when the argumentation could easily be extended to include other animals with selves.

[234] Absolutism here is intended to mean seeing things only in absolutes, i.e. when it comes to selves, only allowing systems that have a full self and systems that have no self at all.

[235] Citation from [Dennett(1991)], page 421, referring to [Eldredge and Gould(1972)], and [Dawkins(1982)], pages 101-109.

[236] Citation from [Dennett(1991)], page 422.

[237] Citation from [Dennett(1991)], pages 426f.

[238] See [Metzinger(2003)], pages 344ff.

- It is necessary to have a self-model in order to be able to plan and carry out directed action at what constitutes which space the self is perceived as being in.
- It is necessary to have a self-model in order to be able to realize that any action is directed at things that are part of the own body.

A self-model also makes various internal parameters (e.g. hunger, thirst, hormone and transmitter levels, intestinal states, skin temparature) available for cognitive processing and thus reflected action. This reflected action, then, is not strictly stimulus-related, but probably better understood as being stimulus-inspired, in that stimuli may result in activating offline phenomenal states that simulate the self-model in certain counterfactual situations. Only after activating several such possible worlds and comparing them with the underlying "world zero,"[239] we decide on one such possible world and perform the actions that we hope and believe will lead to it.

5.3.1 The self is not an illusion

In order to avoid misunderstandings, I regard it important, at this point, to make clear that it is wrong to say that it is wrong to say that the self is *an illusion*. This is because if the self were an illusion, there would need to be somebody or something (a self?) having and owning the illusion. This would necessarily render the entire argumentation circular – which is precisely what Metzinger's self-model theory has been criticized for.

Metzinger is aware of this, and also of the fact that it certainly *seems* to us as if there were a self, and not merely a phenomenal model of a self. In order to explain how this can possibly happen, he draws on Plato's thought experiment of the cave[240] where we are pictured as being chained to a wall, unable to turn our heads, and merely seeing shadows of the things that actually are,[241] cast by a fire in the middle of the cave.

There are many similarities between Plato's cave and Metzinger's self-model theory. The allegory works for quite a few of Metzinger's postulated entities: According to Metzinger, the cave is the complete phenomenal self-model that is formed by the entire organism. The shadows are low-dimensional phenomenal projections of high-dimensional objects,[242] and the fire is neural dynamics – information processing that is constantly perturbed and modulated by sensory and cognitive input. Finally, the cave's wall is merely another aspect of the very same neural process.[243] The parallels between the two thought experiments are indeed considerable, the various entities are directly mappable.

However, there is a fundamental difference between Plato's and Metzinger's view. Plato believes that it is possible to get up, go out into the sun, and then come back and teach people what true forms[244] are. He believes that there can be enlightenment, true knowledge – epistemology demystified, experienced ontology possible. Unfortunately, Metzinger does not follow Plato here – there is no getting up in Metzinger's cave, we have to remain chained.

Metzinger doubts that we can get up for one reason: There is nobody in the cave – it is empty, the phenomenal self-model is the entire cave and all that is contained within it. The self-model

[239] See chapter 3.3.7 for Metzinger's world zero hypothesis.

[240] See [Plato(380 B.C.E.)], and Metzinger's referral in [Metzinger(2003)], pages 547ff. I will assume that you are familiar with Plato's thought experiment of the cave that he put forth in [Plato(380 B.C.E.)], so I will be very brief on its details here.

[241] Platon has two categories of things that actually are: Ideas, which are not epistemically graspable by humans directly, and the entities in the real world that are merely implementations of these ideas. The shadows on the wall of the cave are, in his original work, merely of works of human beings, not even of humans themselves.

[242] The definition of "dimensions" when we are talking about shadows and objects are pretty straightforward: They are actual two-dimensional mappings on a surface, cast by three-dimensional objects. How this translates to Metzinger's theory: The dimensions of entities that are perceived as actual objects in the world are limited by the number and nature of the senses we perceive them with. We can only perceive the properties that our senses are made for, and even then we can only categorize those perceptions in rather rough classifications, with limits in all directions. What other, maybe just as important, properties an entity might have we can not directly perceive.

[243] See [Metzinger(2003)], pages 548f.

[244] True forms are essentially pure implementations of the aforementioned ideas, or even the ideas themselves. These are only rationally analyzable, they are abstract concepts, unlike the shadows in the cave that are physically perceptible.

experiences itself as being in the center, however. Shadows, phenomenal representations, real as they may seem, are merely interpretations of shaded surfaces. The entire cave is nothing but a "continuous online dream about, and internal emulation of, itself,"[245] and what we introspectively perceive as the equivalent of a Platonic shadow of ourselves is nothing but the shadow of the cave itself.

5.3.2 The phenomenal self-model (PSM)

The PSM, the "phenomenal self-model", is how Metzinger formally names this part of his theory, and it is the more fundamental of the two main building blocks of it.[246] The PSM is the logical conclusion from what we have explored so far, and in particular, the consequence which follows from his multilevel constraints and the fact that the system is mistaking the generated transparent self-representation in a naive-realistic self-misunderstanding[247] as a self that actually is not merely a part of an ongoing process but has ontological legitimation that goes beyond abstract entities.

This is Metzinger's main argument for the PSM:

> If all other necessary and sufficient constraints for the emergence of phenomenal experience are satisfied by a given representational system, the addition of a transparent self-model will by necessity lead to the emergence of a phenomenal self. Phenomenal selfhood results from autoepistemic closure in a self-representing system; it is a lack of information. The prereflexive, preattentive experience of being someone results directly from the contents of the currently active self-model being transparent. [...] Under a general principle of ontological parsimony it is not necessary (or rational) to assume the existence of selves, because as theoretical entities they fulfill no indispensable explanatory function. What exists are information-processing systems engaged in the transparent process of phenomenal self-modelling. All that can be explained by the phenomenological notion of a "self" can also be explained using the representationalist notion of a transparent self-*model*.[248]

As a matter of fact, there seems to be no need for a full-fledged perspectivalness constraint for the PSM. Rather, it is the other way around: The PSM *is* a major part of the perspectivalness constraint. There is first the prereflexive proto self-model, that "effortless way of inner acquaintance,"[249] then there is the PSM, and then there is the PMIR (see chapter 6.3). These three parts together form the perspectivalness constraint and thus the full 'self', as discussed in chapter 3.3.10.

As soon as the prereflexive proto self-model is experienced phenomenally and transparently as an entity in the world by a system, it thus logically follows[250] that a PSM forms. All other parts of the self-model, the phenomenal self that is experienced as thinker of the system's own thoughts and doer of the system's deeds, is merely coating: It is necessarily (through genetic disposition) added to that prereflexive self-model during an organism's growth and learning phase.

5.3.3 Genesis of a conscious phenomenal self-model

A phenomenal self-model does not necessarily have to form, even given many of Metzinger's multilevel constraints. He specifies the probable (and intuitively plausible) position of a critic of his model:

> There simply is no conceptually necessary connection [...] from the functional and representational constraints so far developed to the phenomenal target property of

[245] Citation from [Metzinger(2003)], page 550. The main difference between this kind of dream and the ones we have at night is that the latter ones (which we traditionally call "dreams") are *offline* dreams, while the dream we have of our own selves is *online*, influenced and to large extents directed by events external to the system itself.

[246] The other main building block is the PMIR, the "phenomenal model of the intentionality relation", see chapter 6.3.

[247] This is a terminology that Metzinger uses in multiple places – see chapter 5.3.3.

[248] Citation from [Metzinger(2003)], pages 336f.

[249] Citation from [Metzinger(2003)], page 158.

[250] It follows through the argumentation from [Metzinger(2003)], pages 336f, which was cited above.

> selfhood. The representational process of mental self-modeling within a coherent world-model and a virtual window of presence, holism and dynamics conceded as well, does not *necessarily* lead to the existence of a full-blown phenomenal self. [...] A system-model simply is not a *self*-model. [...] What is needed – by conceptual necessity – to take the step from the functional property of centeredness and the representational property of self-modeling to the consciously experienced phenomenal property of selfhood?[251]

Metzinger's argument to counter this then is very similar to the one we already discussed in chapter 3.3.17, where we established ineffability as being the defining characteristic of consciousness. It is exactly the phenomenal transparency (constituted by satisfying the transparency, globality and presentationality constraints) of the contents of our mental self-model which is, as one of two central prerequisites, necessary for the conscious self-model. The other is that the system has to be able to transparently recognize (and identify with) the contents of that self-model as itself. Metzinger puts this as follows:

> We do not experience the contents of our self-consciousness as the contents of a representational process, and we do not experience them as some sort of causally active internal placeholder *of* the system *in* the system's all-inclusive model of reality, but simply as *ourselves, living in the world right now.*[252]

Metzinger terms transparency[253] "a special form of inner darkness."[254] This darkness is constituted in the fact that the following properties are hidden from introspective access:

- Nearly all vehicle properties
- Many content properties[255]
- Many causal connections between all (vehicle and content) properties

A self then is a completely transparent self-representation; the phenomenally perceived representation of a self that the system generates and uses as a kind of "black box" in its processes. Metzinger defines it as follows:

> Completely transparent self-representation is characterized by the fact that the mechanisms which have led to its activation and the additional fact that a concrete internal state exists, which functions as the carrier of their content, cannot be recognized anymore. Therefore, the phenomenology of transparent self-modeling is the phenomenology of selfhood. It is the phenomenology of a system caught in a *naive-realistic self-misunderstanding.*[256]

5.3.4 Switching off the self

Sustaining a self is a hard piece of work for our brains. It thus makes perfect sense that the self is switched off from time to time, when it is not needed – which leads to times of partial or entire lack of conscious experience:

> The brain, the dynamical, self-organizing system as a whole, *activates* the [self] if and only if it needs the [self] as a representational instrument in order to integrate, monitor, predict, and remember its own activities. As long as the [self] is needed to navigate the world, the puppet shadow dances on the wall of the neurophenomenological caveman's phenomenal state space. As soon as the system does not need a globally available

[251] Citation from [Metzinger(2003)], pages 330f, emphasis his.
[252] Citation from [Metzinger(2003)], page 331, emphasis his.
[253] See chapter 3.3.11.
[254] For example, he does so in [Metzinger(2003)], page 331. I refer to this terminology in chapter 3.3.11 as well.
[255] As deficient as this distinction between vehicle and content properties may be in this context – see chapter 3.3.11.
[256] Citation from [Metzinger(2003)], page 332, emphasis his.

> self-model, it simply turns it off. Together with the model the conscious experience of selfhood disappears. Sleep is the little brother of death.[257]

The reason why we have not noticed this fact before is because of the transparent nature of our self-model – we cannot see it for what it is. The self-model is merely an 'illusion' nobody has.[258] This is the price we pay for having a self-model in the first place however; the price we pay for the increased autonomy[259] and thus conscious self-control we have gained through it.

5.3.5 Why "being no one"?

Already the title of Metzinger's book exhibits a problem which is worth discussing. It is not possible to be convinced of his theory of phenomenal self-modeling while remaining a subject.

In order to even understand Metzinger's explanation and the PSM, we have to take on a special epistemological stance:

> "Being no one" in this sense describes an epistemological stance we would have to take toward our own minds in scientifically and philosophically investigating them, an attitude that is necessary to really solve the puzzle of consciousness at a deeper and more comprehensive level, an attitude of research that integrates first-person and third-person approaches in a new way and that, perhaps unfortunately, appears to be strictly impossible and absolutely necessary at the same time.[260]

That special stance is one of a thought experiment which tries to resolve our dilemma. The epistemic irreducibility of any conscious experience, the fact that it is tied to a first-person perspective, may be explainable, but it is not possible to be truly convinced (which is necessary to make it intuitively plausible) by such an explanation:[261]

- Either I attempt to understand that there really are no selves – however, I attempt this while I still perceive myself as "self" in that naive-realistic self-misunderstanding, so there is a performative contradiction[262] between intentional point of view and intentional object.
- Or I let go of my intentional point of view and reach a transcendental state in which it really is epistemically possible that there are no selves – however, having let go of what makes me the 'I' in 'I am convinced,' I cannot be convinced of this.

Let us try to understand Metzinger 's theory about the self anyway, as good as we can. If we can not appreciate it or be truly convinced of it, we can still try and describe our own and our fellow agents' self-models as clearly as possible.

5.4 Minsky: Multiple models

Minsky approaches the problem of selves (or rather, the perception of one self to a person) from a different point of view. He doesn't care that much about the genesis of the phenomenality of a self-model (in fact, he borrows much of the philosophical footwork from Dennett), but rather looks at the actual causes and implications of self-models, and notices that a person usually has a wide array of multi-tier self-models at their disposal. Those include normative just like positive / descriptive aspects, and have a complex internal structure.

First though, we have to look at what it is that makes the brain have multiple selves in the first place.

---

[257] Citation from [Metzinger(2003)], page 558, emphasis his. The context of this sentence is a thought experiment where the self is represented as a pilot, I replaced the occurences of "pilot" with "[self]."

[258] Technically, that makes it not an illusion at all, see chapter 5.3.1. Nevertheless, some form of proto self-model must exist for this kind of process, see chapter 5.5.3.

[259] We mainly gain increased autonomy because we are able to activate offline mental content, independent of the external world, generate counterfactuals, and thus form plans and ideas. See chapter 3.3.12.

[260] Citation from [Metzinger(2003)], page 628.

[261] See [Metzinger(2003)], pages 627f.

[262] It is a performative contradiction and not a logical one because the subject and the object are different kinds of entities, and thus assuming that it is a logical error would be a category error. See chapter 6.4.2.

5.4.1 The six levels of mental activities

According to Minsky, there are six levels of mental activities, the total of which together then constitute consciousness – and produce the self-models.

Minsky, by postulating many levels which seem to be very similar at first sight, consciously disregards Occam's Razor in that he does not attempt to minimize the number of levels. Obviously, that minimization is what science (and in particular, Cartesian philosophy) has done for quite some time, without much success in resolving how consciousness works, or what constitutes a self.[263] Minsky follows exactly this line of argumentation when he states:

> I think [this policy of minimizing complexity] has badly retarded the field of psychology. For when you *know* that your theory is incomplete, then you ought to leave some room for other ideas that you later might need. Otherwise, you will take the risk of adopting a model so clean and neat that new ideas won't fit into it.
>
> I think that this applies especially to making theories about complex structures like brains, for which we still know little about what their functions actually are, or the details of how they evolved.[264]

The levels of mental activities that Minsky describes are:[265]

1. 'Instinctive Reactions': These are reactions to stimuli that we are genetically predisposed to do – and which often do not fit into the fast-paced modern world.[266] There are many instinctive things which still make sense today however; regulating body temparature and heartbeat rate, or squinting when we look into a bright light source.

2. 'Learned Reactions': This involves considerable brain power already, because several aspects of the situation at hand (which parts of it have to be remembered, which steps would lead to success), and in what ways the memorization necessary for learning can happen, need to be analyzed[267] and realized (or, as Metzinger would say, embodied).[268] Many other animals (like rats) possess the ability to learn new things through positive and negative reinforcement as well.

3. 'Deliberative Thinking': At this level, we already must be able to plan ahead. We have to be able to imagine things and situations, including counterfactuals, and have the ability to decide between alternatives. With these features in place, we can decide between multiple possible ways to reach a goal.

4. 'Reflective Thinking': This is a continuous monitoring of our own activities which allows glimpses into both the past and the present within short-term memory, and even allows us to imagine the immediate future. On this level, we are able to think about our own actions, and decide on better ways to handle things the next time we are in a similar situation.

5. 'Self-Reflective Thinking': Here, not just regular activities, but own thought processes are reflected upon – including own mental states (like confusion).

6. 'Self-Conscious Reflection': The final level adds moral values and enables comparisons between who we are (and what we do) and who we want to be (and what this "who we want to be" person would do).

[263] It could be argued that Dennett does the same mistake when he attempts to simplify phenomenality so much that ineffability is not accounted for - see chapters 3.4.2 and 6.4.1.

[264] Citation from [Minsky(2006)], page 147.

[265] See [Minsky(2006)], pages 129ff.

[266] Notably, this modern and fast-paced time period includes not just a few dozen, but at least the last few thousand years – the rise of civilization accelerated our development in various ways.

[267] This analysis does not necessarily need to be conscious. In the case of how learning and memorization happens, we have no introspective access to the corresponding processes whatsoever.

[268] It is not necessary for us to actually *want* to learn something about a situation, even though if we do, it makes learning more efficient. Learning also seems to come natural for us, we do not even need to think about it and it happens anyway whenver we meet new situations.

Interestingly, this bears many similarities with Freud's distinction between Id, Ego and Superego, with the levels roughly corresponding to these entities.[269]

Minsky's levels of mental activities require differently complex self-models. He does not explicitly list these, but I imagine they are roughly as follows:[270]

1. 'Instinctive Reactions' do not require any form of phenomenal self-model at all. The reactions themselves do not need a phenomenal representation, either.

2. 'Learned Reactions' require a relatively simple form of a self-model. It is enough if there is a self-world distinction that is stringent enough to allow remembering which sensory inputs belonged to the self. Furthermore, if they were initiating motor behaviour, the actions leading to the sensory inputs will consequently have to be learned as such.

3. 'Deliberative Thinking' requires a fairly complex phenomenal self-model which includes knowledge about what the body and mind constituting the self[271] can accomplish, and it also requires the possibility for offline activation of this self-model.

4. 'Reflective Thinking' does not add to the required complexity of the phenomenal self-model, since offline activation was already necessary for the previous level. It does however pose bigger challenges to the equivalent of Metzinger's "dynamicity" constraint,[272] in that different time streams (the one experienced as phenomenal Now,[273] and the counterfactual one experienced as phenomenally inspected time in reflection) have to be represented at the same time.

5. 'Self-Reflective Thinking' adds to the required complexity of the self-model in that more contents of it (namely, mental states, goal representations and some emotional states) have to be available for introspection, thus they have to be part of the (potentially) opaque component[274] of the self-model.

6. 'Self-Conscious Reflection' finally requires that multiple self-models can be inspected at the same time. At least the normative *goal* self-model and the positive *descriptive* self-model have to be available[275] for inspection (which is only introspection for the descriptive self-model) and comparison.

5.4.2 Multiple self-models

Similarly to other disciplines, like psychology or physics, one single model (in this case, a single model of a person) often does not seem adequate. According to Minsky, we certainly have multiple models of other persons in different contexts – their "business self" and their "private self," for

[269] See [Minsky(2006)], page 148 for an illustration of the corresponding levels, and see [Freud(1923)] for Freud's introduction of these concepts. Innate, instinctive urges and drives influence the bottom level of instinctive reactions and thus the id, while values, ideals and taboos influence the self-conscious reflection level and thus the superego. Roughly, the intermediate levels correlate as follows, with Freud's concepts overlapping in some levels of Minsky:

- Id: This includes the levels instinctive reactions, learned reactions, and much of deliberate thinking
- Ego: This includes deliberative thinking, reflective thinking, and self-reflective thinking
- Superego: This includes self-reflective thinking and self-conscious reflection

[270] I will borrow quite some of Metzinger's philosophical terminology here.

[271] The mind, for Minsky, is "a cloud of resources" (see [Minsky(2006)], pages 21ff) and includes all those processes that form emotions and feelings. As such, the mind is, using Metzinger's terminology, everything that supervenes on the brain.

[272] See chapter 3.3.9.

[273] See chapter 3.3.6.

[274] I am using Metzinger's terminology again here, with opaque being the opposite of transparent and thus describing mental contents that are accessible by means of introspection. See chapter 3.3.11.

[275] This implies that both these self-models must be opaque.

example.[276] Minsky suggests that the same applies to our models of ourselves; there is not just one self-model, but rather, there is a multitude of them, activated as a selector for different "Ways to Think"[277] at different times.

> Perhaps our most common self-model begins [...] by representing a person as having two parts – namely, a "body" and a "mind."
>
> That "*body-mind*" division soon grows into a structure that describes more of one's physical features and parts. Similarly, that part called "mind" will divide into a host of parts that try to depict one's various mental abilities.
>
> Each of the models that one makes of oneself will serve only in some situations, so one ends up with different self-portraits in which one has different abilities, values, and social roles. [...]
>
> If you tried to represent all those perspectives at once, your model would soon become too complex to use; in each of those realms we portray ourselves with somewhat different autobiographies, each based on using different aims, ideals, and interpretations of the same ideas and events.[278]

Not only are different aspects of a situation bound to different self-models, but many other selves can be modeled as well (social, athletic, mathematical, musical, political, loving, sexual, professional[279]). All of these can come either in normative or positive variants. Switching between multiple of these subpersonalities (which usually are still relatively close to each other) changes what ways to think are available, because different brain centers are active and contributing to the functional processes generating the current self-model, and thus how we act and think.[280]

All of these variants can come in online and offline versions, but additionally, there are also even more offline variants. Future and past selves can be modeled as well, and fantastic selves can be imagined and appreciated without much effort.

Usually, such subpersonalities can quite easily be described by characteristics or character traits.[281] Minsky shows the following different kinds of dispositions which make it possible to neatly arrange a complex personality into one or only a few categories:[282]

- 'Inborn Characteristics' we were born with (genetic predispositions).

[276] See [Minsky(2006)], pages 301f – and as he states on page 303 ibd., "models that people make of their friends are frequently better than the models that people make of themselves." It could be argued that these multiple models are merely parts of one complex model that includes situational aspects, this would mostly be merely a linguistic difference to Minsky's proposal however. See chapter 5.5.5.

[277] Minsky's Ways to Think, his critics and selector models, are unfortunately subjects we can not look at in this thesis – although they would be very interesting as well. See [Minsky(2007)], pages 215ff, and chapter 5.5.4.

[278] Citation from [Minsky(2006)], pages 304f, emphasis his.

[279] See [Minsky(2006)], pages 306f.

[280] See [Minsky(2006)], page 307. It is interesting to see that newer research in entirely different psychological fields supports this theory: Players of online role playing games underwent an MRI scan while they were describing their character both in the real world and their avatars' character in an online game. According to [Moran and Kelley(2009)] as summarized in [Callaway(2009)], the same brain areas that are responsible for modeling the subject's selves were also active when they were modeling their avatars:

> When Caudle's looked for brain areas that were more active when volunteers thought about themselves and their avatars compared with real and virtual others, two regions stood out: the medial prefrontal cortex and the posterior cingulate cortex. That makes sense as prior research has linked the medial prefrontal cortex to self-reflection and judgement.
>
> Interestingly, however, there was "next to no difference" in the activity in these regions when people thought of themselves and of their avatar, says Caudle, who presented the results at the annual meeting of the Society for Neuroscience.

[281] Minsky makes this example, in [Minsky(2007)], pages 301f: "When Charles thinks about Joan in different realms, his descriptions of her might not all agree. For example, his view of Joan as a person at work is that she is helpful and competent, but tends to undervalue herself; however, in social settings he sees her as selfish and overrating herself."

[282] See [Minsky(2006)], page 310.

- 'Learned Characteristics' that correspond to the goals we try to achieve and priorities we have set, which also influence emotive responses to sensory input.
- 'Investment Principle' – if we learned a way of performing an action that works, it is usually not efficient to learn another way that would also work. This is because the way we know works well, and renders another learning period unnecessary or at least inefficient.[283]
- 'Archetypes and Self-Ideals', essentially our normative self-models, that shape positive ones by means of the higher levels of mental activities (those that correspond to the Freudian Super-Ego).
- 'Self-Control', constraining short-term urges for longer-term goals. This serves both being predictable for others, but also becoming self-predictable and thus being able to "depend on yourself."[284] That makes life easier for us, or, in Minsky's words: "It saves a great deal of effort and time to see people or things as stereotypes."[285]

5.4.3 Personal identity

The idea of a self, of being an individual, is certainly a stimulating and treasured one. This has evolutionary origins, or at least is perfectly compatible with our biological constraints: We are constrained to one single, localized body. We have a private mind that nobody else can easily see into. We claim moral responsibility for deeds we do and force others to assume responsibility for theirs, and construct causal attributions from the position of attributing one self to each body.

Social relations without selves would be awkward if we would not imagine our peers as being selves,[286] and it would be much more difficult to maintain attention and focus if we would not have the illusion of experiencing a single, continuous stream of consciousness.[287] But how do we construct such a personal identity, all convenience of having it (and thus teleological explanations) aside?

Minsky closely follows Dennett's argumentation regarding the center of narrative gravity[288] on this point, keeping in mind that we regularly fail to fully appreciate early childhood memories[289]:

> We should ask ourselves what compels us to think of ourselves as Selves – and here is a simplistic theory of this: whatever happens, we're prone to ask ourselves who or what was responsible – because our representations force us to fill the "caused-by" slots [that go with the way we represent all memories]. [...]
>
> However, when you fail to find a plausible cause, that slot-filling hunger may lead you to imagine a cause that doesn't exist – such as the "I" in *"I just got a good idea."* For if your frame-default machinery compels you to find a single cause for everything that you ever do – then that entity needs a name. You call it "me." I call it "you."[290]

[283] Despite the tempting view to see this as a disposition, Minsky sees it as a principle – our brain just works like that, it is not a decision we make, even though we can make the decision to try and override the principle. He introduces the concept in [Minsky(2006)], page 310:

> Once we learn an effective way to do some job, we'll resist learning other ways to do it – because new methods are usually harder to use until we become proficient at them. So as our older procedures gain strength, it gets harder for new ones to compete with them.

[284] Citation from [Minsky(2006)], page 312.
[285] Citation from [Minsky(2006)], page 310.
[286] They would probably not be social relations as such at all, because social relations require persons, autonomous members of a society – and without selves, there can be no such persons. Minsky does not explore this tangent when he only writes in [Minsky(2006)], page 321:

> Other people expect us to think of them as Single Selves, so unless we adopt a similar view, it will be hard to communicate with them.

[287] See [Minsky(2006)], pages 320f.
[288] See chapter 5.2.
[289] "ee chapter 4.4.1.
[290] Citation from [Minsky(2006)], page 309.

5.5 Consolidation

I will first, in chapter 5.5.1, have a look at why dualism seems to be a less desirable option from a scientific point of view. The other three views presented in this chapter might differ considerably in details, but they agree on one central thing: The self is an entity that emerges from mental processes and states, something that locally supervenes[291] on brain states and neural correlates.

Next, various problems which the theories of Dennett, Metzinger and Minsky still have will be explored. Finally, an attempt to consolidate all three theories into one common view that incorporates all their strengths, while trying to avoid their weaknesses, is presented in chapter 5.5.4.

5.5.1 On substance dualism

All the theories mentioned in chapter 5.1 have in common that they postulate that there is more to the world than the physical, and more importantly, they postulate that this additional matter is not traceable or detectable by scientific or naturalistic means. Essentially, they postulate that there is a kind of causation that is not causally analyzable. Looking at the original Cartesian version that puts the *res cogitans*, the "mind stuff", into the pineal gland, Dennett puts it concisely:

> How, precisely, does the information get transmitted from pineal gland to mind? [...] let's ignore those upbound signals for the time being, and concentrate on the return signals, the directives from mind to brain. These, *ex hypothesi*, are not physical [...]. No physical energy or mass is associated with them. How, then, do they get to make a difference to what happens in the brain cells they must affect, if the mind is to have any influence over the body? A fundamental principle of physics is that any change in the trajectory of any physical entity is an acceleration requiring the expenditure of energy, and where is this energy to come from?[292]

It may be true that the world consists of more than what is epistemologically graspable by us human beings. But that does not make this additional "matter" non-physical or even non-naturalist. In particular, it is implied by definition that every causal relationship is causally analyzable. Causes for any kind of effect, whether we can touch and see the effect or not, can be logically scrutinised. As our insights into mathematics or particle physics show, we are perfectly able to perform these kinds of investigations even with empirically nongraspable, completely abstract subjects.[293]

Applied to philosophy of mind, this means the following: There is nothing that stops us from investigating logical interactions not available for introspection. If the Cartesian *res cogitans* does however have causally analyzable interactions with the *res extensa*, it is not a fundamentally different, non-physical kind of matter. Assuming that it is amounts to a category error.

Consequently, I will dismiss substance dualism as not relevant for this thesis.[294]

5.5.2 Criticising Dennett

Even though Dennett's theory on what constitutes a self is fairly thin, it is compatible with Metzinger's, which is much more extensive.[295] Dennett's theory of the self as center of narrative

[291] See chapter 3.3.4.

[292] Citation from [Dennett(1991)], page 34, italics his.

[293] For example, there are major works on the epistemology of numbers and mathematical structures, starting with [Hilbert(1897)].

[294] Conceptual dualism on the other hand is arguably at the core of it. See chapter 8.2.

[295] Metzinger's theory is more extensive for two reasons. First, the PSM (see chapter 5.3.2) is one of the two central building blocks of Metzinger's theory and plays a core role in his explorations overall, compared to Dennett who merely created a chapter on the self being the center of narrative gravity (see chapter 5.2) that seems more like an afterthought. Second, Metzinger also explores tangents and implications, while Dennett's deliberations on the subject do not go any further than that, because he does not attempt to explain how a self comes to be or whether it has ontological legitimation in an abstract space, but why consciousness as a whole (and thus also selves) can be "explained away", as Dennett himself explains and justifies in [Dennett(1991)], page 455:

> Only a theory that explained conscious events in terms of unconscious events could explain conscious-

gravity also implies that the self is merely a model, he calls it a fiction, that the system produces in order to make more sense of the following fact: There is a physically localized and temporally invariant structure that the phenomenal experience is attributed to as belonging to.

This compatibility between Metzinger and Dennett is a good thing, because both theories seem to lack major fundamental flaws. However, Dennett makes a comparison that only holds up to logical scrutiny with some limitations:

> If what you are is that organization of information that has structured your body's control system (or, to put it in its more usual provocative form, if what you are is the program that runs on your brain's computer), then you could in principle survive the death of your body as intact as a program can survive the destruction of the computer on which it was created and first run.[296]

Regarding this analogy however, it must be considered that a program cannot run on just any computer, at least not without an emulator running first: A program that was created for a certain operating system, running on certain hardware, will not be able to run on another basis, unless they are explicitly made to be compatible through additional programs.

Applied to the analogy of consciousness, this means: A self relies on an enormous number of representative and presentative processes supplying it with data about both the environment and the inner state of the body that the self belongs to. The self also is "programmed" in a many different parallel processes with corresponding aligning (neural) pathways – pathways that themselves shape the processes originating from the sensory organs through which data flows.

Splitting the self from the body is thus not easily possible, as the body constitutes the self's current contents through the topology and states of its neurons, and certainly formed the self in the past. A computer is radically different from a brain, so in order to emulate a human brain on one of today's computers, we would require an emulation so complex that it would presuppose more work than the actual 'brain software' transfer itself. The self supervenes on neuronal topology and states, a computer program on the other hand supervenes on zeroes and ones – and translating one to the other is not trivial.

Moreover, if the "software" were to change during execution time, the hardware has to change as well, because the hardware (neurons and their connections to other neurons) defines the software's contents – there are apparently around 400 specialized and quite static (on a large scale, but changing on a small scale) neural areas,[297] all of which contribute to what ultimately constitutes the self.

Thus, the hardware-software distinction does not quite work when it comes to human selves, even though Dennett is fairly convinced otherwise.[298]

This stays true unless somebody takes the time and effort to identify and translate every possible low-level function call into one that works on the hardware to which the program should be transferred. Of course this is possible in principle: Today, it is possible to run 'ancient' Commodore 64 programs[299] on an emulator running on current-day hardware and operating systems, but the more different those hardware bases get, the more work has to be done in order to produce an emulation of the original hardware.

Consequently, when Dennett argues that *in principle*, it could be possible to survive the death of the body, I agree – however, saying that the body's control system is merely the software that runs on the brain's computer is inadequate, because the brain itself is not only the "hardware," but also an important constituent of that "software."

---

ness at all. If your model of how pain is a product of brain activity still has a box in it labeled "pain," you haven't yet begun to explain what pain is, and if your model of consciousness carries along nicely until the magic moment when you have to say "then a miracle occurs" you haven't begun to explain what consciousness is.

[296] Citation from [Dennett(1991)], page 430.
[297] See [Minsky(2007)] around 1:14:36.
[298] See [Dennett(1991)], pages 210ff.
[299] The Commodore 64 is of course only ancient by information technology standards.

5.5.3 Metzinger's prereflexive proto self-model

There is anecdotical evidence that makes Metzinger's prereflexive proto self-model[300] intuitively plausible. For example, some people experience phantom limb pains yet never had a the limbs in question in the first place. There seems to be some hardwired (genetically hardcoded and necessarily grown) structures in our brains that correspond to certain body parts.[301]

However, the functional components that may form such a self-model are something that warrants more research, because, as we have seen,[302] once there is phenomenal experience, and a pre-reflexive proto self-model, the rest that forms our phenomenal self-model (or, with Minsky, multiple self-models – see chapter 5.5.4) comes naturally and by logical necessity.

That research would do well to answer the following questions – assuming Metzinger's teleofunctionalism[303]:

- What types of teleological properties are necessary for the formation of a human-like phenomenal self-model, given the ten constraints[304] besides perspectivalness and a phenomenally experienced human body? In other words, how exactly do we have to define the lower levels of the perspectivalness constraint to make it both necessary and sufficient for the generation of a phenomenal self-model in a phenomenally experienced process in a human brain?

- What are the functional properties that enable those teleological properties? In other words, how exactly do we have to define the functional structures that enable our now found exact definition of the perspectivalness constraint?

- What kinds of neural correlates (and functional links of the mental contents supervening on them) have to exist in order to create those functional properties, and how can they form during a human organism's growth process? In other words, how is the perspectivalness constraint implemented in the human brain?

To date, we have made little progress with regard to these questions. There are hints and pointers that make the idea of a prereflexive proto self-model seem reasonable. It appears that the proto self-model is capable of explaining many things that remain mysteries in other theories and belief systems.

We know for example that there are brain regions that invariably are responsible for a certain body part's sensory inputs. We also know which brain regions are activated when subjects are modeling selves. However, there are no fleshed-out theories yet that fit these discoveries together. I believe that this can be a worthy topic for future research.

5.5.4 Multiple phenomenal self-models

Most parts of Dennett's view appear to hold up to logical scrutiny. The same thing applies to almost all parts of Metzinger's view, and almost all parts of Minsky's view. Considering that they contradict each other only in minor points, this can only mean one thing: I propose that we blend them together and form a theory that incorporates them all. The following discussion can be taken as my attempt at doing precisely that.

Since Dennett's ideas on subjectivity are not as detailed as, and can mostly be expressed by a subset of, Metzinger's ideas, I think we can for now drop Dennett. This leaves us with

[300] The proto self-model is a special form of inner acquaintance with ourselves that provides the basis for a PSM and thus must have existed before a PSM could form – see chapters 3.3.10, 3.4.4 and 5.3.2

[301] Metzinger has an extensive section concerning phantom limbs in [Metzinger(2003)], pages 461-488. In particular, he writes on page 478, referring to physicians and neurologists who researched phantom limbs: "In his discussion Poeck agrees with Sidney Weinstein and Eugene Sersen, who in 1961 published a substantial paper containing five case studies describing phantom limb experiences in children with *congenital* absence of limbs, that is, phantoms for a limb which had never *existed*, that the assumption of a 'built-in' component of the conscious body image has to be made."

[302] See chapter 5.3.2.

[303] See chapter 3.4.6.

[304] See chapter 3.3.

Metzinger's generation of phenomenal self-models, with the reservation that not one such self-model is generated during the lifetime of a system. Rather, from one moment to the next, different "critics" and "selectors"[305] are active. Thus different parts of the brain provide functional processing abilities, leading to at least some lower-level parts of the phenomenal self-model supervening on different neural correlates from one moment to the next, and thus a different positive phenomenal self-model is transparently[306] experienced.

Modeling *normative* selves on the other hand always happens in an opaque manner, and they (just like other kinds of imagined offline self-models) are available for introspection. However, the fact that our positive phenomenal self-model changes from one moment to the next is not necessarily available to us, since our currently active positive self-model is transparent, and some parts of it are even ineffable.[307] We can, as with other things, introspectively distance ourselves from our deeds and look at what we are doing; as exemplified in when we take a step back and realize we are just acting so rudely because we are jealous when we really should not be.

By default however, we do not reflect like that, and often do not realize that we are indeed not the same today as we were yesterday.[308] The common saying "I'm not quite myself today" obtains an entirely different and way deeper meaning with this theoretical background.

These multiple self-models can, to some extent, be stored transtemporally. This is always true for a description of an opaquely experienced phenomenal self-model. However, there is a reservation: For some of these stored self-models, it is no longer possible to fully appreciate them in the sense that I constrained appreciation in chapter 4.4.1. Some childhood memories might still be available, but while I know that I am still the very same person, I do not know anymore what it was like to be me back then.

5.5.5 Multiple aspects of one self-model?

According to the theory we now reached,[309] a person models multiple selves, normative and descriptive ones, and has the capability to store aspects of them transtemporally. This person then always has one active and transparent online self-model,[310] and can have several opaque offline self-models available for introspection at the same time. All these entities supervene locally[311] on brain structures.

Despite the apparent simplicity of this theory, it does not appear to be fully adequate for some situations we regularly find ourselves in, and all the different kinds of phenomenal self-models we possess:

- There will be overlapping areas where multiple phenomenal self-models share similar or

[305] As I mentioned before (see chapter 5.4.2), unfortunately we do not have the space to look at Minsky's "critics" and "selectors" – essentially, they are brain centers responsible for activating different other brain centers that then process sensory data and other mental contents in different ways. This then results in different "Ways to Think" being available to the system. See [Minsky(2006)], where Minsky's "Critic-Selector Machines" are introduced, first on page 29:

> So this book will suggest that to deal with hard problems, our brains augmented their ancient Reaction-Machines with what we'll call *"Critic-Selector Machines."* [...] The "Critics" of Critic-Selector Machines will also detect situations or problems inside the mind such as serious conflicts between active resources. Similarly, the "Selectors" of Critic-Selector machines don't just perform actions in the external world, they can react to *mental* obstacles by turning other resources on or off – thus switching to different Ways to Think.

[306] See chapter 3.3.11.

[307] See chapter 3.3.17.

[308] Nevertheless, despite having a different *active* phenomenal self-model, we are of course still the same person, and thus the multiple self-models can well be perceived as being merely aspects of one all-encompassing self. See chapter 5.5.5.

[309] See chapter 5.5.4.

[310] This statement is greatly simplified: In rare cases, several self-models can be transparently active, for example when we are switching from one activity to the next and are reorienting ourselves. Furthermore, in pathological cases, several such self-models can be active concurrently, sharing the same body. Finally, it is not necessary to have a self online at all times, as chapter 5.3.4 showed.

[311] See chapter 3.3.4.

identical traits. It is possible that a person acts completely different in varying surroundings, but that does not seem to happen all the time; we recognize our friends whether they are playing sports or studying.

- The various phenomenal self-models are not pure. The business self will mix with the parent self when a businesswoman is talking with her kids on the phone during her lunch break. In such a situation, it is not so much the case that two phenomenal self-models are active concurrently, but that the resources required by multiple phenomenal self-models are simultaneously active in our brains and thus form a new, *merged* phenomenal self-model.
- While there may be multiple phenomenal self-models, they are still part of the same conceptual self, and except in pathological cases will also be judged as being the same person by the social environment.

Consequently, what I called "multiple phenomenal self-models" could really be said to be merely aspects of *one* complete phenomenal self-model, not all of which are active at all times. The difference between these two ways of naming the theory, however, appears to be of merely linguistical nature.

6 Intentionality

Importantly, "intentionality" is one of those "suitcase words"[312] that can have many different meanings depending on their context. What I intend to analyse in the following is a rather specific meaning of the word: The thesis will narrow down the meaning of "intentionality" to the peculiar way in which our mind can focus on a concept, or a group of concepts perceived as one entity, and then seemingly forms a direct connection between us and this target entity – at least, this is how we often phenomenally experience it.[313]

Intentionality can be imagined, as a probably naive but intuitively graspable model, as the arrow of attention, pointing from inside our heads to an object in the world (or at an imagined object inside our heads), while carrying a meaning under a certain aspect.[314]

6.1 Brentano: Immanent objectivity

Without naming it "intentionality" like more recent philosophers do, Brentano already introduced and researched the concept:

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as an object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.[315]

The idea is very intriguing. It certainly seems that there always is some kind of object component of any mental phenomenon, something the mental phenomenon is about. There are suitcase words[316] which can describe an entire class of entities. Consequently, even if we can always linguistically refer to different kinds of intentionality as entities belonging to one class, that does not necessarily make them the same from an ontological point of view.

Nevertheless: The fact that there is something unifying about mental phenomena, in that they all share having some kind of 'aboutness', has to be accounted for. Brentano rightly calls attention to that fact.[317]

---

[312] "Suitcase words" are a concept that Minsky introduces. Essentially, an entire suitcase of concepts fits onto one word, which we learned to know without thinking about it; dissecting that word is hardly possible because the concepts the word implies do not even necessarily have to belong to the same categories. He talks about them in various places in [Minsky(2006)], for example asking on page 12, after having introduced plenty of meanings of the word "love": "Why do we pack such dissimilar things into those suitcase-like words?"

Later on, he looks at the concept from different angles, and offers a psychological explanation in [Minsky(2006)], page 110: "We often use those suitcase words to keep us from asking questions about ourselves. Just having a name for an answer can make us feel as though we actually have the answer itself."

[313] I believe that it is a slight bit of a philosopher's delusion to presume that everybody thinks about it with that intentional arrow pointing out of our heads. When talking with fellow computer scientists, I often find it hard to persuade them that it is not merely a case of 'wandering focus.' I believe that both descriptions hold merit, but will stay true to the philosophical nature of this thesis and use the intentional experience of *directedness* wherever possible.

[314] As Searle put it in [Searle(1992)], "All intentionality is aspectual."

[315] Citation from [Brentano(1995 (original 1874))], page 88, their translation. The original quote goes: "Jedes psychische Phänomen ist durch das charakterisiert, was die Scholastiker des Mittelalters die intentionale (auch wohl mentale) Inexistenz eines Gegenstandes genannt haben, und was wir, obwohl mit nicht ganz unzweideutigen Ausdrücken, die Beziehung auf einen Inhalt, die Richtung auf ein Objekt (worunter / hier nicht eine Realität zu verstehen ist), oder die immanente Gegenständlichkeit nennen würden. Jedes enthält etwas als Objekt in sich, obwohl nicht jedes in gleicher Weise. In der Vorstellung ist etwas vorgestellt, in dem Urteile ist etwas anerkannt oder verworfen, in der Liebe geliebt, in dem Hasse gehasst, in dem Begehren begehrt usw." Oskar Kraus refers to the same immanent objectivity in a footnote on page 89, saying that "it is not to be interpreted as a mode of being the thing has in consciousness, but as an imprecise description of the fact that I have something (a thing, a real entity, substance) as an object, am mentally concerned with it, refer to it."

[316] I explained suitcase words, Minsky's concept, in a footnote in chapter 6.

[317] See [Chrudzimski(2005)], which goes beyond Brentano and also includes Husserl and Ingarden.

Brentano already classifies mental phenomena into three kinds: presentations, judgements, and those "of love and hate" – by which he means any kind of emotional phenomena. He is well aware that there can be other classifications, too, and he explicitly mentions those of Aristotle (thought and appetite) and the apparently prevalent classification of his time (presentation, feeling and will).[318]

6.2 Dennett: Only derived intentionality

Dennett does not believe in intrinsic or original intentionality. The essence of the thought experiment that he starts his analysis with reads as follows:

> Suppose some human being, Jones, looks out the window and thereupon goes into the state of thinking he sees a horse. [...] Suppose the planet Twin Earth were just like Earth, save for having schmorses where we have horses. (Schmorses [...] are well-nigh indistinguishable from horses by all but trained biologists with special apparatus, but they aren't horses [...].) If we whisk Jones off to Twin Earth, land of the schmorses, and confront him in the relevant way with a schmorse, then either he really is, still, provoked into the state of believing he sees a horse (a mistaken, nonveridical belief) or he is provoked by that schmorse into believing, for the first time (and veridically), that he is seeing a schmorse. [...] However hard it may be to determine exactly which state he is in, he is really in one or the other [...]. Anyone who finds this intuition irresistible believes in original intentionality [...]. Anyone who finds this intuition dubious if not downright dismissible can join me [...] in the other corner [...].[319]

This thought experiment is interesting because it challenges our intuitions. Schmorses clearly are distinct from horses, from an intrinsic point of view, while their external appearance certainly seems to be the same.

If there is such a thing as an intentional arrow (with ontological legitimation beyond that of a helpful metaphor), it thus clearly has to point at either a schmorse or a horse, because it has to point at *something* or (since it is defined through a start point and an end point) it can not exist at all. If an intentional arrow is only a helpful metaphor however, it does not matter what it really points at, because it does not really point at anything at all in a literal sense, and consequently the question whether it actually points at a horse or a schmorse makes no sense.

Dennett goes even further: Not only does he doubt that intentionality can be intrinsically about something in particular, he also does not believe that there is anything beyond derived intentionality.[320] That is what his argument for the intentional stance is all about.

6.2.1 The intentional stance

It is striking how often humans ascribe intentional acts and beliefs to less developed animals and even machines, both of which obviously do not have any feelings or even states that could properly be classified as "mental" (or rather, phenomenal) states. We also ascribe intentional acts and beliefs to each other, with apparently more justification and less metaphorically.

This ascription of intentionality to other systems is, according to Dennett, the intentional stance. It is a stance we enter in which we ascribe intentionality to other systems, or even ourselves. What makes us ascribe those intentional acts and beliefs to each other, however? Dennett makes out three principles:

1. A system's beliefs are those it *ought to have*, given its perceptual capacities, its epistemic needs, and its biography.

[318] See [Brentano(1995 (original 1874))], pages 194ff.

[319] Citation from [Dennett(1987)], pages 294f. There, he also lists many philosophers who either agree that there is intrinsic intentionality, or agree that there is not. Pardon all the ellipsis, I gave my best not to leave out anything important for Dennett's extensive introduction of the thought experiment.

[320] What Dennett and others call "derived intentionality" has been called "observer-relative ascriptions of intentionality" by Searle in his Chinese Room discussion (see [Searle(1980)]).

2. A system's desires are those it *ought to have*, given its biological needs and the most practicable means of satisfying them.

3. A system's behavior will consist of those acts that *it would be rational* for an agent with those beliefs and desires to perform.[321]

Thus, essentially, evolutionary fitness and rationality are the sole criteria for reasonable ascription of intentionality, and ascription of intentionality is all there is to true intentionality. In other words, according to Dennett, there is no intrinsic intentionality. I will discuss this view in chapter 6.4.1.

Dennett proposes that whenever a system acts in a rational way according to constraints set upon it by the environment (namely, what Dennett calls its *"raison d'être"*[322] and its biological needs), it possesses all the intentionality it can possibly possess.

Strictly speaking, this definition also means that only biological systems (but no artificial or postbiotic systems) can possess any form of intentionality, as they strictly need to satisfy *biological* needs. However, and I believe we would not run counter to Dennett's intentions with this project, I think we can safely replace those "biological needs" with a similar concept, although I find it hard to express that similar concept in such a distinct wording. It would have to contain both extrinsic and intrinsic requirements for survival, both short- and long-term. In systems that eventually stop working, such as death-ridden biological systems, this also includes reproduction.

6.2.2 Further notions of intentionality

It is important to note that Dennett goes beyond my narrowed-down definition of intentionality.[323] He often talks about consciousness and mental states in general, and intentionality as "possession of mental states." In particular, he is concerned with the following:

- Dennett means to disprove that there is a meaning intrinsically attached to such individual arrows of attention we direct at the world.[324]
- Dennett looks at higher-order mental states.[325]
- Dennett even examines the "language of thought" in the bigger context of aforementioned mental states.[326]

In no way do I want to disqualify or doubt the importance of those examinations by not looking at them here, yet I merely want to analyze intentionality in a rather narrow sense. I believe that if we want to understand consciousness, it makes sense to explore small puzzle pieces individually. And although of course we cannot afford to lose sight of the big picture, I will postpone looking at consciousness as a whole until chapter 7.

6.2.3 No intrinsic intentionality

As has been shown, Dennett does not believe that there is any such thing like original or intrinsic intentionality. Similar to Searle not finding anything that could give his Chinese Room's homunculus any intrinsic intentionality he could pass on to the system he is in, Dennett does not find any

[321] Citation from [Dennett(1987)], page 49 – all three points are quotes, emphasis his. As he clarifies ibid., "In (1) and (2) 'ought to have' means 'would have if it were ideally ensconced in its environmental niche."'

[322] For example in [Dennett(1987)], page 298.

[323] See chapter 6.

[324] See chapter 6.2.3.

[325] See [Dennett(1987)], pages 242ff.

[326] In particular, he points out (in [Dennett(1987)], pages 230f) how said language of thought would require connectionist networks: "There must indeed be a higher level of description at which we can attribute external-semantical properties to relatively global features of the network's activities, but at such a level the interations and relationships between semantic elements are not computational but [...] statistical, emergent, holistic. [...] In Connectionist models, the (typically simulated) hardware does add something: just which content-relative effects actually occur (something that is only statistically describable at the high-level) depends on low-level features of the history of operations. The different flavors of cognition emerge from the activity, without being specifically designed to emerge."

part of the human mind that could give it intrinsic intentionality. Referring to Dawkins,[327] he describes what we humans really are, from an evolutionary point of view.

Dennett designs a thought experiment in which we would want to experience the year 2401, and then build a robot that could contain and protect our hibernating body until then. Since the world is an inherently dynamic place, merely sitting always in the same spot probably could not ensure our survival. Consequently, that robot would have to be able to classify things[328] which are good or bad for its continued functionality, and it would also have to be able to classify and filter those things which do not affect the system.

The system would require some forms of autonomous self-control with goals and subgoals, would have need for "quick and dirty" approximations[329] in its calculative appraoches to problem solving. It would thus show equivalents of our mental states of wondering, seeing, deciding; it would essentially exhibit a sophisticated form of derived intentionality, similar to what a machine passing the Turing Test would have to show.

With this analyzed and the robot thus designed, Dennett points out:

> If we cling to this view [that this artifact would only possess derived intentionality], the conclusion forced upon us is that our own intentionality is exactly like that of the robot, for the science-fiction tale I have told is not new; it is just a variation on Dawkins' vision of us (and all other biological species) as "survival machines" designed to prolong the futures of our selfish genes. We are artifacts, in effect, designed over the eons as survival machines for genes that cannot act swiftly and informedly in their own interests. [...] So our intentionality is derived from the intentionality of our "selfish" genes! *They* are the Unmeant Meaners, not us![330]

This indeed raises the question: Assuming we do not believe in magic or dualism, how does the derived intentionality (which we certainly have) become intrinsic intentionality? Why do we believe that we are so special?

Dennett certainly seems to be right in this regard. Intrinsic intentionality, as intuitively plausible as it may be, does not seem to have any justification that goes beyond mere intuition.

6.3 Metzinger: Just a phenomenal model (PMIR)

PMIR stands for "phenomenal model of the intentionality relation", and it is the higher-level one of the two main building blocks of Metzinger's theory of consciousness.[331] It is also part of the "perspectivalness" constraint,[332] just like the prereflexive proto-self and the PSM are. Metzinger considers the perspectivalness constraint as one of the three defining factors of subjective consciousness:

> The concept of a PMIR will, finally, give us a more precise understanding of *constraint 6*, the perspectivalness constraint for conscious processing [...]. Together with the idea of a transparent global model of the world, as activated within a window of presence, this third major theoretical entity will [...] allow us to offer a more informative version of the minimal concept of *subjective* consciousness.[333]

Similarly to the PSM (see chapter 5.3), the PMIR is merely a phenomenal model, merely an emergent, transparent, phenomenal structure. It is a conscious mental model, presented phenomenally in a transparent manner and thus experienced as unquestionably real.

The content of that model is an ongoing, episodic subject-object relation, representing the system as subject through its PSM, and the object in an asymmetrical relation to that subject:

[327] See [Dawkins(1976)].
[328] 'Things' here denotes events, objects, hazards, etc. – all the entities in the system's environment that it is designed to be able to make out and discern.
[329] See [Dennett(1987)], page 297.
[330] Citation from [Dennett(1987)], page 298, emphasis his.
[331] The first building block of Metzinger's theory is the PSM, see chapter 5.3.2.
[332] See chapter 3.3.10.
[333] Citation from [Metzinger(2003)], page 411.

> What is the phenomenal model of the intentionality relation? It is a conscious mental model, and its content is an ongoing, episodic *subject-object relation.*[334]

6.3.1 Kinds of perceived intentional relations

This definition of the PMIR as a mere model of an arrow of attention applies to a wide variety of cases, and "the number of possible object components is almost infinitely large."[335] These possible object components usually fall into one of the following categories:[336]

- A perceptually deliberately attended object, given through sensory input.
- A consciously deliberately attended opaque kind of conscious content, like a cognitive self-model.
- An object forcibly perceptually attended, given through sensory input.[337]
- A complex mental motor self-simulation, usually prior to "embodying"[338] it by executing those motor behaviours.

Of course, all those categories can be represented with different kinds of relations.[339] These relations include attending, thinking, desiring,[340] fearing, and many others. They also have a direction – the "arrow of attention" points outwards (in regular perception) or inwards (in introspection), from the first-person perspective, but always originates from the subject component, thus putting subject and object into the aforementioned asymmetrical relationship.

6.3.2 Phenomenalizing intentionality

Metzinger wants to naturalize intentionality, and phenomenalizing it is the first necessary step for this naturalization. He explains:

> Phenomenalizing intentionality, I would submit, may be a necessary detour, an indispensable first step in the project of *naturalizing* intentionality *tout court*. Meaning and the conscious experience of meaningfulness have to be separated. Generally speaking, mental representations possess two kinds of content: phenomenal content and intentional content. Phenomenal content supervenes locally. Intentional content, in many cases, is determined by external and nonlocal factors. [...] It is important to note how intentionality [...] is *itself* depicted on the level of phenomenal content.[341]

Parts of his statement are obvious: Meaning and the conscious experience of meaningfulness (which corresponds to intentionality) are not identical, although they are phenomenally experienced as being intuitively and unquestionably the same. Metzinger argues that while we do know that phenomenal content supervenes locally,[342] we do not necessarily know that intentional content, which refers to external entities, also supervenes locally. If we can show that intentional content is just a special form of phenomenal content, it goes without saying that intentional content supervenes locally and thus potentially can be naturalized.

Obviously, showing that intentional content is merely a special form of phenomenal content is the chief purpose of the PMIR theory. Since according to Metzinger, the phenomenal experience

[334] Citation from [Metzinger(2003)], page 411.
[335] Citation from [Metzinger(2003)], page 411.
[336] See [Metzinger(2003)], pages 411f.
[337] The object in question being forcibly attended means that the subject is "finding yourself forced to automatically attend," as [Metzinger(2003)] specifies on page 412.
[338] Metzinger uses this word as an active verb to express that the self-simulation is executed online, coupled to effectors, and not only offline. The simulation is thus also (in addition to being an offline simulation) made real in the actual body it is simulated for. See [Metzinger(2003)], page 412.
[339] These relations would be called "attitude specificator" in propositional attitude psychology.
[340] Metzinger uses the word "willing" in [Metzinger(2003)], page 412.
[341] Citation from [Metzinger(2003)], page 414, emphasis his.
[342] See chapter 3.3.4.

of an intentionality relation is all there is to intentionality, phenomenal experience could exist completely without ontological counterparts to the entities represented by said intentionality relation's phenomenal contents. Intentionality could be entirely virtual, and if it were, it would not be possible for us to notice. In fact, it is quite possible that we misrepresent many things out there in the world, because this misrepresentation has proved to be more beneficial for the continued survival of our species:

> This global effect [of "global immersion"] is achieved by continuously activating dynamic and transparent representations of a subject-object relation, which episodically integrates the self-model and those perceptual, cognitive or volitional objects, which cause the changes in its content, by telling an internal story about how these changes came about. This story does not have to be a true story. It may well be a greatly simplified confabulation, which has proved to be functionally adequate.[343]

6.3.3 PSM without PMIR?

There are pathological cases of humans having damaged brains that are still able to construct a PSM, but no PMIR anymore.[344] They can walk and sit and look, but lack intention or motive in whatever they do. They do not have a conscious representation of the arrow of intentionality, not even towards themselves.

These patients are still embodied selves, but not agents: They lack goals they could pursue, things they could want, and do not have conscious representations of motor simulations anymore either. As Metzinger points out,

> [Such tragic cases] very clearly demonstrate what it means to say that the phenomenal first-person perspective [including the PMIR] is the decisive factor in turning a mere biological organism into an agent, into a willing subject.[345]

This is mainly because phenomenal volition is a form of phenomenal intentionality, and thus not possible without a PMIR. Without volition, even goal-driven movement cannot be experienced as agency by the would-be agent, as the alien hand syndrome (where patients experience that usually one of their hands moves on its own, without the patient able to intervene or guide it) demonstrates:

> The central point is that many such arm movements clearly seem to be goal-directed actions, although no such goal representation is available either on the phenomenal level in general or on the level of conscious self-representation. The underlying goal representations are not phenomenally owned, and therefore are not *functionally appropriated*.[346]

So while it is possible to have a PSM without a PMIR, a complete phenomenal first-person perspective requires both.[347]

6.4 Consolidation

It seems hard to be talking about intentionality while avoiding the matter of consciousness or subjective experience in general. The two concepts are very closly related – or rather, they are interwoven. Brentano already pointed this out when he revived the concept of intentional inexistence. Intentionality is the prime precondition for consciousness, because all conscious mental processes are also directed at something (directed at an intentional object) and originate from an intentional subject. Metzinger puts it as follows:

[343] Citation from [Metzinger(2003)], page 416.
[344] See [Metzinger(2003)], pages 416f, where Metzinger also refers to case studies in [Damasio(1999)].
[345] Citation from [Metzinger(2003)], page 419.
[346] Citation from [Metzinger(2003)], page 425, emphasis his.
[347] This will become clearer once we look at the bigger picture in chapter 7.

> We can now see how a full-blown subjective consciousness evolves through three major levels: The generation of a world-model, the generation of a self-model, and the transient integration of certain aspects of the world-model *with* the self-model. What follows is a minimal working concept of subjective experience: Phenomenally subjective experience consists in transparently modeling the intentionality relation within a global, coherent model of the world embedded in a virtual window of presence.[348]

Interestingly, this functional coupling of the intentionality relation with consciousness prompted many philosophers (even Dennett, as we saw in chapter 6.2.2) to look at intentionality mainly in the bigger picture as well. Metzinger's analysis of the intentionality relation is among the most thorough I found, and the only one that distinctly looks at the intentionality relation itself in greater detail and without being distracted by implications for consciousness; even misnomers, using "intentionality" synonymously to "consciousness," seem to be common.[349]

With intentionality sitting at the core of consciousness, so much so that it is often used as the defining characteristic of or even synonymously to consciousness, it is very important that we have a clear picture of what it actually is. Metzinger, and to a lesser extent Dennett, offered us a possibility to naturalize intentionality without losing the information which an intrinsic intentionality could provide, while at the same time allowing us better insight into how it is produced and what its effects on the physical brain are. This certainly seems to be a worthy approach to take.[350]

6.4.1 Dennett's intentionality should be phenomenal

I am talking about Dennett's argumentation that intentionality is nothing but an ascription that happens in a "stance" we assume when looking at entities that might exhibit inentionality.[351] We remember, he argues that ultimately, whatever intentional states we ascribe to a target system have to be what the system's intentionality truly is.

The reason why this appears to be inadequate is two-fold:

---

[348] Citation from [Metzinger(2003)], page 427, emphasis his.

[349] For example, when Dennett goes back to Brentano when undertaking a similar project to my explanations in [Gloor(2007)], showing that the brain is merely a special kind of Turing Machine. In [Dennett(1987)], page 67, Dennett writes:

> Consider that warhorse in the philosophy of mind, Brentano's Thesis that intentionality is the mark of the mental: all mental phenomena exhibit intentionality and no physical phenomenal exhibit intentionality. [...] But given the concept of an intentional system, we can construe the first half of Brentano's Thesis – all mental phenomena are intentional – as a *reductionist* thesis of sorts, parallel to Church's Thesis in the foundations of mathematics.

Note how "intentionality" stands for "consciousness" here. Admittedly, I mixed up the two for the sake of better short-term clarity in chapter 5.1, too.

[350] Whether naturalizing consciousness is the only approach that makes sense is of course to some extent a matter of opinion, but it seems to me that reductive naturalism (i.e. everything is reducible to purely physically analyzable components) is the most reasonable and logical approach available.

This does not imply that I believe that our current scientific theories (or theoretical frameworks, or fundamental scientific entities) contain the answers to all the questions, or even that our current scientific paradigms need to be the ones that prevail. But arguably, scientific theories could in theory be formed that contain all the answers, using rational argumentations as their foundation. Some basic principles of science (mostly, conclusive logic, causal analysis and semantical coherence) will necessarily lead us to a better understanding of the world that we have only just just started to analyze.

Sidestepping these basic principles and dogmatically defining some mysteries as insolvable on the other hand (and even ignoring evidence that hints at a solution, particularly when that solution conflicts the dogmatic 'truths') will not lead to a better understanding of anything, because understanding is thus not sought in the first place.

I have argued for this approach before, in [Gloor Modjib(2008)]. Dawkins, in another context, explains the problem with dogmatic truths nicely in [Dawkins(2006)], page 19:

> If all the evidence in the universe turned in favour of creationism, I would be the first to admit it, and I would immediately change my mind. As things stand, however, all available evidence (and there is a vast amount of it) favours evolution. It is for this reason and this reason alone that I argue for evolution with a passion that matches the passion of those who argue against it. My passion is based on evidence. Theirs, flying in the face of evidence as it does, is truly fundamentalist.

[351] See chapter 6.2.1.

- Less importantly, the argumentation can lead to a circular argumentation culminating in an infinite regress. If something (me, for example) ascribes intentionality to something, it has to do so intentionally. There must be an intentionality ascribing the intentionality, because ascription is an intentional act. Dennett would need a more careful choice of words at least to avoid this circularity.[352]

- Intentional acts for us humans have what Nagel might call a what-is-it-like-ness to them. They are phenomenal in nature. This is implicitly included in the mere notion of intentionality – intentional acts appear familiar to us, because we all know what it is like to conduct an intentional act. We also know that for example a computer which is merely following a script, this following of the script can not possibly be anything like us conducting such an intentional act. In fact, it seems improbable that for today's computers, following a script has any what-is-it-like-ness to it at all.[353]

The same reasons apply for Dennett's heterophenomenology. Dennett defends heterophenomenology, and with it, his form of intentionality, against the attacks from Nagel in [Nagel(1991)]:

> Well, then, what does rotting chicken smell like to a turkey vulture? [...] We can uncover the corresponding family of reactive dispositions in the vulture by the same methods that work for me, and as we do, we will learn more and more about the no doubt highly idiosyncratic relations a vulture can form to a set of olfactory stimuli. But we already know a lot that we won't learn. We will never find a vulture being provoked by those stimuli to wonder, as a human being might, whether the chicken is not just slightly off tonight. And we won't find any amusement or elaborate patterns of association or Proustian reminiscence. Am I out in front of the investigations here? A little bit, but note what kind of investigations they are. It turns out that we end up where we began: analyzing patterns of behavior (external and internal – but not "private"), and attempting to interpret them in the light of evolutionary hypotheses regarding their past or current functions.[354]

Dennett does not ascribe any special meaning to phenomenality (see chapter 3.2). Arguably however, for an observer-relative ascription of intentionality to be intuitively intentional, it has to be conducted by a system that has phenomenal states. The state has to be a specimen of *phenomenal* intentionality for the system, and it has to satisfy at least Metzinger's perspectivalness and transparency constraints.

There are two argumentative approaches for disproving Dennett's claim that intentionality can exist without phenomenality:

- Dennett does not account for the special kinds of ineffability (see chapter 3.3.17) that goes with phenomenal experience, and thus does not appear to account for actual, real-world intentional experience at all. Dennett's entire argumentation builds on the assumption that there are no strictly private such experienced patterns.[355] And Metzinger clearly shows that not only are there such private patterns,[356] but also why they are private: Because they are causally active (and thus can not just be ignored), but not introspectively accessible – which is precisely Metzinger's definition of them being transparent.

[352] This argument was brought up by Thomas Nagel, see [Nagel(1991)].

[353] This does make Dennett's analogies to *today's* computers somewhat inadequate. Computers need more structures in place than they commonly have today before they can phenomenally experience content.

[354] Citation from the online version of [Dennett(1995)]. Personally, I find it puzzling that Dennett should say that consciousness is a gradual phenomenon and then assume that vultures have no dispositional stances in their intentional relations that are similar to our amusement or wondering. Of course, even if they are similar, it is very improbable that they are exactly the same despite all the "hardware" differences, so it appears indeed that they will not have the *exact* same dispositions as us humans do. It remains to be proven, however, that vultures have no dispositional stances at all.

[355] Dennett does agree that we have priviledged access to internal states. This is, however, not enough, as this does not account for ineffability.

[356] See chapter 3.3.17.

- The ineffable parts could of course be described as well in principle – finding and properly describing them would be incredibly hard, but not impossible. However, not even heterophenomenology would be sufficient there, as this phenomenal presentational content may be available for introspection, but is not symbolic – we would need both a full functional description of the human brain and a complete neurological map of it in order to even have a chance at uncovering that description. However, this is where appreciation stops working – as soon as we merely have a description, we can not relive the experience, and thus it can not be fully appreciated anymore (see chapter 4.4.1).

Dennett's position appears to be too good at what it aims at: Simplifying the subject at hand. His theory simplifies intentionality so much that essential parts of it get lost in the process. This is why his theory appears to be counter-intuitive: It does not account for the fact that we experience intentionality as a phenomenal entity, and parts of it are ineffable and subsymbolic. Merely describing the symbolic parts does not account for the entire experience.

6.4.2 Cogito, ergo sum

This chapter goes beyond the minimal notion of a self and offers an analysis of how a *cognitive* first-person perspective can come from a merely *phenomenal* first-person perspective, expanding on the findings from chapter 5.5. The difference between the two perspectives is that beings who only possess a phenomenal first-person perspective may be conscious subjects of experience, but not all of them also have first-person concepts of themselves which then enable a cognitive first-person perspective.

Metzinger explains this transition by depicting Baker's interesting analysis of Descartes' "cogito, ergo sum" ("I think, therefore I am"), or rather Baker's more specific reformulation, defining I* as the own self perceived as thinker of first-person thoughts:

> I am certain that I* exist.[357]

Metzinger starts analyzing the components of this sentence one by one, assuming that the system in question is capable of phenomenal experience and thus satisfies most constraints we saw in chapter 3.3:

- "I*": The content of the transparent self-model, under the principle of autoepistemic closure[358].
- "I* exist": Metzinger explains this as follows: "Not only the fact that the world-model is a *model* but also the fact that the temporal internality of the contents of the window of presence is an internal construct is not introspectively$_3$ available to the subject."[359] In other words, the autoepistemic closure extends into the entire world model; the world and the existence of the transparent self-model inside it are naively misrepresented as epistemically given.
- "I": The entity that holds the belief that this sentence expresses, the thinker of the I*-thought – this corresponds merely to the opaque portions of the current self-model, but internally models the system as a whole.

[357] Citation from [Baker(1998)], as cited in [Metzinger(2003)], page 398. Metzinger then goes beyond Baker and adds the distinction between [active phenomenal content] and <linguistic expressions> with corresponding symbols (square and angle brackets, respectively), which further clarifies the ways in which we refer to mental content – but for the purpose of this chapter here, we will disregard that distinction and merely look at phenomenal content.

[358] The principle of autoepistemic closure says that the transparent self-model cannot discover its own representational nature by principle. Metzinger defines it in [Metzinger(2003)], page 131:

> [Autoepistemic closure] is constituted by (a) the existence of a comprehensive representation of reality as a whole, and (b) the fact that this representation cannot be recognized *as* a representation by the system itself [...]. Put differently, in standard situations and from a first-person perspective the contents of phenomenal states always are *in a world* – they are part of *my world* [...]. This world is presented in the mode of naive realism.

[359] Citation from [Metzinger(2003)], page 400, emphasis his.

- "am certain that": The existence assumption regarding the cognitive content (the "I* exist" from above), which directly follows from the phenomenological experience of the intentionality relation – the object component is represented as immediately and unquestionably given.[360]

Put differently: The expression "I am certain that I* exist" denotes how our mind is fooling itself into believing that a self representing our mind exists. Using Metzinger's concise words again:

> The object component of the phenomenal first-person perspective is transparent and the respective person is, therefore, on the level of phenomenal experience, forced into an (epistemically unjustified) existence assumption with respect to the intentional content of the object component. The same is true of the subject component. The second defining characteristic is the transparency of the self-model, yielding a phenomenal self depicted as *being* certain.[361]

6.4.3 Empathy and intersubjectivity

Moving one step further, beyond mere subjectivity and intentionality, this part provides an outlook on what Metzinger's theory implies for sentient beings with a phenomenal self-model.

As a matter of fact, once we agree that a self is, ultimately, a phenomenal self-model, it follows that systems which have a self necessarily have the prerequisites for forming a model of a person. After all, they are already modeling *themselves* with these prerequisites. The implication is that these prerequisites could also be used for other purposes than modeling the *own* personhood – and exactly that, according to Metzinger, is what enables "offline simulations" of "first-person perspectives" even for other minds.[362] Those representations of other minds show some interesting properties:

> These experiences are interesting, because they do not satisfy the transparency constraint. When thinking about the mental states of fellow human beings in this manner, we subjectively experience ourselves as manipulating mental *representations*. On the other hand, all this activity is integrated into the transparent background of a phenomenal self, as embodied and as being the initiator of these cognitive activities.[363]

Subsequently, this leads to the possibility of modeling the phenomenal "you," meaning modeling the experience of being confronted with another agent, another person. This task really is not only one of simulation, but also one of emulation. We are able to 'read the minds' of other humans, to some extent, by constructing complex models of their selves which include goal structures and predicted motor behaviour. It has been shown that some of the same neural structures that an observed human activates, for example, for moving a finger, are also being activated in observers' brains when other agents execute these actions. Metzinger explains this as follows:

> The motor representation embedded in this partition [...] underlies the conscious experience of *being a self in the act of imitating*. The motor representation not embedded in the PSM is neither opaque nor transparent. It is a functional property, possibly not even directly reflected on the level of phenomenal experience. It is a part of the unconscious self-model which, however, is currently used as a model of a certain part of external reality, of the social environment, namely, as an *other* self-model.[364]

So, having a phenomenal self-model like we do is a very efficient way of enabling intersubjectivity by means of enabling empathy. Only if we perceive of other persons as agents can we act according to that fact, and only if we can model some of their goals and mental states (or at least goals

[360] See chapter 6.3.
[361] Citation from [Metzinger(2003)], page 401, emphasis his.
[362] See [Metzinger(2003)], pages 182-184.
[363] Citation from [Metzinger(2003)], page 365, emphasis his.
[364] Citation from [Metzinger(2003)], pages 367f, emphasis his.

and mental states that we would have if we were them) can we take these into account and act accordingly.[365]

Notably, there are parallels from this theory to Dennett's intentional stance.[366]

[365] Note that the PSM alone does not suffice to generate such relations: The relation itself is modeled as a form of intentional direction, so it does also require some form of intentionality. So it also requires Metzinger's way of forming intentional relations from chapter 6.3, the PMIR. Metzinger says in [Metzinger(2003)], on page 420: "A phenomenal first-person perspective allows for the mental representation of a phenomenal *second-person* perspective. The PMIR is what builds the bridge to the social dimension."

[366] See chapter 6.2.1.

Part III
Conclusions

The goal of this thesis is an analysis of the semantic field that surrounds consciousness. Phenomenology, qualia, subjectivity, intentionality – they all have common sense meanings (see part II) and were subject of much discussion, but appear to be disconnected and overlapping at the same time. We started out with rough definitions of the four parts, and then looked at each of them, trying to put together the opinions of more traditional exponents with those of Dennett and Metzinger, and then attempting to find a consensus among the sometimes conflicting theories that we found.

What is still missing is a summary of these individual puzzle pieces, and an outlook on what new achievements could be reached in this matter in the near future. These questions are addressed in the current part of my thesis.

7 Summary and the bigger picture

Let us take a step back from the four partial problems that I identified and analyzed in part II, and attempt to look at all of them at once. With all the single components in place, we now know that full phenomenality partly relies on subjectivity, while subjectivity itself relies on some of the multilevel constraints that make up the rest of phenomenality. We have established that intentionality relies on some basic form of self, the prereflexive proto self-model,[367] while it simultaneously is what enables the higher levels of the self-model.[368]

Furthermore, we know that the most basic form of mental content are not qualia, but phenomenal representational content.[369] Moreover, qualia have an ontological justification as well after all, because they group regions of presentational quality space for symbolic *re*presentation.[370]

With this knowledge in place, it becomes apparent that we need to structure our summary with some differences to the analysis itself, and that we have to introduce some additional levels.

Below, I present all the entities that have shown be important for phenomenal subjective intentional experience. Again, I follow the natural order of apparent increasing complexity, but no longer do I have to rely on intuition for that order like I did in part II. After all, we had a thorough look at the pieces of the puzzle by now, and have found rational ordering criteria.

This means, however, that what is more and what is less complex has changed. It also means that since so many things are interwoven, untangling them requires intermediate steps between the earlier four (too) easy levels.

This chapter summarizes and clarifies the new seven levels that have become apparent in our analysis.

7.1 Prereflexive proto self-model

There is a requirement for the PSM and the PMIR which is actually part of Metzinger's perspectivalness constraint[371] as well. I am referring to the still not fully understood prereflexive proto self-model (see chapter 5.5.3). It appears that further research into what exactly constitutes this prereflexive proto self-model is warranted and necessary. Nevertheless, it follows from Metzinger's argumentation that at least the following two factors are important prerequisites for the capability of our bodies and brains[372] to form it:

[367] See chapter 5.5.3.

[368] This was hinted at in chapter 5.3.2, but will be further analyzed in chapter 7.7

[369] See chapter 4.3.5.

[370] For this step, the PMIR is also necessary, because it is what enables the symbolic represenetation through the means of an intentional subject-object relation (see chapter 6.3).

[371] See chapter 3.3.10.

[372] Other systems capable of creating a prereflexive proto self-model would require similar prerequisites. This implies that for example a distributed neural network can not create the same sense of "self" that we can.

- Functional presentation centers in the brain that are responsible for receiving nerve cell input from various body parts, including inner organs. These functional presentation centers are grown, as phantom limb experiments (see chapter 5.5.3) show, and thus obviously genetically hardcoded.
- A single localizable body as region of maximal stability and invariance (see chapter 3.3.10) that allows the system to get accustomed to taking a first-person point of view, and also assigning entities and properties to the 'first person' this point of view originates from.

Arguably, this prereflexive proto-self is not only a prerequisite for the PSM and the PMIR, but also a prerequisite for phenomenal experience itself: Without such a first-person point of view, the possibility of localizing sensory inputs is not given, and thus any higher concepts that refer to locations in any way cannot possibly be formed.

A consciousness that has no such proto self-model may well have coherent (albeit very abstract) thoughts, but only the multimodality in which we can experience the world, as something that is all around us, allows us to form a world model at all.

7.2 Basic phenomenality

Out of the levels of consciousness we encountered in chapter 3.3.16, the one that is required for a very basic form of phenomenality[373] is probably Metzinger's "Differentiated Consciousness."

This includes the presentationality constraint[374] (the phenomenal Now in a window of presence), the globality constraint[375] (integration into a coherent global state), their expansions in the convolved holism constraint[376] (subdivisions within subdivisions that split this global state apart) and the dynamicity constraint[377] (that allows a system to experience past and future, and the Now embedded in them). Definitely, we also need the transparency constraint[378] (epistemic closure concerning inner workings of mental processes), because the world zero has to be ineffably experienced, in a naive-realistic misunderstanding as unquestionably real, or it is not phenomenal.

We recognized in chapter 3.3.16 that the ultrasmoothness constraint[379] (homogeneity of simple content and the non-existence of a grain problem) really is a prerequisite for convolved holism. This is the case because we get into an infinite regress otherwise when attempting to subdivide former wholes infinitely, which is exactly what the grain problem is about. Consequently, it will have to be included in this step as well, although it will only be required for its true strength, concept formation, when we introduce qualia.

As we found in chapter 3.3.16 as well, we probably need the global availability constraint[380] (mental contents are available for deliberately guided attention, cognitive reference and control of action), too. Minsky's reservations (see chapter 3.4.3) still apply, however: Not all the processess have access to all the data, but it certainly seems that way because we are wired in ways that make all the *necessary* data usually available to different processes happening in different parts of the brain.

At this point we have a system that has sensory organs that can potentially reach out into the world and into the system itself as well, with all the limitations that the transparency constraint sets. It is ready for input.

[373] At least if we eventually want to reach consciousness – of course, we could insert an intermediate step between the prereflexive proto self-model and basic phenomenality with Metzinger's "Minimal Consciousness" here as well.
[374] See chapter 3.3.6.
[375] See chapter 3.3.7.
[376] See chapter 3.3.8.
[377] See chapter 3.3.9.
[378] See chapter 3.3.11.
[379] See chapter 3.3.14.
[380] See chapter 3.3.5.

7.3 Phenomenal presentational content

As we saw in chapter 4.3.5, qualia are not the atomic entities they were made out to be. Phenomenal presentational content is the most basic form of experience, because there are kinds of nonconceptual experience that are so subtle and fine-grained that they cannot be categorized. These special sub-qualia entities are not accessible for concept formation or transtemporal reference.

Notably, while the step of *presenting* phenomenal content to a system is somewhat trivial from a philosophical perspective, it is very complicated from a computational point of view. Most of today's attempts at producing weak artificial intelligence are struggling with exactly this step. It mostly includes pattern matching tasks: Speech recognition, facial recognition, three-dimensional object perception, orientation, memory formation, and plenty of other, similarly complicated things.[381]

The framework and procedural possibility for these processes, in the case of biological systems, has to be genetically hardcoded and built into the prereflexive proto self-model,[382] as it is not possible to build them from scratch, not even for an efficient self-learning system like our body and brain. However, in order to be able to perceive the data from these perceptive and categorizing processes as presentational content, we also require basic phenomenality.[383]

Arguably, the representation of intensities contraint[384] (analogue representation of sensory contents) is necessary for phenomenal presentational content at least in biological systems, as probably all the important biological sensors are analogue. Digital representation of intensities, only "on" and "off" switches, are thinkable, but finding fields of application for generic consciousness where they make sense will be hard – consequently, it seems to be a safe bet to include the representation of intensities constraint in this step as well.

7.4 Basic subjectivity

I attempted to merge the theories of Minsky and Metzinger in chapter 5.5.4. According to this combination of their argumentations, subjectivity is really about multiple phenomenal self-models that the phenomenal system forms – out of phenomenal presentational content that is invariant and thus perceived as internal (which essentially is the way the proto self-model is transparently presented to the system), plus social influences and conventions.

These self-models consist of the six levels of Minsky that somewhat correspond to Freud's distinction between Id, Ego and Super-Ego.[385] However, for the higher levels among them we require an intentionality relation, so only two levels (those correlating to the Freudian Id) are part of basic subjectivity:

1. 'Instinctive Reactions' - requiring no self-model yet.
2. 'Learned Reactions' - forming a simple self-model that includes the bipartition between self and world.

These levels are necessarily entirely transparent, because as soon as we want to add opacity, we need a way to relate to the opaque parts – which is only possible if we can model intentionality, and intentionality is only added in chapter 7.5. Basic subjectivity, however, is more than the prereflexive proto self-model, in that it incorporates sensory input and allows processing it in various ways.

7.5 Intentionality

In order to progress any further from basic subjectivity, it is necessary to consciously model the relation of the system depicted by such basic subjectivity to both the world model and its own

[381] Further such subjects include integration of semantic and episodic memory into the system, learning algorithms, goal representations and modeling creativity.
[382] See chapter 7.1.
[383] See chapter 7.2.
[384] See chapter 3.3.13.
[385] See chapter 5.4.1.

self-model. Otherwise, we have to remain at a level that is barely conscious at all – a purely reactive system in many ways, and certainly not one to which we could make reasonably meaningful ascriptions of intentionality.

As Dennett showed, intrinsic intentionality, as intuitively plausible as it may be, does not seem to have any justification that goes beyond mere intuition.[386] Metzinger also insists that if we want to naturalize intentionality, we have to phenomenalize it.[387] Thus, if we want a rational theory, we have but one choice: The theory has to refrain from relying on intrinsic intentionality, and consequently, we need a replacement.

Conveniently, we have a replacement in Metzinger's phenomenal model of the intentionality relation:[388] The intentionality relation is nothing but a phenomenal model, it only has an abstract conceptual legitimation, every intentional relation supervenes locally on brain structures.

Metzinger's model of the intentionality relation is an entity which is continuously being rebuilt in a dynamic process, hardwired in the brain, that allows us to *re*present entities which formerly were mere presentata. It is phenomenal in that we experience it as being meaningful, and it depicts the relation from the transparently represented subject component to an object component which is one among a huge selection of entities that would qualify for filling the "object component" slot[389] in a system's intentional modeling framework. Notably, this object component does not need to correlate to any entities external to the system, we might be mistaken about all the things we experience.

That relation between subject and object component is then presented as being a certain *kind* of such relation – having an attitude specificator (for propositional attitude psychology) or intentional content (for Husserlian phenomenality). These kinds of relations in turn are presented as equivalence classes that the system can represent, and consequently analyze, by making them object components of a phenomenal model of an intentionality relation of their own.

This means that the PMIR has to be able to accomodate large classes of both object components and attitude specificators. It has to be able to present all of these object components as the same kind of object component – the kind of "the entity that I currently relate to." And it has to present all of these attitude specificators as "the way in which I relate to that entity."

7.6 Qualia

As we established in chapter 4.3.5, qualia are not atomic, they are merely arbitrary subdivisions of quality space that allow us to conceptualize phenomenal presentational content.[390] Still, they retain conceptual importance, exactly because they allow us to subdivide that quality space into ultrasmooth[391] subdivisions.

While qualia are thus the true test for ultrasmoothness because it is now *required* for us forming categories of phenomenal presentational content, we did introduce the ultrasmoothness constraint for basic phenomenality already[392] and can thus merely widen its application to include phenomenal presentational content now.

In chapter 4.4.3, we concluded that even though traditional qualia as truly atomic entities cannot stand when faced with phenomenal presentational content that in fact *does* subdivide it (meaning that qualia cannot possibly be atomic and undivisible), qualia still have legitimation as the concept of volumes of qualitative space, as borders of the concept formative capabilities of a system.

As a matter of fact, there are other prerequisites for qualia as well. We need to be able to form concepts, and for that we require ways to relate to phenomenal presentational content. Thus, we need both basic subjectivity (or we would not know what it is that relates to the presentational content) and intentionality (or we could not relate to it at all).

[386] See chapter 6.2.3.
[387] See chapter 6.3.2.
[388] See chapters 6.3 and 6.3.2.
[389] See chapter 6.3.1.
[390] See chapter 4.3.
[391] See chapter 3.3.14.
[392] See chapter 7.2.

7.7 Full subjectivity

There is more to full subjectivity than to basic subjectivity,[393] because as soon as we also have an intentionality relation, we can potentially relate to our basic multiple phenomenal self-models[394] (which in turn present the one proto self-model).

Again we resort to Minsky's levels of self-models.[395] The first two levels are listed in chapter 7.4, this chapter focuses on levels 3-6 exclusively:

3. 'Deliberative Thinking' - adding phenomenality and offline activation to the requirements for the self-model, and adding the requirement for a model of the intentionality relation to the system as a whole.

4. 'Reflective Thinking' - not really adding complexity, but requiring the dynamicity constraint's processes[396] to be able to be activated independently by different streams of thought.[397]

5. 'Self-Reflective Thinking' - requiring parts of the self-model to be opaque, or at least the possibility to make them opaque.

6. 'Self-Conscious Reflection' - requiring that multiple self-models have to be available in parallel and can be compared.

These six levels together form a particular self-model. One or several of these self-models is always active when there is conscious experience. There has to be at least one active self-model, or there is no experiencing self (see chapter 5.3.4) – and only one of them is fully transparent at a time, in standard cases.[398] Yet, as we saw in chapter 5.4.2, there is no complete all-encompassing entire-self-model, because that would lead to too much computational complexity for our brain for no benefit. Rather, only selected memories and patterns of behaviour are available to each individual self, both for practical (simplification through information hiding) and continuity reasons; it makes sense to be able to "depend on yourself."[399]

Such self-models are then created for descriptive context-dependant positive analysis tasks, but also for normative goal-relative tasks. The normative tasks can be ethical, moral, or serving other goals that the agent itself can form – they can be entirely altruistic or entirely egoistic in nature. Human agents are a superb example for this because they are really creative[400] concerning which goals their normative self-models serve.

The same self-modeling routines that are used for creating the self-models are then also borrowed for intersubjectivity and empathy (see chapter 6.4.3). If it is possible to model yourself, you can also model your neighbour, and as soon as you can do that, you can understand him and lead meaningful[401] discussions. As such, the possibility for us to form a self-model is also one of the necessary conditions for meaningful (in the same experiential sense) language generation, because only if you can model somebody as a person is it possible to truly tell them something.

[393] See chapter 7.4.
[394] See chapter 5.5.4.
[395] See chapter 5.4.1.
[396] More specificially, what I am talking about here are the processes currently executing those process descriptions that enable the system to satisfy the dynamicity constraint.
[397] A stream of thought is just such a process in the brain, often initiated by some form of sensory input (which is experienced as presentational content and conceptualized as qualia). Arguably, all such processes are originally initiated by either another process in the brain or some kind of sensory input (maybe triggered by the autonomic nervous system), as something *must* be their origin.
[398] There certainly are several such self-models at once transparently active in the brain of a patient with a multiple personality disorder.
[399] Citation from [Minsky(2006)], page 312.
[400] Minsky prefers using the term "resourceful" instead of "creative", because according to him that is why our lines of thought are so unpredictable: We have many resources at our disposal, various "Ways to Think," and when we are creative we merely select the "Way to Think" that appears to be most promising or tempting for reaching the goal at hand. See [Minsky(2006)], pages 275ff.
[401] Meaning here is thought of in a context-relative, not intrinsic, sense. There is a difference between meaning and the experience of meaningfulness. We encountered this before, in chapter 6.3.2.

Let us recapitulate the parallels to Metzinger's levels of consciousness in chapter 3.3.16: By now, we have added nearly all of his constraints, including the perspectivalness constraint[402] (including the proto self-model, the PSM and the PMIR), the full transparency constraint[403] (epistemic closure concerning inner workings of mental processes with the possibility of making some of these processes opaque) and the offline activation constraint[404] (allowing representational content that is not directly driven by presentational content).

8 Conclusion

With the characteristics and prerequisites for full subjectivity from chapter 7.7, together with the constraints we already added before that for basic phenomenality,[405] we now have a fully conscious system that has, to quote Searle, "causal powers equivalent to those of the brain."[406]

Even better, we even have some slight idea what those causal powers could be! They are the constraints Metzinger suggested, with some minor variations, and in a slightly more complex layering. The special causal powers of the brain that make us conscious subjects are the very same that make us able to experience the world phenomenally from a first-person point of view.

The only one of Metzinger's constraints that we did not add yet in our new seven levels from chapter 7 is the adaptivity constraint,[407] the reason for this will become apparent in chapter 8.1. Ignoring this constraint for now, this means that all the constraints Metzinger listed as prerequisites for a "Cognitive, Subjective Consciousness"[408] are accounted for. Not only that, but the added layers clarified and made explicit some interdependences between the four rough terms that we started out with, and hopefully contribute to a more sound foundation for the philosophy of consciousness overall.

8.1 The special nature of the adaptivity constraint

According to the adaptivity constraint[409], everything has to make sense from an evolutionary perspective. Metzinger designed it as a constraint that specifies that a system needs a proper history, but as shown in chapter 3.4.6, this is not necessary for the system to be conscious. Still, I proposed a modified fitness criterion for the adaptivity constraint as follows:

> Phenomenal representata are *good* representata if, and only if they successfully and reliably depict those causal properties of the interaction domain of an organism that are important for successful survival of the genotype.

The modified form of the adaptivity constraint built upon this criterion is certainly necessary for a conscious system which is also perceivable as such by us: It seems doubtful that we could classify a system as having phenomenal subjective experience if it does not exhibit any behaviours that we would expect from a system capable of that, which is, behaviours that are beneficial for the survival and integrity of the system.

Conversely, an instinct-driven system that exhibits all the characteristics of one, but also has phenomenal subjective experience that is unrelated to the capability to survive, would be classified as merely instinct-driven and again, we would not perceive it as being conscious.

The constraint is, however, not a necessary part of any of the steps in chapter 7. Consequently, the modified adaptivity constraint must be the difference between a system that merely experiences

[402] See chapter 3.3.10.
[403] See chapter 3.3.11.
[404] See chapter 3.3.12.
[405] See chapter 7.2.
[406] Citation from the online version of [Searle(1980)]. Notably, he used the term to attempt to show what it is that a computer necessarily lacks for producing intentionality and ultimately consciousness.
[407] See chapter 3.3.15.
[408] See chapter 3.3.16.
[409] See chapter 3.3.15.

phenomenally from a first-person point of view,[410] and one that we would classify as a conscious system. The modified adaptivity constraint is what consciousness visible from outside the system.

8.2 Conceptual dualism?

Chapter 5.1 introduced substance dualism, and chapter 5.5.1 explained my reasons for not including it in the summary: If we assume that a causal connection between mental entities and physical entities exists, which is necessary for our thoughts and wishes having any influence over the way our body acts, it follows necessarily that they need potentially *traceable* causal links. This means that if there are domain borders to cross between the two kinds of entities, there need to be causally analyzable ways to cross these domain borders – both bottom-up from neural pathways to mental concepts, and top-down from phenomenal experience to brain structures and states.

Now, phenomenal entities, according to Metzinger, supervene locally on brain properties. As we found in chapter 3.3.4, this means that phenomenal presentational content, which is a kind of abstract conceptual content, is realized in electrical currents and layouts of neural networks. However, while it is possible to map from one of these kinds of content to the other, this is not easy, and both levels can be analyzed rather individually:

- We can discuss and analyze biological neural networks and how the brain works, figuring out the interdependences that various brain centers have, and the functional responsibilities which can be attributed to different areas of the brain.[411]

- We can (as this thesis does) discuss and analyze phenomenal experience and mental states, abstract processes that lead to consciousness, and analyze the concept formation and modeling tasks that the system is capable of.

In theory, it is possible to 'translate' from one level to the other. Metzinger already did some research concerning the physical-neurobiological implementation of each of his constraints,[412] but for many of them he could only offer speculations concerning where and how exactly they are implemented.

The interesting thing now is that this does not stop us from analyzing causal dependencies between different abstract and conceptual mental entities. We can base research about the phenomenal level entirely on observations and even, to some extent,[413] personal experience. On the other hand, we can also research neural connections and brain structures without even having to think about what supervenes on them. In this way, the two levels are, while interconnected, also quite independent of each other.[414]

This is a more thorough way of stating what chapter 4.4.3 hinted at: Phenomenal presentational content might be reducible in principle, but not necessarily easily so. Thus qualia simplify talking about the abstract, supervening level, without having to worry about the physicalist and complex details.

This conclusion in fact leads to this thesis defending a special kind of conceptual dualism which, I believe, is implicit all of the work of all three philosophers[415] that are the foundation of my argumentation.

[410] Such a system would probably vanish within a relatively short time frame.

[411] At this level of description, what supervenes on the brain is, for the most part, merely represented as abstract data contained in electric charges of nerve cells. The supervening 'meaning' however is not important.

[412] In fact, Metzinger wrote a chapter for each of his multilevel constraints, attempting dissect its physical-neurobiological level, see chapter 3.4.1.

[413] The limits to basing research on personal experience were made explicit by Dennett, see chapter 3.2.

[414] This is, interestingly, similar to how a programmer thinks about his computer programs. When I program, I do not think about how exactly my higher-level code will be translated to bits and bytes, processor instructions and memory contents. I do not need to know how the compiler exactly translates them. A researcher who does not know the history of computing might take a seemingly unbiased approach and say "there is a spirit in this machine that allows it to process words and images, for this is more than merely mathematical calculation" – particularly when he does not have the source code of programs at his disposal, but only the compiled, cryptic and hardly decipherable end product. Again, failure of imagination is not an insight into necessity (see [Dennett(1991)], page 401).

[415] These three philosophers are of course Metzinger, Dennett and Minsky.

8.3 The hard problem demystified?

Certainly I do not mean to claim that I have given all the answers, or that I have been able to give more than just some hints of what exactly it is that makes subjective consciousness possible and phenomenally interesting. Yet I believe that I have made some first pointers to that end explicit.

From a naturalist or rational-causalist background, the traditional theories about the mind are not really satisfactory. It always amazed me how dualist tendencies,[416] untranscendability by principle and dogmatic explanations through apparently arbitrary claims are still prevalent in today's philosophy of mind.

Personally, I never understood why we should be so convinced that we are special in some way, nor why other, artificial or postbiotic systems should be excluded from the club of the entities that can potentially be intelligent. There is just no rational reason for that. There is a *legitimation* of course, the all too human fear that we could suddenly be superfluous:

> What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control.[417]

However, all we can do is try to find out whether these other systems can be intelligent, and then talk about what exactly the premises are (if there are any) that make it impossible for them to be intelligent, based on information and not doubt. We should admit that as long as no proof either way exists, we do not know for sure which systems can potentially be conscious. This thesis and the theory presented certainly points towards more kinds of systems having this potential than previously suggested.

Arguably, Searle's "special causal powers" do not explain a lot, particularly not if they are defined as the inexplicable things that make it impossible for these other systems to be intelligent.

8.3.1 A novel approach

We need to be on the way towards a better understanding of the hard problem of the philosophy of mind, one that allows us to conduct an informed talk on the things we do not know, and to narrow down what it is that we do not yet understand. I think that Dennett, Metzinger and Minsky provide important guidance. Some incompatibilities had to be clarified, but I also have called attention to many areas where the fusion of their theories proved to be fruitful and able to explain more than the parts individually.

I hope that the summary from chapter 7 in particular will be appreciated as a novel approach to the mysterious matter of subjective consciousness. Future research to put my theory to test is warmly welcome.

[416] I am here again talking of substance dualism, not conceptual dualism like in chapter 8.2.
[417] Citation from [Joy(2000)].

References

[Baker(1998)] L. R. Baker. The first-person perspective: A test for naturalism. *American Philosophical Quarterly*, 35:327–346, 1998.

[Blackmore(2008)] S. Blackmore. *Memes and "temes"*. TED Talks, 2008. http://www.ted.com/index.php/talks/susan_blackmore_on_memes_and_temes.html.

[Brentano(1995 (original 1874))] F. Brentano. *Psychology from an Empirical Standpoint*. London and New York: Routledge, 1995 (original 1874). Translated by Antos C. Rancurello, D. B. Terrell and Linda L. McAlister.

[Callaway(2009)] E. Callaway. *How your brain sees virtual you*. New Scientist, 2009. http://www.newscientist.com/article/dn18117-how-your-brain-sees-virtual-you.html.

[Chrudzimski(2005)] A. Chrudzimski. Brentano, husserl und ingarden über die intentionalen gegenstände. *Existence, culture, and persons*, 2005. Available online at http://www.roman-ingarden.phils.uj.edu.pl/pliki/arkadiusz_chrudzimski_brentano_husserl_ingarden.pdf.

[Damasio(1999)] A. R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace, 1999.

[Darwin(1859)] C. Darwin. *On the Origin of Species*. 1859. Available online at http://darwin-online.org.uk/EditorialIntroductions/Freeman_OntheOriginofSpecies.html.

[Dawkins(1976)] R. Dawkins. *The Selfish Gene*. Oxford: Oxford University Press, 1976.

[Dawkins(1982)] R. Dawkins. *The Extended Phenotype*. San Francisco: Freeman, 1982.

[Dawkins(2006)] R. Dawkins. *The God Delusion*. Boston, New York: Houghton Mifflin Company (2008), 2006.

[Dennett(1988)] D. Dennett. *Consciousness in Modern Science*, chapter Quining Qualia. New York: Oxford University Press, 1988.

[Dennett(1987)] D. C. Dennett. *The Intentional Stance*. Massachusetts: The MIT Press, 1987.

[Dennett(1991)] D. C. Dennett. *Consciousness Explained*. London: Penguin Science, 1991.

[Dennett(1995)] D. C. Dennett. Animal consciousness: What matters and why. *Social Research*, 62(3):691–711, 1995. Available online at http://instruct.westvalley.edu/lafave/dennett_anim_csness.html.

[Dennett(2003)] D. C. Dennett. Who's on first? heterophenomenology explained. *Journal of Consciousness Studies*, 9-10:10–30, 2003. Available online at http://ase.tufts.edu/cogstud/papers/jcsarticle.pdf.

[Descartes(1642)] R. Descartes. *Meditationes de prima philosophia*. Apud Danielem Elsevirium, 1642.

[Eldredge and Gould(1972)] N. Eldredge and S. J. Gould. Punctuated equilibria: An alternative to phyletic gradualism. *Models in Paleobiology*, pages 82–115, 1972.

[Evans(2003)] V. Evans. *The structure of time: language, meaning, and temporal cognition*. Amsterdam: John Benjamins Publishing Co, 2003.

[Freud(1923)] S. Freud. *Das Ich und das Es*. Wien: Internationaler Psychoanalytischer Verlag, 1923.

[Gloor(2007)] G. Gloor. *The Chinese Chatroom*. 2007. Available online at http://www.haslo.ch/philosophy/chinesechatroom.pdf.

[Gloor Modjib(2008)] G. Gloor Modjib. *The Legitimation of Traditional Phenomenology in a Rational-Causalist World.* 2008. Available online at http://www.haslo.ch/philosophy/phenomenology.pdf.

[Hilbert(1897)] D. Hilbert. Die theorie der algebraischen zahlkörper. *Jahresbericht der Deutschen Mathematikervereinigung*, 4:175–546, 1897.

[Hochberg et al.(1951)Hochberg, Triebel, and Seaman] J. E. Hochberg, W. Triebel, and G. Seaman. Color adaptation under conditions of homogeneous visual stimulation (ganzfeld). *Journal of Experimental Psychology*, 41:153–159, 1951.

[Husserl(1984)] E. Husserl. *Husserliana*, volume XXIV. Dordrecht/Boston/Lancaster: Mertinus Nijhoff Publishers, 1984.

[James(1890)] W. James. *The Principles of Psychology.* New York: Simon and Schuster, 1890. Available online at http://psychclassics.yorku.ca/James/Principles/index.htm.

[Joy(2000)] B. Joy. *Why the future doesn't need us.*, volume 8. Wired, 2000. http://www.wired.com/wired/archive/8.04/joy.html.

[Kowalsky(2003)] G. Kowalsky. *Science and the Search for God.* New York: Lantern Books, 2003.

[Kurzweil(2001)] R. Kurzweil. *The Law of Accelerating Returns.* 2001. http://www.kurzweilai.net/articles/art0134.html?printable=1.

[Leopold and Logothetis(1999)] D. A. Leopold and N. K. Logothetis. Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, 3:254–264, 1999.

[Lewes(1875)] G. H. Lewes. *Problems of Life and Mind: First Series: The Foundation of a Creed*, volume 2. London: Trübner, 1875.

[Lewis(1929)] C. I. Lewis. *Mind and World Order: Outline of a Theory of Knowledge.* New York: Charles Scribner's Sons, 1929. Citation page numbers as per the 1991 reprint by Dover.

[Libet(1981)] B. Libet. The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to p. s. churchland. *Philosophy of Science*, 48:182–197, 1981. Available online at http://www.jstor.org/pss/187179.

[Locke(1690)] J. Locke. *Essay Concerning Human Understanding.* London: Basset, 1690.

[Marais(1937)] E. N. Marais. *The Soul of the White Ant.* London: Methuen, 1937.

[Metzinger(2003)] T. Metzinger. *Being No One: The Self-Model Theory of Subjectivity.* Massachusetts: The MIT Press, 2003.

[Minsky(2006)] M. Minsky. *The Emotion Machine.* New York, London, Toronto and Sydney: Simon & Schuster Paperbacks, 2006. Available online (in a prelimary version) at http://web.media.mit.edu/m̃insky/.

[Minsky(2007)] M. Minsky. *Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* 2007. http://mitworld.mit.edu/video/484.

[Moran and Kelley(2009)] H. T. Moran, J.M. and W. Kelley. Modulation of cortical midline structures by implicit and explicit self-relevance evaluation. *Social Neuroscience*, 2009.

[Nagel(1974)] T. Nagel. What is it like to be a bat? *The Philosophical Review*, 83 (4):435–450, 1974. Citation page numbers as per the version available online at http://organizations.utep.edu/Portals/1475/nagel_bat.pdf.

[Nagel(1991)] T. Nagel. What we have in mind when we say we're thinking. *Wall Street Journal*, November 7 1991.

[Newman(1995)] J. Newman. Reticular-thalamic activation of the cortex generates conscious contents. *Behavioural and Brain Sciences*, 18(4):691–692, 1995. Available online at http://imprint.co.uk/online/new1.html.

[Orwell(1949)] G. Orwell. *Nineteen Eighty-Four*. New York: Harcourt, Brace & Co, 1949.

[Plato(380 B.C.E.)] Plato. *The Republic*, volume VII. 380 B.C.E.

[Pöppel(1994)] E. Pöppel. Temporal mechanisms in perception. *International Review of Neurobiology*, 37:185–202, 1994.

[Raffman(1995)] D. Raffman. *Conscious Experience*, chapter On the Persistence of Phenomenology. 1995.

[Searle(1980)] J. R. Searle. Minds, brains, and programs. *Behavioural and Brain Sciences*, 3(3):417–457, 1980. Available online at http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html.

[Searle(1992)] J. R. Searle. *The Rediscovery of the Mind*. Massachusetts: The MIT Press, 1992.

[Sehon(2005)] S. Sehon. *Teleological Realism*. Massachusetts: The MIT Press, 2005.

[Sellars(1963)] W. Sellars. *Science, Perception and Reality*. London: Routledge & Kegan Paul, 1963.

[Telegraph.co.uk(2008)] Telegraph.co.uk. *Otto the octopus wreaks havoc*. 2008. http://www.telegraph.co.uk/news/newstopics/howaboutthat/3328480/Otto-the-octopus-wrecks-havoc.html.

[Watts(1993)] L. Watts. *Cochlear Mechanics: Analysis and Analog VLSI*. PhD thesis, California Institute of Technology, 1993. Available online at http://www.lloydwatts.com/thesis.html.

[Watts(2009)] L. Watts. *Lloyd Watts Neuroscience*. 2009. http://www.lloydwatts.com/neuroscience.shtml.

[Zahavi(2003)] D. Zahavi. *Husserl's phenomenology*. Stanford: Stanford University Press, 2003.