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

July 16, 2009

Abstract

The hard problem of consciousness is a very fascinating, but also a very disputed matter. It is the one thing that we don't 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 actually four important parts to this question, all of which warrant a closer look: First, how does the brain produce phenomenality? What are qualia? The most mystifying part: How does subjectivity emerge? And finally, what is intentionality, and how does it all fit together?

When looking for answers to these questions, I encountered two philosophers with interesting proposals. Daniel C. Dennett was one of them. His works Consciousness Explained (1991) and The Intentional Stance (1987) break not only with Cartesian dualism but also with Husserlian phenomenology and its descendants. Thomas Metzinger was the other, he published Being No One in 2003 as his first big 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 for us seems to be at the core of subjectivity; the self.

Both Dennett and Metzinger have a somewhat novel approach to the questions at hand, in that they look at them in an interdisciplinary manner and always 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, and will assist us in our climb towards a better overview.

I will try to tackle the hard problem of consciousness in three parts. First, a short introductory section that tries to find the context the discussion takes place in, and points out which great thinkers the giants upon whose shoulders we stand today were, and how some important terms are to be understood. Then, the main section, in which I want to focus on how Dennett and Metzinger (and, for subjectivity, Minsky) try to answer the four partial questions I mentioned, and try to critizise what I perceive to be shortcomings of their theories. I will also attempt to consolidate both of them into one model where this is possible.

The third section will then consist of a summary of all the parts we will have looked at until then, putting them in a bigger picture, and a short outlook towards what the future might bring for philosophy of mind and its hard problem.

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 through all this time. I also want to thank the rest of my family for all their support.

At University, this thesis would not have been possible without the aid of Prof. Dr. Gerd. Graßhoff and Prof. Dr. Eduard Marbach. I 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 all too often only 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 Simon Bünzli, Stefan Aeschbacher, and Daniela Reist, who were (and, at the time of this writing, are) the proofreaders who I'm sure will find gazillions of errors, typos and mistakes.

Contents

Part I
Introduction

Before we can get to the heart of the matter of what consciousness is all about, we have to make sure that we're all on equal footing. Often, different philosophers use the same terms with different meanings. Even if I take great care not to talk about Wittgensteinian *Scheinprobleme*, it is important to get potential misunderstandings due to linguistic inconsistencies out of the way before they have a chance to rear their ugly heads. For that end, and to aid readers with less of a philosophical background, I will introduce and define the major terms that I think are important for the discussion.

1 Definitions

Major misunderstandings in philosophy are usually due to differing views on what common terms are supposed to mean. So as a very first step, I will introduce the way I understand some of the important concepts I will use throughout this thesis, and I will hopefully show the implications this understanding has for philosophy of mind in general and the hard problem of consciousness in particular.

1.1 Epistemology

Epistemology, the philosophical look at knowledge. Seeing how it is always conscious minds that are interesting systems when we look at what knowledge is, and how nearly all epistemological research so far has been done from a human first-person perspective, we can really say that epistemology in an interesting sense essentially is about *human* knowledge.

Theories as to what exactly knowledge is are varied and disputed. A commonly accepted view, dating back to Plato's Theaetetus[1], is that knowledge is a justified true belief. In formal terms: "A subject S knows that a proposition P is true if, and only if:

1. P is true
2. S believes that P is true, and
3. S is justified in believing that P is true"[2]

This approach does have various problems and shortcomings,[3] but since epistemology really isn't a main subject of this thesis, we'll accept this definition of proper knowledge for now. I want to point out important implications of this definition however.

[1] See [60].
[2] Citation from [8].
[3] See [6].

1.1.1 The Difficult Relation to Ontology

Obviously, we can be wrong about the world. We can misrepresent various aspects of happenings and entities around us – shapes that appear to be ghosts,[4] optical illusions,[5] and other, less reproducible cases.

All of us are frequently wrong about one aspect or another of our environment, and often we can't subjectively make out whether we are right or wrong about any particular aspect – we need the assistance of other agents, an intersubjective look at the entities in question.

The implications for ontology (the science of the things that exist, thus the science of the things of which existence assumptions are true) are that we cannot necessarily deduct ontological statements from empirical observations – the epistemology of empiricism is at least complicated.

1.1.2 The Difficult Relation to Introspection

Similar arguments come into play when we talk about introspection. If we are to believe that we can introspectively gain proper knowledge about all aspects of our mental world we direct introspective attention at, we have to believe with Descartes that we cannot be wrong about our own mental states.[6]

There are various counter-examples to that view however. Trivially, there are pathological cases where patients believe they're blind when they really aren't or vice versa, or phantom limb syndromes where patients believe that they feel a limb that is not there (anymore), or even stranger symptoms like patients that do believe that they do not exist at all, defying all logic.[7] But there are other, non-pathological cases that show up the fallibility of introspective access as well:

> In one study, subjects were presented with four pairs of stockings and asked to indicate which pair had the highest quality. The leftmost pair was preferred by a factor of almost four to one. However, unbeknownst to the subjects, all four pairs of stockings were identical. Though position effects were clearly playing a role in the subjects' choice, none of them identified position when asked to explain their reasoning, and those who were asked explicitly whether position played any role in their reasoning process all denied it. [...]
>
> However interesting this result, Nisbett and Wilson's work might not seem especially threatening to most proponents of infallibility, since it concerns introspective access only to higher order reasoning processes [...]. In contrast, empirical work on "changeblindness,"

[4] Many examples of this can be found at [71].

[5] Quite a few of those can be found at [14].

[6] See [44], and my look at Descartes' "cogito, ergo sum" in chapter 4.5.3.

[7] See [52], pages 454ff. The question whether those patients are still rational agents can reasonably be asked – arguably, they are, since their rational argumentations are sound if one accepts the premise that they don't exist. As we will see, this is transparently given as unquestionably immediate for them, just like existence is given for us. See chapter 1.5.3.

> which calls into question our introspective access to our current perceptual states, seems to pose a deeper threat. According to work done by Kevin O'Regan (who works, ironically, at the Universite Rene Descartes in France), subjects typically fail to notice even large changes to objects in their visual field, as long as the change occurs simultaneously with some other "disruption," such as a blink or a mudsplash on a windshield.[8]

I think it is safe to infer that the epistemology of introspection is just as complicated as the epistemology of empiricism.

1.1.3 The Difficult Relation to Ineffability

Ineffable are those things where a symbolic representation is not available. To say it with Wittgenstein, "Wovon man nicht sprechen kann, darüber muß man schweigen."[9]

This means that ineffable entities in consciousness are by definition not intersubjectivizable.[10] So for introspectively gained ineffable insights, the epistemological standing is even worse than for regular introspection; Not only can we be wrong about them, but we can't even intersubjectively verify them.

1.1.4 The Impossibility of Knowledge

I think that epistemology is, in fact, not such a complicated subject as we tend to make it out to be – and obviously, not as complicated as it seems from my discussion so far here. The only way in which we can know if a belief is not only justified, but also true, is if we assume things that imply verity in their very definition (like intrinsic intentionality or dualism) and only due to that have verifying force (making that verifying force circular). It is the classic Münchhausen trilemma we fall into, of course. When attempting to justify any argumentation, we have the choice among the following three alternatives:[11]

1. An infinite regression, which appears because of the necessity to go ever further back, but isn't practically feasible and doesn't, therefore, provide a certain foundation;

2. a logical circle in the deduction, which is caused by the fact that one, in the need to found, falls back on statements which had already appeared before as requiring a foundation, and which circle does not lead to any certain foundation either; and finally:

3. a break of searching at a certain point, which indeed appears principally feasible, but would mean a random suspension of the principle of sufficient reason.

[8] Citation from [44].

[9] Odgen translated this as "Whereof one cannot speak, thereof one must be silent," according to [11]. German citation from [72], page 85.

[10] Dennett disagrees, see chapter 2.2 and my discussions in chapters 3.4.1 and 5.4.1.

[11] Citation from [9], which translated [13].

If we do not have any dogmatic definitions of truth that provide the foundation for the third alternative of just breaking the search while also maintaining rationality, we can't possibly be directly in contact with anything in this world – it is just models that are in direct contact with other models, only influenced and partially created by (probably) true things out there in the world.[12] Such foundations of truth are of course proposed by many, mostly religious, thinkers, but I find it hard to accept entities that I am not allowed to question for principial reasons.[13]

So, in order for a belief to be knowledge it necessarily has to be true, and we cannot verify whether any belief we hold is true or not at all. Consequently, we cannot know whether any beliefs we hold are cases of knowledge or not. We can merely find out which of our beliefs are *adequate*, not which ones are true.

This of course is only true in a strict, epistemological sense. In day-to-day talk, we often talk about "knowledge" when we don't mean those strict epistemological definitions. Some beliefs are so adequate, and so justified, that intuitively we think that they *must* be knowledge – Newtonian physical laws for example, or Descartes believing that "cogito, ergo sum". This belief that those things must be knowledge,[14] in turn, is again highly adequate.

But I consciously chose the two examples for this to point out that even if we deeply *believe* that those beliefs have to be knowledge, we can't *know* whether they are. Einstein's theories show us how Newtonian laws don't apply for really large and really small scales, quantum mechanics do the same for even smaller scales, and I'll show up Metzinger's dissection of Descartes' belief in chapter 4.5.3.

We hold justified beliefs that are adequate for all situations we might encounter them in, and they seem so natural to us that we are certain that they are cases of proper knowledge. I would go so far to actually call them "folk knowledge." However, we cannot find out whether they are cases of *epistemic* knowledge in a stricter sense, or not.

1.1.5 Justified Adequacy is Criticizable

Justified adequacy is by definition criticizable, and there are two places where an apparently justified adequate belief can be attacked:

- Justification can and should always be questioned, and is an argumentative construct. Beliefs can be called into question by other beliefs with different justifications. It then has to be the goal of any rational discussant to find out which justification relies on more reasonable (and probable, given our frameworks of belief) basic premises.

- Adequacy can be proven wrong by means of simple examples of inadequacy

[12] See chapter 2.4.6.

[13] See chapter 1.1.5.

[14] Interestingly, Newtonian laws are produced by a posteriori reflections on nature, while Descartes' insight on the other hand follows from a priori introspection – both of them however are perceived as unquestionably true, thus apparently cases of knowledge.

> for any given belief. Keep in mind however that a belief can still be adequate for the majority of cases, barring (sometimes explicit) exceptions, and that this is actually the most common kind of adequate belief we hold as such folk knowledge. We all know the saying, "the exception proves the rule."

This is somewhat opposed to us putting blind faith into any kind of last authority. Personally, I do believe that any such last authorities like books or god-like entities that are not merely widely accepted as such, but actually provide an ontological foundation are highly improbable – and even if they exist, we cannot possibly know which of the various such entities that are offered as being the one true last authority happens to be the correct one.[15] On the other hand, I do believe that rationality and logic are what the world is built on (and that it is perfectly possible, albeit improbable, that god-like entities exist that are *not* last authorities). This to me seems to be one of those two basic disagreements that cause most cultural struggles these days – the other one being different views on which last authority is the correct one.

It seems to me that merely relying on rationality and logic rather than complex last authorities makes a huge lot more sense from a philosophy of science point of view, and it is both more justifiable and more adequate. This does however not mean that this view is necessarily correct. We cannot know whether it is.

1.1.6 The Search for Justified Adequate Beliefs

My point and the motivation for these observations: I do believe that the search for an adequate representation of the innermost workings of consciousness is possible – whether we can introspectively access those things or not.

Furthermore, I do also believe that introspectively (empirically) gained beliefs about those innermost workings of consciousness do not necessarily have more justification than a priori gained beliefs through reasoning and analogy (probably relating to other facts that then again can be empirical or a priori). It is the test of time, thought experiments and psychological studies only that can show us which of our theories make more sense.

Since many introspectively gained beliefs have proven to be inadequate for some border cases of consciousness or other, it appears that there are cases where a priori reasoning is better than introspection at gaining knowledge about our inner workings.

I want to explore the hard problem of consciousness with this epistemological background: It is not epistemic knowledge I am after. It is the next-best thing, and the only thing us humans can possibly gain: Folk knowledge, jus-

[15] Strictly speaking, it's rarely the god-like entities that are considered to be the last authority, but rather the various memes that cultures built around their own interpretations of these god-like entities. The rational legitimation of those memes does rest on an understanding of the memes being given by the god-like entity itself – and this is what seems to give the religious books, the holders of common and accepted memes, their perceived normative power.

tified adequate beliefs that still make sense even after we throw the strangest pathological border cases of consciousness at them.

1.2 Dualism

Substance dualism, which is the kind of dualism I want to talk about, essentially is the postulation of two kinds of essences: A physical, nonthinking essence (which builds the world, body and brain) and a nonphysical, thinking essence (which builds the mind). Thus, it is the separation of body and mind, and didn't really start with Descartes – it dates back to Zarathushtra, Plato and Aristotle.[16]

There are subtheories of dualism.[17] They all 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 naturalist 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 "mind stuff" into the pineal gland, Dennett puts it pretty 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?[18]

While as we already saw in chapter 1.1, I perfectly agree that the world consists of more than what is epistemologically graspable by us human beings, that doesn't make this additional matter non-physical or even non-naturalist. And in particular, it is a very trivial finding that every causal relationship is causally analyzable. Causes for any effect, whether we can touch and see it or not, can be scrutinised in a logical manner. And as mathematics or particle physics show, we are perfectly able to perform these kinds of investigations even with epistemologically nongraspable subject matters.

Applied to philosophy of mind: There is nothing that stops us from investigating logical interactions not available for introspection.

[16] See [3].

[17] According to [3], those 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).

[18] Citation from [30], page 34, emphasis his.

1.2.1 Hidden Forms of Dualism

In addition to openly dualist theories, many recent philosophers still have hidden dualist (intuitive) premises when looking at a world they perceive as naturalist. Searle for example posits special causal powers that are not necessarily replicable by a formal program. His argument is:

> I offer no a priori proof that a system of integrated circuit chips couldn't have intentionality. That is, as I say repeatedly, an empirical question. What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that.
>
> [...]
>
> Now I argue at some length that [the circuit chips] couldn't have intentionality solely in virtue of instantiating the program. Once you see that the program doesn't necessarily add intentionality to the system, it then becomes an empirical question which kinds of systems really do have intentionality, and the condition necessary for that is that they must have causal powers equivalent to those of the brain.[19]

I explored this argument in-depth in an earlier paper on exactly this subject matter,[20] and while my understanding of the specifics may meanwhile have grown a bit and I do believe that some of what I wrote back then, I still stand behind the main argument I put forth:

Yes, there is an empirical question whether a particular combination of circuit chips (including memory chips) and a formal program (including memory contents) has the causal powers of a brain that make intentionality[21] (and thus strong artificial intelligence) possible. However, assuming the correctness of the Church-Turing thesis,[22] there is no fundamental obstacle that would stop circuit chips from being able to implement a formal program that produces intentionality.

---

[19] Citation from [63], page 453.

[20] See [37] – Dennett makes a very similar argument regarding mystical causal powers of the mind in [29], pages 323ff, and regarding Turing machines in [30], pages 209ff. My argumentation would probably have been more succinct if I had read Dennett before that paper of mine, admittedly.

[21] Used in a wider sense here than I will narrow it down to, see chapter 5.4.

[22] The Church-Turing thesis essentially says that "what is effectively calculable is computable" (see [4]) – I tried to flesh this out in [37], page 15, and came up with this: "What is produced by any means whatever that uses a method each step of which is precisely determined and which is certain to produce the answer in a finite number of steps, can be produced by a Turing-machine or equivalent mechanical device."

A critic might say here that we know that there are incomputable functions - for example, there is no way to calculate most real numbers. However, if we are to assume that human-like intelligence is such an incomputable function, we also have to assume that the human brain is more than a mere machine – since no mere machine can calculate incomputable functions.

So unless we assume mystical powers (yet again), the brain's behaviour can be reproduced in any other Turing-complete machine.

The empirical question isn't whether circuit chips are able to implement the program, but whether a particular program enables those causal powers. That is, unless Church, Turing, Searle, Dennett, Metzinger, Minsky and others are wrong and the brain isn't merely a machine after all.

1.2.2 Possible Origins of Dualism

Metzinger points out how it is intuitively highly attractive to think in dualist terms, particularly when out-of-body experiences come into play:

> For anyone who has actually undergone that [out-of-body] experience, it will be almost impossible not to become an ontological dualist afterward [...]. In all their realism, cognitive clarity, and general coherence, these phenomenal experiences will almost inevitably lead the experiencer to later conclude that conscious experience can, as a matter of fact, take place *independently* of the brain and the body: what was phenomenally possible in such a clear and vivid manner simply must be metaphysically possible.[23]

His theory does offer an explanation for this:

> However, the information that is not subjectively available to them, of course, is that all of this is just a *model* of reality generated by their central nervous system.[24]

This just as an interesting anecdote. We will look at the nature of that model in chapters 4.3 and 5.3.

1.2.3 The Cartesian Theater

Dennett also tries and succeeds to find further fragments of dualism, anachronist remainders of dualist ideas. One such remainder is the Cartesian Theater. It is the idea that there is a central, clearly localizable place of "meaning happening" in the brain. The idea that there is a locus of understanding, 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 understanding experience a lot easier:

> 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.[25]

The trouble is that, as we will see, such a clear locus is not available, but rather, there is a "spatiatemporal smearing of the observer's point of view in

[23] Citation from [52], page 503, emphasis his.
[24] Citation from [52], page 502, emphasis his.
[25] Citation from [30], page 107, emphasis his.

the brain"[26] – which is what gives rise to Dennett's model of multiple drafts. Since there is no such clearly definable point of understanding, neither in space nor in time, there can't possibly (from a scientific perspective) be a Cartesian Theater of any kind anyway.

But a Cartesian Theater additionally has metaphysical problems, too:

> The Cartesian Theater may be a comforting image because it preserves the reality/appearance distinction at the heart of human subjectivity, but as well as being scientifically unmotivated, this is metaphysically dubious, because it creates the bizarre category of the objectively subjective – the way things, actually, objectively seem to you even if they don't seem to seem that way to you![27]

Those things would, of course, be the ways in which content is produced in that Cartesian Theater locus, if it would be interpreted (by the audience) to have another meaning than the presenting mechanisms (the actors) intended it. You see, the trouble goes deeper than the quote itself even mentions, as this would also require intention and possibly intentionality on both sides of the Cartesian Theater stage, in both the actors and the audience. The Cartesian Theater, even if it existed, would thus fail to explain anything.

Dennett explains how there really can't be a Cartesian Theater if we do not commit to dualism, and conversely, how if we do not want to commit to dualism we cannot believe that there is a Cartesian Theater:

> When you discard Cartesian dualism, you really must discard the show that would have gone on in the Cartesian Theater, and the audience as well, for neither the show nor the audience is to be found in the brain, and the brain is the only real place there is to look for them.[28]

Note how this doesn't disqualify the Pineal Gland from having a special role in intentionality. It just disqualifies it from being the one central choke point where *all* understanding happens at once.

1.3 Supervenience

Supervenience is an attempt at finding a concept that can explain the apparent dualism between mind and body. Supervenience as a term was introduced by Donald Davidson:

> Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental

[26] Citation from [30], page 126.
[27] Citation from [30], page 132.
[28] Citation from [30], page 134.

> respect, or that an object cannot alter in some mental respect without altering in some physical respect. Dependence or supervenience of this kind does not ential reducibility through law or definition; if it did, we could reduce moral properties to descriptive, and this there is good reason to *believe* cannot be done; and we might be able to reduce truth in a formal system to syntactical properties, and this we *know* cannot in general be done.[29]

This definition does imply that Davidson meant to say that they weren't reducible at all – resorting to intrinsic intentionality. Leaving intrinsic intentionality aside[30] for now, the rest of the definition leaves it somewhat open whether those mental events are actually reducible in principle and merely practically irreducible, or irreducible in principle already – and the word has been used with both meanings by various philosophers since. I want to pick out but two:

- Scott Sehon on one hand[31] uses it in a way that bears striking similarities to the concepts of epiphenomenalism, parallelism and occasionalism minus divine hands. He strongly opposes all reductive tendencies and property dualism, and uses the fact that no one-to-one reductionist mappings of complex mental states have been found to date as inductive "proof" for the claim that mental events are irreducible in principle to physical events.[32]

- Thomas Metzinger on the other hand agrees that it is impracticable to attempt to reduce mental events to physical events, but does not say that they are irreducible in principle. Quite the opposite, he proposes a model that could enable such a reduction: functional clusters of neurons, "combining a high internal correlation strength between its elements with the existence of distinct functional borders,"[33] and measurable by means of a cluster index. As he writes about such functional clusters, "this island of causal density is constituted by a certain subset of neural elements that are strongly coupled among themselves, but only weakly interactive with their local environment within the system."[34] He proposes such functional clusters for many of his multilevel constraints, while openly admitting that research has not progressed far enough for all of them yet.[35]

I tend to agree with Metzinger's point of view in many respects, and this is one of them. Practical irreducibility doesn't have to be the same as principial irreducibility, and in this case, it seems like it really isn't.

[29] Citation from [24], page 214.
[30] With a reason, see chapter 5.2.3.
[31] See [65].
[32] Arguably, that is a bad case of Dennett's philosophers syndrome. See chapter 1.5.2.
[33] Citation from [52], page 141.
[34] Citation from ibid.
[35] He is also backed up by even more recent research like [36].

1.4 Evolution

Evolution is the theory that tries to make sense of the fact that all living species on our planet seem to be related in many ways.[36] All living cells use the same basic set of nucleotides and amino acids, and there are large similarities in DNA between for example humans and chimpanzees (depending on the study, between 95% and 98.5%[37]).

What Darwin essentially said:

> 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.[38]

Note how this is not a theory about random things just miraculously happening in very improbable ways that eventually produce humans. Rather, it is a very brutal theory about a big sample of small, incremental changes being implemented and tested all the time, and most of these changes being dismissed as not fitting – through death of the individuals carrying unfitting gene changes. Given a large enough sample, there have to be some changes that make the offspring better adapted. Those (and the ones that are don't influence the chances of survival at all) are the only ones that persevere. All the others die.

I will assume that this thesis holds merit, and that our scientific methods are justified. This does, to some extent, mean that the contents of this thesis are at least in parts incompatible with creationism and generally religion.

Particularly the teleological argumentation that we will see when looking at Metzinger's teleofunctionalism and Dennett's explanation of the genesis of intelligence are not possible without referring to scientifically grounded theories. I will require evolution to explain things in chapters 1.4.3, 2.3.13, 2.4.5, 4.2.1, 4.3 and 5.2.3. Which goes to show: Evolution is able to explain way more things than Darwin had in mind when he proposed the theory.

Personally, from a philosophy of science perspective, I do see those the wide areas that require evolution to be properly explainable more as a reassurance that evolution indeed does hold merit than as a reason to dismiss both evolution and the explanations resulting from it. After all, I believe that most of us strife to understand more of the world than what little we do now.[39]

[36] See [7].

[37] See [18]. It is interesting how a similarity of 95% is considered to be low in some deeply religious circles. To quote [33]: "Will evolution be called into question now that the similarity of chimpanzee and human DNA has been reduced from >98.5% to ~95%? Probably not. Regardless of whether the similarity was reduced even below 90%, evolutionists would still believe that humans and apes shared a common ancestor. Moreover, using percentages hides an important fact. If 5% of the DNA is different, this amounts to 150,000,000 DNA base pairs that are different between them!" I will leave the number 2,850,000,000 as an exercise for the reader.

[38] See [23], as paraphrased in [16].

[39] See chapter 1.5.3.

1.4.1 Selfish Genes

Eventhough it is not entirely undisputed, I will also adopt Dawkins' view on natural selection and evolution not being a purpose in and of itself, and animals being mere vehicles that enable the survival of our genes. As he writes:

> We, and all other animals, are machines created by our genes. [...] Our genes have survived, in some cases for millions of years, in a highly competitive world. [...] I shall argue that a predominant quality to be expected in a successful gene is ruthless selfishness. This gene selfishness will usually give rise to selfishness in individual behaviour. However, as we shall see, there are special circumstances in which a gene can achieve its own selfish goals best by fostering a limited form of altruism at the level of individual animals. [...] Much as we might wish to believe otherwise, universal love and welfare of the species as a whole are concepts that simply do not make evolutionary sense.[40]

So, animals are essentially selfish, unless altruism benefits the survival of its genes. In some cases, altruism goes very far: Sometimes, an individual animal acting suicidal actually benefits the reproduction of its genes, like when bees die after stinging an enemy of the hive, or birds risk their lives to warn the flock of an approaching hawk.

It is very important to note that Dawkins did not mean to say that genes are selfish in any self-aware sense. They merely are selfish in that only those genes survive that are predisposed in ways that make them produce forms of life that has high odds of reproduction, and thus is able to further the population of a particular genotype and produce more of these genes. In this sense, only selfish genes survive, while all genes that are not selfish necessarily die. So, all genes that actually do survive must be selfish. But this does not imply any kind of intrinsic intentionality of genes or the likes.

Dawkins also introduces memes,[41] small bits of knowledge or ideas or cultural heritage (I find it hard to find a fitting word besides, well, "meme") that get passed on from individual to individual, allowing way faster progress and a kind of secondary evolution.[42] Susan Blackmore goes even further and introduces temes as a third replicator after genes and memes.[43] These are very interesting subjects, but unfortunately, I do not have the space to pursue them any further.

1.4.2 Deficiencies of Bio-Evolution

Evolution might have worked a treat for life on earth and possibly elsewhere, but it certainly wasn't the best way to go about it. Minsky points out the following flaws of genetic, Darwinian evolution, saying that "it's so easy to improve

[40] Citation from [25], page 2.
[41] See [25].
[42] See [45] for how fast our progress currently actually is – and [43] for a scarier outlook.
[43] See [16].

evolution it's not funny":[44]

- Genomes have no explicit goals – so they are unable to selectively search for structures with specified properties. Such knowledge could reduce a search by many orders of magnitude.
- Genomes do not represent what their products do or how they work, because they do not describe the structures that they construct or the processes those structures support. *This makes it hard for evolution to further improve the structures that actually do the work.*
- *Evolution selects the ones that survive, but forgets why all the others died.* Thus it can avoid the most common mistakes, but cannot accumulate knowledge about less common ways to fail; only Cultures can accomplish that.

So, only a few additional constraints can make assisted evolution way better than Darwinian one, and for some questions, evolution is not the best answer. This mainly applies to artificial intelligence where we attempt to reproduce consciousness, but since that is not the subject of this thesis, I will not follow this line of thought any further.[45]

Interestingly, biology also has mechanisms of adaptation on reproduction that go beyond mere basic genetics. The field of epigenetics is very interesting, and shows up that in biological contexts, it is often the circumstances that make an organism activate some parts of its genetic code and not others.[46] Still, culture and memes have to step in and improve basic genetics to the point where they allow today's rapid development.

1.4.3 What Consciousness has to do with Evolution

Both Dennett and Metzinger rightly point out the important role that evolution plays when we're looking at consciousness. Metzinger explicitly makes teleofunctionalism a background assumption of his work,[47] while Dennett actually has an entire chapter on the evolution of consciousness.[48]

Metzinger: Teleofunctionalism in the way Metzinger uses it appears to be the following assumption: Given there is a system that developed in an evolutionary process. Every functional property of that system was itself developed in this evolutionary process, and was selected due to evolutionary fitness in the particular niche the system was adapting to. So for every functional property of a system, there was a goal that its implementation served, and that goal was in turn beneficial for the reproduction of the genotype in question.

[44] This list is a verbatim copy of a slide in [54] starting 33:15 – the quote is from ibd. 34:10.
[45] See chapter A.1.
[46] See [68].
[47] See [52], page 26.
[48] See [30], pages 171-226.

This view might seem a bit radical, and it finds its culmination in the adaptivity constraint that we'll look at in chapters 2.3.13 and 2.4.5. Personally, I am rather certain that there are mutations and functional properties that do not have a teleofunctionalist justification like that – mutations that are *not detrimental* (as opposed to *strictly beneficial*) for a genotype's continued existence can just as well survive reproductive and adaptive activity. For consciousness itself however, being very complex and very fragile, I imagine that teleofunctionalism applies to a large extent, and as such I believe that I can and will adopt it for my thesis.

Dennett: Dennett offers a few important observations and speculations, looking at consciousness not as a black box, but from the inside out – since that approach is radically more comprehensible:

- Selfishness must have developed with the first organisms already, although it wasn't conscious yet – "The first reasons preexisted their own recognition,"[49] and the main reason for any actions was and is self-preservation. This requires a tripartition of events into good, neutral, and bad.
- Nature and evolution never nicely split functions apart as humans tend to do (to avoid unwanted side-effects), but just tries long enough.
- The requirement of being able to get out of harm's way invokes a radically new question for organisms: "Now what do I do?"[50] – thus giving rise to subjectivity, then anticipation and alarm states where "the animal stops what it is doing and does a quick scan or update that gives every sense organ an opportunity to contribute to the goal of available and relevant information."[51]
- Only after this happened was there the possibility for conscious awareness – gathering information for future use. This gave rise to another problem: There was no captain, so the various volounteers (all the parts of the brain that had goal representations and produced action dispositions) had to organize themselves into one stream of thought, since the body can only do one thing at a time.
- Autostimulation (for example, talking to oneself as a way to solve a problem) is a way to invoke resources that were only used when our ancestors were stimulated by fellow humans – but can now also be used for solving problems we have ourselves.

Notably, Dennett doesn't offer explicit proof or even models for every one such step, but he drafts a reasonable theory regarding how evolutionary pressure made consciousness what it is today. He also notes that while evolutionary pressure was long the driving force behind adaptations of the genome for new

[49] Dennett, page 174
[50] Citation from [30], page 177.
[51] Citation from [30], page 180.

tricks a species learned, "the tricks [of consciousness] have so altered the nature of the environment for our species that there is no longer significant selection pressure calling for further hard-wiring."[52] He does of course speak (implicitly) about memes.

1.5 Some further Concepts

There are some concepts that I think are important, but don't fit into any of the bigger chapters. I will present them in a condensed form here.

1.5.1 Suitcase Words

In true Wittgensteinian tradition, Minksky points out how often, we seem to see problems where there really wouldn't be any if we would just use fitting terminology.

The particular problem that Minsky sees is that in psychology, we often use just one word to describe an entire suitcase of concepts. He first introduces the term of "suitcase words" early on in his book, asking: "Why do we pack such dissimilar things into those suitcase-like words?"[53] He then goes on to define them further, saying that "*emotion* is one of those suitcaselike words that we use to conceal the complexity of very large ranges of different things whose relationships we don't yet comprehend."[54]

Only later in the book does he offer speculations on why we seem to have so many suitcase words (and not just in psychology), looking at it from different angles:

> Psychologist: Suitcase words are useful in everyday life when they help us to communicate. But we won't know what each other means unless we share the same jumbles of ideas.
>
> Psychiatrist: 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.
>
> Ethicist: we need the idea of consciousness to support our beliefs about responsibility and discipline. Our legal and ethical principles are largely based on the idea that we should only censure "intentional" acts, that is, ones that have been planned in advance, with awareness about their consequences.
>
> Holist: Although many processes may be involved, we'll still need to explain how they combine to produce our stream of conscious thoughts – and our explanations will need some words to describe the phenomena that emerge from this.[55]

[52] Citation from [30], page 190.
[53] Citation from [53], page 12.
[54] Citation from [53], page 17, emphasis his.
[55] Citation from [53], page 110. Minsky does use "intentional" in a completely different sense from the one I use in this thesis.

The important philosophical lesson we can draw from this is relatively simple: We should stay the course after the linguistic turn, and keep questioning everything that our language implies. If that means breaking up tried and treasured concepts, so be it. Although Minsky does have a word of caution for us as well:

> I am not suggesting that we should try to dissect and replace all our suitcase-words, because they incorporate ambiguities that have evolved over centuries, to serve many important purposes – but also, they often handicap us by preserving outdated concepts.[56]

Well, replacing might be a bit too much indeed. But dissecting them while keeping the newfound parts' suitcase categories in mind, in my opinion, can't hurt.

1.5.2 Intuition

There are two things that I want to point out with regard to our intuitions about things in general, but consciousness in particular:

1. Failure of imagination is not an insight into necessity – what Dennett calls "Philosopher's Syndrome."[57] It is not the case that something is necessarily not true merely because we cannot imagine it being true. We are built a certain way, with certain dispositions that helped us survive for long enough to stand where we are now. Those dispositions were (and most of them, are) adequate for survival, and not necessarily epistemical insights into reality.[58]

2. There is a fundamental problem about us looking at what constitutes consciousness and in particular selfhood: The epistemic irreducibility of 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.[59]

An example for the evolutionary pressure that made intuition just the way it is: When a huge grizzly bear comes charging at a caveman, he doesn't survive if he ponders whether he's just imagining or dreaming that, or whether it's actually a bear after all. The cavemen that survive are the ones that experience phenomenally perceived events as immediate and unquestionably real.

And genetically, we are still cavemen.

[56] Citation from [53], page 111.

[57] See [30], page 401.

[58] See chapter 1.1.

[59] This is because either we agree with the possibility and coherence of the explanation, but do that from a still experienced phenomenal first-person perspective that we cannot drop, thus forming a conflict that cannot be resolved without internal contradictions in our belief system. Or we abandon the first-person perspective completely in some form of enlightenment, but then (even if us humans are even able to reach that state) have no subject anymore that can be convinced of the truth of any such theory. See [52], pages 627f.

1.5.3 Reductive Naturalism

I will adopt a stance that can well be described as reductive naturalism, although I took a slightly different approach and called it a "rational-causalist world view" in [38] – which probably has less negative connotations.

This does not imply that I believe that our current scientific theories (or theoretical frameworks, or fundamental scientific entities) contain all the answers, or even that our current scientific paradigms[60] need to be the ones that prevail. But I do believe that scientific theories in a wider sense could in theory be formed that contain all the answers. It also means that I believe that some basic principles of science (mostly, conclusive logic, causal analysis and semantical coherence) will lead us to a better understanding of the world that we have just started to grasp.

Reductive naturalism does seem to have legitimation even when we look at the innermost workings of consciousness, our models of ourselves as thinking subjects:

> Today, it only seems safe to say two things: Phenomenal content as such is not epistemically justified content, and it locally supervenes on brain properties. A brain in a vat could generate the conscious experience of thinking *[I am certain that I* exist]*. Cotard's syndrome patients can generate the conscious experience of thinking *[I am certain that I* do not exist]*. It is therefore plausible that a minimally sufficient neural correlate for the phenomenal content characterizing the specific conscious experience of thinking *[I am certain that I* exist]* does exist.[61]

My reasons for this are very similar to Dawkins' that he clarifies like this, concerning evolution – I believe the maxim of his thought here should hold true for all aspects of our world views:

> 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.[62]

1.5.4 Legitimation for the Naturalization Project

Dawkins writes in another context: "Mystics exult in mystery and want it to stay mysterious. Scientists exult in mystery for a different reason: it gives them

[60] See [61], which is among my favourite philosophical works and always was a big inspiration for me.

[61] Citation from [52], page 404. I will look at I* thoughts briefly in chapter 4.5.3.

[62] Citation from [27], page 19.

something to do."[63] While what he wrote applies to religion, I do believe that it also applies to philosophy of mind.

The naturalization project that Dennett and Metzinger undertake, facing such strong opposition, is not one that attempts to make consciousness any less fascinating – as a matter of fact, knowing how exactly it works in intricate detail might well make it even more fascinating than when it has to resort to what I called "mystical causal powers"[64] and Dennett dubbed "a sort of alchemy."[65] Intrinsic or original intentionality,[66] from my point of view, is merely such an unnecessary mystification, a glorified gap where some philosophers attempt to go with St Augustine's advice:

> There is another form of temptation, even more fraught with danger. This is the disease of curiosity. It is this which drives us to try and discover the secrets of nature, those secrets which are beyond our understanding, which can avail us nothing and which man should not wish to learn.[67]

My heartfelt apologies, but I will stay curious about the world.

[63] Citation from [27], page 152.
[64] See [37], page 36.
[65] See [29], page 70.
[66] See chapter 5.2.3 for Dennett's dismissal of them.
[67] Citation from St Augustine, as cited in [27], page 159.

Part II
The Four Questions

When facing a really difficult 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. The hard problem of consciousness, the question how it can possibly be that our brain, being a mere clump of cells, can develop things like feelings, thoughts, plans and fantasies, is one such difficult problem, and it *is* possible to break it up like that.

As I mentioned before, there really are four major aspects to it. These four aspects are in some ways arbitrary – I have not found another such list that satisfied me. I do feel however 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 looking at 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 the the foundation for the next.

Of course, the four subjects cannot be looked at merely in isolation, but I will postpone looking at the bigger picture until chapter 6 – where we will also abandon these four parts again and look at the hard problem from a different perspective, ordering the resulting new parts with slight differences from the intuitions that led me to the order in this part here, too.

The partial questions for now thus are:

- *Phenomenality:* The special way things are experienced by agents like us. Of course, for a first-person perspective to exist, we do 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 will need in order to understand how consciousness happens. What is that phenomenality, and what entities or processes produce it?

- *Qualia:* The particular non-subdivisable entities of atomic quality we perceive when we experience things. 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 that 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 those smallest entities of quality?

- *Subjectivity:* The way we see everything from a first-person perspective. We do experience both ourselves and our fellow humans as persons, and we do believe that our fellow humans experience themselves as persons as well. How is such a first-person perspective generated, and what else is included when you speak 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 that

stands in a relation to these inner workings. How and why is that so? How comes that we feel we are in direct contact with an apple tree in the garden by means of some kind of "intentional arrow" or "mental focus"?

I will dedicate a chapter to each of these four parts, defining what really is puzzling about them and then looking at possible explanations for their origins and constitution.

When looking at the four big questions of what phenomenality, qualia, intentionality and subjectivity 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 shows up important counterpoints to those 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 (being a computer scientist and artificial intelligence researcher) comes from a more pragmatic and solutions-oriented background, bringing an interesting and fresh point of view into the debate.

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

- For *phenomenality*, Edmund Husserl's phenomenology pretty much was the origin of the idea, and has followers to this very day.
- For *qualia*, we will look at how 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, often apparently without the proponents knowing it themselves – as I exemplified in chapter 1.2.
- 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 this point.

2 Phenomenality

There seems to be something special about the way we see the world and ourselves, something that's maybe ineffable, maybe inexplicable or mystical, but

definitely hard to understand. I go ahead and call it "phenomenality", breaking with Dennett and Searle who include it in the notion of "intentionality" – the latter will be redefined in a much narrower way when I look at it in chapter 5.

I'd 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.

I will try to go another way here, showing what is special about phenomenality itself, the special nature in which we (and probably other conscious agents) see things, independent of what else there is to consciousness. 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.

2.1 Husserl: Systematic Reflection

Edmund Husserl started a radical paradigmatic shift with his (mostly unpublished, at least during his lifetime) works. 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 intuitively better explanation for the subconscious, although as we'll see it's not necessarily more correct), Husserl put the focus on how we might be able to describe our own phenomenal experience, and the immediate way events are given to our consciousness. Importantly, events are given immediate indeed, but occasionally only partially and inadequately – Husserl agrees that it is perfectly possible to misrepresent things.

The immediate directedness is given through three parts of an intentional act:[68]

- Immanent content, which is raw perceptive data.
- 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's not phenomenal. Particularly

[68] I already used this list in [38] – 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.

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 2.4.1.

2.2 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.[69]

Essentially, Dennett defines heterophenomenology mostly by what it's not: It's not subjectivist, but intersubjective. It's not first-person pilot studies, but third-person scientific research. And he claims that everything that would be phenomenology if it were really interesting 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.[70]

In the end, it comes down to the following: In heterophenomenology, phenomenological reports of individual agents are not 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: Collecting data and pointers for later research into the true nature of how consciousness happens:

[69] Citation from [32], emphasis his.
[70] Citation from [30], page 68, emphasis his.

> 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.[71]

Notably, for Dennett, heterophenomenology is not (unlike his notion of traditional phenomenology) 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 3.2.

2.3 Metzinger: Multilevel Constraints

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

The constraints are multilevel in that Metzinger attempts to capture all the important levels of description he makes out to be important for consciousness research.[72] These include the phenomenological level (gained introspectively, analyzed epistemically), the representationalist level (intentional content), the informational-computational level (computational function and goal), the functional level (purely causal properties independent of physical realization) and finally the physical-neurobiological level (direct neural corellates in usually human brains).

I will not go into 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.[73] 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 2.3.14, 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 introduce two other concepts that Metzinger improves our understanding of: Introspection, and how consciousness can't be anything but a process.

[71] Citation from [30], page 98.

[72] See [52], page 110.

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

2.3.1 Introspection

Metzinger makes it clear that there isn't just trivial introspection. Rather, there are varying degrees of introspective access to mental states – which results in phenomenally represented information.[74] Metzinger uses ordinals to differentiate them. They are:[75]

Introspection$_1$: "External attention" – a subsymbolic metarepresentation of the world model. Representing an internal system state, but referring to a part of the world. This "corresponds to the folk-psychological notion of attention."[76]

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

Introspection$_3$: "Inward attention" – a subsymbolic metarepresentation of the self-model. Representing an internal state that also refers to an internal state, a part of the experienced self-model.[77]

Introspection$_4$: "Consciously experienced cognitive self-reference" – a conceptual metarepresentation of the self-model. This generates self-knowledge and includes all situations in which we think about ourselves as ourselves.

While introspection$_3$ is just as 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."[78] Arguably, it is closest to our intuitive notion of "introspection."

I will however 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, even beyond the principial troubles concerning introspection that we saw in chapter 1.1.2.

2.3.2 Consciousness as a Process

With the brain being an information processing machine, mental states cannot be states in a traditional, rigid sense. Rather, they have to be reformed and reworked all the time through that process, in order to be experienced as transtemporally constant.

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

[74] See [52], page 31.
[75] See [52], pages 36f. This list is very close to Metzinger's.
[76] Citation from [52], page 36.
[77] See chapter 4.3 for how Metzinger fleshes out that self-model.
[78] Citation from [52], page 37.

> 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.[79]

The object of representation can obviously be an internal state or a counterfactual as well.[80] Also, not all entities in the world have to have a representation – as we will see in chapter 3.3, there are presentata just like representata. But the biggest 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 constant remodeling process in which our brains are constantly engaged.

> 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.[81]

Mental states are not the most fundamental entities that are experienced as ontologically given however – rather, phenomenal presentational content is. We will look at this a bit more in-depth in chapter 3.3.5. For now, it is important to realize that this presentational content is what drives the processes that 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.[82]

With these important foundations in place, let us now turn towards Metzinger's multilevel constraints, and try to understand them.

2.3.3 Constraint 1: Global Availability

> Phenomenally represented information is precisely that subset of currently active information in the system of which it is true that

[79] Citation from [52], page 20.
[80] See [52], pages 43f.
[81] Citation from [52], page 555.
[82] Citation from [52], page 98.

> it is globally available for deliberately guided attention, cognitive reference, and control of action [...].[83]

The availability is global because every part of consciousness has access to them. These three are indeed the basic kinds of global availability:

- 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 (that are then globally available as well).

In addition to these, speech control, autobiographical memory, phenomenal cognition and others are mentioned in the text – all of which are merely subcategories of the more basic three kinds however.

Every such availability will likely have different neural correlates, but their whole is experienced as embedded into one whole 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 flexibility, selectivity of content, and a certain degree of autonomy.[84] This makes the large number of specialized modules our brain has able to interact with those phenomenal states that are globally available – and 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 (more about this when we talk about qualia in chapter 3, we can't form a mental concept of "exactly this colour" for example), or occasionally not even for deliberately guided attention in early processing stages (like when we pull away a hand from a flame "instinctively").[85] Contents can be transparent[86] or opaque.

But 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 the way we experience it phenomenally.

[83] Citation from [52], page 117f.

[84] See [52], page 119.

[85] I will discuss different kinds qualia, with different degrees of global availability of sensory contents, in chapter 3.3.3.

[86] See chapter 2.3.9 for what Metzinger means with "transparency."

2.3.4 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.[87]

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"[88]), 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 have a concept of past or future presents, nor is it able to form a concept of coherence at all.

For human beings, the concept of Now has a culturally invariant duration of maximally three seconds, so there must be biological reasons for that – which nicely fits in with neurophysiological research. On the other hand, 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"[89] 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.[90]

2.3.5 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.[91]

This is the general case of constraint 1, global availability. Only if there is a coherent global state can there be something that is available within that coherent global state. This constraint is very fundamental: Being situated in a coherent representation of a world allows us to see the world naive-realistically as "my single world" from a first-person perspective. Since consciousness can't be anything but a process,[92] 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"[93], and that is pretty much what Descartes describes with

[87] Citation from [52], page 126, emphasis his – I will employ his capital Now to denote the concept he is talking about.
[88] Citation from [52], page 126.
[89] Citation from [52], page 128, emphasis his.
[90] See [52], page 129.
[91] Citation from [52], page 131, emphasis his.
[92] See chapter 2.3.2.
[93] Citation from [52], page 132.

the concept of indivisibility, the unity of consciousness. We aren't aware that it is only a model, a representation (more on this yet again when we talk about transparency in chapter 2.3.9). 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.

Having just one single world model is a pretty efficient way of reducing chaos from incoming sensory data, 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,[94] 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.

2.3.6 Constraint 4: Convolved Holism

> Consciousnes 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.[95]

Everything is part of the big world model, but within that, there are obviously substructures. These can even be multimodal, including vision, hearing, time, even social contexts. Metzinger calls them "levels of phenomenal granularity,"[96] 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."[97]

It is necessary for convolved holism to be possible at all that we do have one all-encompassing world model, and we experience all those parts and wholes as being in the phenomenal Now, so 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 that is so important for us humans.

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 that systems have to be able to react quickly to small shifts in their perception of their environment, or even themselves – so introspection is essentially nothing but a special kind of experienced convolved holism.

[94] See [52], page 136, where he refers to [46].
[95] Citation from [52], page 143.
[96] Nomenclature by[52], page 144.
[97] Citation from [52], page 145.

2.3.7 Constraint 5: Dynamicity

> Our conscious life emerges from integrated psychological moments, which, however, are themselves integrated inot the flow of subjective time.[98]

When convolved holism was 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 past and future Nows. 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.[99]

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 that we experience is always perceived as being a property of such experientally presented, dynamical objects.

It is not easy to form a concept beyond this, and intuitions seem to differ widely. 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.[100]

2.3.8 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,

[98] Citation from [52], page 151.

[99] Citation from [52], page 153, emphasis his.

[100] See [52], 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."

> phenomenal *self*? Second, what is a consciously experienced phenomenal *first-person perspective*?[101]

Metzinger does point out how this is among the most interesting phenomena when it comes to phenomenology, and I'm tempted to agree. It was among the chief things that made me read Metzinger's work after having been disappointed when it comes to exactly what subjectivity actually is after reading Dennett – the latter merely glosses over the functional aspect of this interesting point (as we'll see in chapter 4.2), not offering a real solution to the problems it poses.

Metzinger makes out several levels on which perspectivalness comes into play:

1. The phenomenology of being someone, of consciously experienced selfhood. Part of this is what the phenomenal self-model that we will look at in chapter 4.3 is all about; "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."[102]
   A self-model (potentially being both cognitive and conceptual) necessarily requires a more basic automatic and fundamental self-consciousness however, as we will see in chapter 4.5.2. This proto-self is the other part of this phenomenology of being someone, and the prime prerequisite for the PSM.
   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 – a point where Dennett would probably disagree, since he does not believe in irreducibility in any form.

2. 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*"[103] intentionally, which is what we will talk about in chapter 5.3 – the phenomenal model of the intentionality relation.

3. Social cognition (the untranscendable "we") is only possible if there is a phenomenal first-person perspective (the untranscendable "me"). We will talk about this in chapter 4.5.4, and it is the third level of perspectivalness.

Functionally, there is an embodiment constraint in all conscious organisms we know to date: They are experientially centered in that they only have one body and limited reach, so 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*."[104]

[101] Citation from [52], page 157, emphasis his.
[102] Citation from [52], page 158, emphasis his.
[103] Citation from [52], page 159, emphasis his.
[104] Citation from [52], page 161, emphasis his.

Interestingly, perspectivalness does show up among the other constraints here, but actually partially follows from some of the other constraints by necessity.[105] While I do see the justification for placing it among the other constraints, I do believe that Metzinger could have stated this more clearly in his book.

While Metzinger does list 11 constraints, they are actually merely 10 plus the aforementioned prereflexive proto self-model, and the PSM and PMIR follow from them and form the perspectivalness constraint.

2.3.9 Constraint 7: Transparency

> Transparency [...] is a property of active mental representations already satisfying the minimally sufficient constraints for conscious experience to occur. [...] The second defining characteristic postulates that what makes them transparent is the *attentional unavailability of earlier processing* stages for introspection.
>
> [...] For any phenomenal state, the degree of phenomenal transparency is inversely proportional to the introspective degree of attentional availability of earlier processing stages.[106]

Metzinger's definition of transparency is different from other philosophers' definitions. The standard definition only defines that vehicle properties of certain mental states are not available for introspection whereas content properties are. But as Metzinger points out, there clearly are cases where 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"[107], which makes definitions of transparency using those 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."[108]

Metzinger goes on to describe three common potential misunderstandings of the concept:[109]

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 faults when thinking about one's own consciousness, exactly because the core processes are hidden from introspection.

[105] Deducing this is what his chapters about the PSM and the PMIR, and consequently chapters 4.3 and 5.3 in this thesis, are about.
[106] Citation from [52], page 165, emphasis his.
[107] Nomenclature by[52], page 166.
[108] Citation from [52], page 170.
[109] See [52], pages 166ff.

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 probably has, because transparency will end up being a prerequisite for phenomenality in chapter 2.3.14) transparent mental states.

3. 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"[110] 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*."[111] These kinds of technical transparency are however not phenomenal and thus (at least until this changes) not interesting for the discussion at hand anyway.

Transparency does seem to imply the existence of the entities represented. This leads to the fact that a naive realism concerning those entities is intuitively very attractive for us humans by necessity – as it is the way we experience the world. We don't see what we're looking through, the medium that the experience takes place in is indetectable for introspective access.

Metzinger calls transparency "a special form of darkness."[112] The opposite of transparent states would then be opaque ones – they are the ones where darkness is made explicit, where we represent that something is merely a representation, 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.

In opaque cases of experience, we know while we know that we could be wrong, as opposed to knowing while we know that we know in standard, transparent cases of experience.

2.3.10 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.[113]

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

[110] I'm using quotes here because there is no subject that could mean to transcend them.
[111] Citation from [52], page 168, emphasis his.
[112] Citation from [52], page 169.
[113] Citation from [52], page 179, emphasis his.

bodily actions and their bodily consequences,"[114] independent of its sensory organs.

This enables both memory and future planning ("internal representation of goal states"[115]) and generally counterfactuality (possibility versus reality) and internal experiments. The implications of this are vast. 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:

> 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*.[116]

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.

2.3.11 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.[117]

Metzinger explicitly excludes metaphorical uses of intensity relations and only includes those associated with the experience of qualia.[118] 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."[119]

In some cases, it is even possible to experience intensity without associated dimensions. For color, that would be hue and saturation – Metzinger refers to Ganzfeld experiments,[120] where subjects experience a colorless, formless visual experience after looking at a uniform field of color (said Ganzfeld) for some minutes.

[114] Citation from [52], page 183.
[115] Nomenclature by [52], page 181.
[116] Citation from [52], page 180, emphasis his.
[117] Citation from [52], page 184.
[118] See chapter 3.3 for Metzinger's distinction between different kinds of qualia.
[119] Nomenclature by [52], page 185.
[120] See [52], page 101f, where Metzinger refers to [40].

What is not possible however is that there is no analogue representation of a certain stimulus, for 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 zero.

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.

2.3.12 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, structureless *density* of simple phenomenal color experiences and the like that has traditionally supported antireductive theoretical intuitions.[121]

This constraint is also called "Ultrasmoothness". What Metzinger talks about with this constraint is what Sellars called the grain problem[122] – the ultimately uniform experience of things like color in a certain area, an absence of internal structure. Sellars states this problem:

> Putting it crudely, colour expanses in the manifest world consist of regions which are themselves color 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.[123]

However, as Metzinger points out, the entire area in question appears as directly given, it can well be called atomic itself and not further divisible. As soon as we start thinking about subregions, the concept of those subregions forms (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."[124] So there isn't really a grain problem, because the experience of such surfaces itself

[121] Citation from [52], page 189f.
[122] See [52], page 189, where Metzinger refers to [66] and others.
[123] Citation from [66], page 26, as cited in [52], page 190, emphasis as in [52].
[124] Citation from [52], page 192

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. 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,"[125] processes that lead to a coherent percept 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 doesn't have to be, split into

2.3.13 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.[126]

A common intuition when it comes to non-natural intelligences (artificial, postbiotic, any kind really) is "But *none* of these things is ever going to have genuine *emotions*!"[127]

The underlying issue is that artificial systems do not require any teleofunctionalist properties; their goals and ideals (or rather, the corresponding representative mental states) don't necessarily have to make sense. Here is one of the reasons why I stressed the importance of evolution in chapter 1.4: Evolution sees to it that the development of goal states that 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 necessarily make sense.[128] 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.[129]

This, by the way, should in no way trivialize the what-is-it-likeness of emotions. 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

[125] Citation from [52], page 195.
[126] Citation from [52], page 198, emphasis his.
[127] Citation from [52], page 199, emphasis his.
[128] 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 [25] and [16]) makes some of the goal states that were a result of genetic evolution occasionally outdated.
[129] That doesn't 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.

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.[130]

I will discuss the adaptivity constraint further in chapters 2.4.4 and 2.4.5.

2.3.14 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 do dodge how hard the hard problem actually is, 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[131] of consciousness and phenomenality. And we don't need to be afraid of solutions that appear to be counter-intuitive.

It is clear that while us humans 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 makes out which of these constraints have to be satisfied for a system to be called conscious to what extent:[132]

- *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

[130] Citation from [52], page 200f.

[131] See [5], where emergence is defined by means of a citation from Lewes in [47], 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."

[132] This list is very similar to the one in [52], pages 204ff.

content here yet). This is equivalent to "the presence of a world"[133] – no subjectivity, no differentiated representation of causality, space or time, 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."[134]

- *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. Arguably, many animals are on this stage of consciousness.
- *Cognitive, Subjective Consciousness:* Here, we add offline activation (8), assume 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"[135] – and decouple those from current external input. Such systems could also represent themselves as representational systems, would be true thinkers of thoughts, and able to escape naive realism at least in thought experiments.

Note how 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,[136] which might be necessary for us to even recognize a system as being conscious at all.

Homogeneity of simple content (10) is interesting, because as we found out 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 think I would add this constraint as a requirement for differentiated consciousness already.

2.3.15 Phenomenality for Metzinger

What really makes phenomenal experience special is its ineffability:

> 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

[133] Citation from [52], page 204.
[134] Citation from [52], page 204.
[135] Citation from [52], page 205.
[136] Or at least the weaker adaptivity that I suggest in chapter 2.4.5.

> 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.[137]

Ineffability then comes naturally with the combination of at least perspectivalness and transparency, with other constraints (like global availability, activation within a window of presence, or convolved holism) making experiences only more ineffable.

Phenomenality, just like consciousness, is not an all-or-nothing attribute of systems. It seems to me 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.

2.4 Consolidation

I think it's fairly clear, considering how much space I gave the three different points of view regarding phenomenality: I am fairly partial to Metzinger's views. I think that he offers the first true explanation of consciousness in general (despite Dennett in [30] 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."[138] Nevertheless, it seems to be all we have for now, and I believe that it's quite a good start.

2.4.1 What Husserl Might Answer to Dennett

We saw in chapter 2.2 what Dennett's biggest gripe with traditional phenomenology is: It is not necessarily intersubjective. He makes this rather 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.[139]

However, I am not sure if Dennett's position is really so different from Husserl's when it comes to what is science and what isn't. Maybe Heidegger's continuation of Husserl's theories, putting way more focus on how phenomenality is a subjective phenomenon, would fall victim to this argument of Dennett's. Husserl however would probably even agree with him:

[137] Citation from [52], pages 94f.
[138] Citation from [52], page 117.
[139] Citation from [32].

> 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.[140]

So individual phenomenal experiences only gain a scientific meaning if they are also looked at in the context that the subject experiencing them was in – and, it is important 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:

> 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.[141]

So, considering these passages from Husserl, I am uncertain as to what it is that actually makes heterophenomenology so different from phenomenology. Both Dennett and Husserl agree that we have some (more or less limited) privileged access to our mental states. They both agree that proper scientific research can only be performed from a third-person, intersubjective point of view, taking phenomenological accounts not at face value and forming theories as to what it is that makes agents have those particular phenomenal experiences.

Considering neither of them actually offers more insight into what phenomenal experience itself is however, I want to leave this side scene, and will 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 get to make autophenomenological reports in the first place.

[140] Citation from [41], page 224. My translation: "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."

[141] Citation from [41], page 231. My translation: "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."

2.4.2 Another Look at Global Availability and Convolved Holism

When Metzinger talks about his constraints global availability (1) and convolved holism (4), he does look mostly at the phenomenal side of it. Minsky agrees with him in that things certainly *seem to be* globally available, and quotes Newman:

> [In the Global Workspace theory] The theater becomes a workspace to which the entire audience of "experts" has potential access ... Awareness, at any moment, corresponds to the pattern of activity produced by the then most active coalition of experts, or modular processors. ... 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. ... Each expert has a "vote", and by forming coalitions with other experts can contribute to deciding which inputs receive immediate attention and which are "sent back to committee". Most of the work of this deliberative body is done outside the workspace (i.e., non- consciously). Only matters of central import gain access to center stage.[142]

However, we cannot push global availability too far:

> 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.
>
> 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.[143]

So while it certainly may *seem* like we have 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 (that Minsky calls "critics" and "experts," but unfortunately we do not have the space to look at these more in-depth) have access to parts of these processes and can potentially further process them.

However, this is neither magical nor entirely global, but rather, what brain centers have access to what processes is stored in the dynamical hardware[144] of the brain. I say this without claiming that domain borders between different

[142] Citation from [57] as found in [53], page 125.
[143] Citation from [53], page 345.
[144] 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.

processes can't be crossed – quite the opposite, that possibility is indeed (as Metzinger claims) one of the prime advantages of our kind of consciousness. Rather, these domain borders can only be crossed in certain ways that are predetermined in the same dynamical hardware that implements the processes themselves.

That certainly doesn't 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.

2.4.3 The Trouble With Perspectivalness

Metzinger's perspectivalness constraint (constraint 6) seems to be fairly straightforward. It is a first-person point of view that is nonconceptual, happens before cognitive processing, it is automatic and requires no effort whatsoever. It consists of a nonconceptual prereflexive proto self-consciousness, a PSM[145] and a PMIR.[146] If the prereflexive proto-self is there, we will be able to naturally reach the agreement that a full-fledged self-model has to arise from the system.

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

I have no solution to offer to this, and I guess it will require empirical research with actual postbiotic or artificial systems to truly progress in this matter. It is certainly a worthy topic for future research.

2.4.4 Adaptivity Questioned

Metzinger heeds another warning, when it comes to adaptivity: He says that in order to be truly intelligent, a postbiotic or artificial system also has to fully satisfy the adaptivity constraint. He even goes further and quotes Davidson's "The Swampman,"[147] with an argument similar to the ship of Theseus. 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 and everything).

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

[145] See chapter 4.3.
[146] See chapter 5.3.
[147] See [52], page 206.

> [...] 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 [...].[148]

I'll be very blunt: I do not believe this argument has much merit. If it had,[149] 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 does not have the right kind of history built-in. All those things that we take as signs of proper intelligence beyond basic empathy (including being able to play chess, to talk with each other in a common language, to read and write) 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.[150]

Thinking about it, humans are necessary to trigger the second-level evolutionary process that eventually leads to childbirth as well.[151]

2.4.5 A Weaker Adaptivity Constraint

When Metzinger formulated his adaptivity constraint, he seems to have been mainly driven by one of his background assumptions, one that he explicitly states at the very beginning of his book: Teleofunctionalism.[152] He introduces his commitment like this:

> [...] 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

[148] Citation from [52], page 206.
[149] I do owe this point to my good friend Simon Bünzli.
[150] Citation from [52], page 207.
[151] I like the take of the movie "Robots" from 2005 on this. Herb Copperbottom: "Making the babies is the best part!"
[152] See [51].

> 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.[153]

Teleology may not be the best approach for everything (and in fact, when Sehon used it along with the term "supervenience" that Metzinger also uses, he constructed a theory that I can't find myself agreeing with at all). But when Metzinger uses it, he seems to mean that he looks at the things our brains do in a way that assumes they must be good for something – because if they weren't, they wouldn't have been a fitness criterion for evolution, and consequently wouldn't have survived the last few million years.

I do agree with that. I do not agree with million years of evolution being 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) it suddenly loses some kind of metaphysical fitness *pre*condition. But that 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 reminescent of dualist theories despite all our efforts to banish it.[154]

I do not deny the importance of evolution for today's kind of intelligence. It was 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 has. "Fitness for purpose" however can be, even if we're talking about intuitively non-replicable things like emotions (that, essentially, are nothing but "logic of survival" indicators) or qualia – and as such, I'd 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 proper consciousness.

Essentially, I want to take Metzinger's normative approach to what good representata are and slightly alter it. It is:

> Phenomenal representata are *good* representata if, and only if they successfully and reliably depict those causal proerties of the interaction domain of an organism that were important for reproductive success.[155]

[153] Citation from [52], page 26, emphasis his.

[154] This is particularly odd because Metzinger is quite good at finding "subtle residues of Cartesian dualism" elsewhere, see chapter 2.3.9.

[155] Citation from [52], page 203, emphasis his.

And, the change I propose is simple:

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

2.4.6 Rational-Causalist Phenomenology Reconsidered

In [38], 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. Those 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 (that are in turn created by things in the world that enter the "mental space" via physically analyzable pathways, for example light that hits light-detecting nerve cells in the eye), we can causally analyze their interdependence, and form models of this directedness, because the kinds that are now directed at each other are of the same kind.

I don't think I deviated that much from Husserl's ideas with those suggestions even.[156] Husserl realized that there is a cognitive closure, that we cannot perceive what Metzinger would call transparent properties of mental events. Husserl merely didn't attempt to look behind what is introspectively accessable because he thought there would be no point in attempting to explain something where we do not even know what the explanandum itself is. And he did not attempt to form an ontological theory, but rather an epistemologically plausible one.

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

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

3 Qualia

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

Examples are distinct shades of red and green in a picture we look at, or the specific nature of the sound of a piano, or the particular way a glass of good wine tastes, or the enjoyment we can get out of a particular way sunlight is cast through treetops in a forest. Dennett makes an example of the plethora of qualia that 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 anyway:

> 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?[157]

Well, how could it? Let's see.

3.1 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

[157] Citation from [30], pages 26 and 406, emphasis his.

> 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.[158]

Qualia for Lewis are ineffable[159] 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.""[160]

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 (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 do know which qualia we experience. However, it is possible that the concepts the experienced qualia relate to are wrong and (while part of our world model) not part of the external world. As Metzinger points out,[161] Lewis qualia are available for attentional, behavioural and cognitive processing.

This work of Lewis 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. We do face huge problems when we attempt to imagine what it is like to be a bat:[162]

> 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,

[158] Citation from [48], pages 60f.
[159] See [48], pages 123f.
[160] Citation from [48], page 124.
[161] See [52], page 84.
[162] It is well worth noting that Nagel was not anti-physicalist, nor anti-reductionist, even-though this quote seems to imply this – he 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 [55], 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. I do believe that it is possible to explain this on a functional level, as well as through neurophysiology that Nagel himself admitted faces no such obstacles. Nevertheless we indeed do face obstacles right now, because in our current position, we have to rely on empathy for emulation and proper appreciation of other minds.

> or by imagining some combination of additions, subtractions, and modifications.[163]

We will confront Lewis' notion of qualia with Ruffman's and Metzinger's in chapter 3.3.3. Nagel's observations, I believe, 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 that pretends to be a bat.

3.2 Dennett: No Qualia Whatsoever

Dennett adopts quite a radical stance when it comes to qualia: For him, they (at least in their traditional reading) don't exist.[164] There is nothing irreducibly, inexplicably special about atomic entities of phenomenal qualities of subjective experience. He proposes a fresh start, one without the history of the discussion about qualia and their special atomic 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.[165]

Let us look at what that Multiple Drafts model is, and see how Dennett believes it can replace qualia.

3.2.1 Orwellian vs. Stalinesque Revisions

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

[163] Citation from [55], page 3.
[164] See [28].
[165] Citation from [30], page 370.

never delivered, subjects feel all five wrist taps at the wrist in the expected manner.[166]

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

- *Orwellian revisions:* Like the Ministry of Truth in Orwell's novel 1984[167] rewrites history all the time: This theory would suggest 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 gouvernment misrepresented facts through the censored media to its populace, complete with false evidence and elaborate productions: This theory would suggest that 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 (formed, as Metzinger rightly emphasizes, by ever-changing physical representations where the current form of the vehicle is actually part of the content) just like long-term memory.

3.2.2 Multiple Drafts

The fact that both Orwellian revisions seem to be clearly correct in some situations,[168] while Stalinesque revisions seem to be clearly correct in others,[169] it appears that the answer must lie somewhere in the middle (yet again). Furthermore, what we've seen in chapter 1.2.3 (whihc is that there is no Cartesian theater, and mainly the "spatiatemporal smearing of the observer's point of view in the brain"[170] that Dennett finds) 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 exactly those two observations and finds a very elegant solution to the problems they pose to dualist or epiphenomenal views:

[166] Citation from [30], page 143.
[167] See [58].
[168] 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.
[169] 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 colour inbetween. Whenever they were probed, there was no experience of flashing and not moving dots, at least if they were flashing with a certain minimum speed.
[170] Citation from [30], page 126.

> 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."[171]

This leads to all kinds of things 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.[172]

The great thing about this theory is that it stands up to empirical scrutiny, and that timings for example for processing of visual inputs can be found with experiments. Judgements can be distinguished between "the mere onset of stimulus, then location, then shape, later color (in a different pathway), later still (apparent) motion, and eventually object recognition."[173] And all those judgements don't necessarily come together somewhere 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), but are rather directly available for further processing and potentially actions.[174]

Dennett goes further and even suggests a few more things:

- There is no "optimal time of probing,"[175] since while there can be a version that is stored in long-term memory, earlier probings (when more parallel processes are still running) usually yield more information about the specific nature of the running processes and their "mental contents."
- 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

[171] Citation from [30], page 111.
[172] Citation from [30], page 113.
[173] Citation from [30], page 134.
[174] Metzinger would of course distinguish between different kinds of availability here, as we saw in chapter 2.3.3. Dennett's focus lies elsewhere of course, but I think it is important to see where their two theories can benefit from each other – this to me seems to be one such place.
[175] See [30], page 136.

Jane came earlier than both of them"[176]) – there is the possibility of backwards projection of perceived events in time.[177]

Being able to spin a narrative is what "a something it is like something to be"[178] is all about for Dennett. A portion of the world becomes an observer as soon as it starts composing a skein of narratives.

3.2.3 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 is that feature detectors that make us experience something as "blue" or "bitter" co-evolved with the features they detect. Color vision evolved because there were features distinguishable by colour, and vice versa, starting small and growing from there - and the fact that a butterfly's wings have the same blue as cobalt does is mere coincidence. "Why is the sky blue? Because apples are red and grapes are purple, not the other way around."[179]

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 we detect with color vision, is to say, uninformatively, that we detect the properties we detect.[180]

This doesn't explain the particular way things feel for us. The common term that describes what makes this feeling special seems to be "enjoy" – it stresses the difference between neuroanatomy and experience, between information and

[176] Citation from [30], page 137.

[177] A particularly striking example and one of my favourites among the experiments Dennett brings forth to support his theory is one by Libet that he introduces in [30], pages 154ff (of which alledgedly a reproduction was never attempted) – see [49], where Libet also arguments against opposition he faced from Churchland.

In that experiment that was 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 [30], 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.""

[178] See [30], page 137.

[179] Citation from [30], page 378.

[180] Citation from [30], pages 382f.

qualia. Qualia (the ones Dennett doesn't support) cannot be divided into mechanical parts, while mere information can be split up without losing anything. Dennett believes that Qualia don't 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 explainable in their entirety, Dennett proposes another thought experiment. He has people want to experience Bach as if it was the very first time they were, without being tainted by more recent musical developments or other advances in the meantime. He goes to claim:

> 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.[181]

Another interesting point is how inverted qualia don't seem to exist: People who wear goggles that turn the world upside down (all the time) adapt to that after some time, and are able to ski downhill or drive bicycles through city traffic. 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 places in the brain will have different views on that question, and some (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. Parts of the brain will necessarily disagree with each other whether it's adaptation or mental content modification.

Essentially, Dennett's position boils down to this: It is, at least in theory, possible to describe any experience in its entirety, and that 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 would be a necessary part of our arsenal for investigating consciousness.

3.3 Metzinger: Phenomenal Presentational Content

Metzinger, like Dennett, doesn't believe that the classical concept of qualia holds up to scientific scrutiny. He even 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."[182]

The traditional notion of qualia for Metzinger 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

[181] Citation from [30], page 388. Note the wording, I will argue in chapter 3.4.1 that appreciation is more than merely description.
[182] Citation from [52], page 64.

subjective states. Judgements about qualia cannot be false, because they are defined through the subjective determination that judges them.[183]

It is under these background assumptions that we will now look at his argumentation.

3.3.1 Qualia are Inefficient

From a computational point of view, it makes sense that not all experiences have such 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.[184]

This is of course a teleological observation that does not necessarily mean that qualia do not exist themselves. Metzinger however does have an argument for that.

3.3.2 Most Simple Forms of Content don't Exist

To the end of showing that there is no such thing as always implied most simple forms of content when we talk about mental events, Metzinger 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. While we can certainly distinguish two particular, just noticeably different shades of red when faced with them (they are available for attention), we cannot identify or recognize them transtemporally. Metzinger thus concludes:[185]

> 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.[186]

[183] See [52], pages 66ff.

[184] Citation from [52], page 79.

[185] He refers to [62], where 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 as per [52], pages 69f.

[186] Citation from [52], page 70.

This leads to some obvious conclusions. This maximally simple content as such does not exist, because there is no way for us 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*.[187]

However, unlike Dennett, he 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) do exist[188] – but they do not exist for all kinds of experience, and thus they can't possibly be the basis of all phenomenal experience.

3.3.3 Lewis Qualia, Raffman Qualia, Metzinger Qualia

Metzinger looks at the discussion about qualia to date, and notices how there really are entirely different readings of the term that are often used interchangeably. He offers the following distinction, looking at them in descending order of number of constraints they satisfy:[189]

- 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 less real than Lewis qualia, and Metzinger qualia are even less real than Raffman qualia. Nevertheless, there are examples for all three kinds; 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."[190]

[187] Citation from [52], page 72.
[188] See [52], page 73.
[189] See [52], page 74 for Lewis and Raffman qualia, and page 75 for Metzinger qualia.
[190] Citation from [52], 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

There are more logical alternatives, only some of which make sense for us humans (for example, the instinctive movement when we are catching a ball is available for motor output, but not necessarily for cognition or attention). The ones mentioned here however are "identifying the phenomenologically most interesting terms."[191]

3.3.4 Qualia are Reducible

For Metzinger, qualia (understood as noncategorizable, but attentionally available forms of sensory content) are reducible in principle. He suggests going about that by finding the physical structures that sensory content supervenes on, and by functionally analyzing the causal role of these structures.[192]

However, he does not banish them from the ontological landscape entirely, he even states that "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."[193] But they are not the most simple form of mental content – rather, *presentational content* is the most simple form of mental content.

3.3.5 Phenomenal Presentational Content

Presentational content is nonconceptual, cognitively unavailable, homogenous and fully transparent. This is what it means that it appears to be ineffable. It also 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 that is able to explain what seems to be so special about qualia, while it is always incorporated into a higher-order whole.

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

Metzinger goes beyond traditional models and finds 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."[194] Instead, he offers this view:

---

philosophische Steckdose gesteckt." ("Andrea Borsato in the flow of thought, feels as if he had stuck the finger into a philosophical power outlet.") That's the kind of feeling Metzinger must mean.

[191] Citation from [52], page 90. The three 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?

[192] See [52], page 85.

[193] Citation from [52], page 86.

[194] Citation from [52], page 93.

> 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.[195]

This is a rather mathematical description. However, the idea that qualia as such don't exist, but rather every recognizable quale corresponds to a (non-divisible by introspection) volume in quality space, does make sense and also takes into account that there certainly are differently wide or narrow interpretations of, for example, redness. That quality space then is continuously regenerated in a dynamical process, and no actual entities are generated - rather, ceteris paribus, always the same input drives always the same presentational processes (that are then experienced as always the same objects being perceived in the same ways).[196] As such, presentational content is always temporal content, too, in that it presents what happens right now.

Presentational content like that has interesting properties. Phenomenal states generated in this way are particularized by three important implications:[197]

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

2. They are fully transparent.

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

For Metzinger, the approach (and particularly the order of the chapters) I took here 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 order will be one of the main changes I will make for the order of partial concepts in the summary in chapter 6.

[195] Citation from [52], page 93, emphasis his.

[196] 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 classical music.

[197] Adopted from [52], page 95.

3.4 Consolidation

Qualia were a somewhat diffuse topic when they were first introduced, and they haven't gotten that much clearer since. Dennett felt this was sufficient to totally abandon the term, while Metzinger on the other hand gives them ontological legitimation, while still attempting to analyze them and saying that there is nothing atomical about them – they are mostly merely paradigmatic definitions of arbitrary abstraction. Let us look closer at how exactly their views differ. They both share some similarities:

- Both concepts of qualia are in principle reducible.
- Both concepts of qualia have distinct ways things seem to feel.

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

- Dennett abandons the term "qualia" altogether, while Metzinger gives the term "qualia" an ontological legimitation, namely, that of a specific categorizable and transtermporally comparable volume in quality space.
- Dennett's concept of what traditional qualia are is (in addition to reducibility in principle) also practically reducible into purely syntactical terms, stating 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, that can't be intersubjectively specified or conceptualized.

Metzinger does allow an intersubjective view at qualia, and even presentational content: By means of the neural correlates they supervene on, and their functional role. However, the specific what-is-it-like-ness is only available from the first person point of view. Dennett on the other hand says that all such experiential content is not only intersubjectivizable, but understandable in its entirety by any other conscious system, even if it has completely different dispositions and experiential contents.

3.4.1 The Difference Between Appreciation and Description

Essentially, Dennett postulates that appreciation (a word he even uses himself, see chapter 3.2.3) is the same as understanding a description – that a description is able to grasp all aspects of an experiential content, and that reading such a description enables a conscious system to "relive" (which is the proper way to read "appreciate," I believe) those contents.

Dennett disagrees with Nagel and believes that it would be perfectly possible for us to know what it is like to be a bat. This implies that both the description can completely capture the experience, and that reliving based on such a complete description is possible. Let us look at the citation from chapter 3.2.3 again:

> 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.[198]

I'm not sure I can agree with Dennett. Metzinger's view looks much more reasonable to me. There simply is mental content that is available for attention, but not concept formation. And that mental content itself cannot be made intersubjectively available – the best we can ever hope to achieve is to map which neural correlates the ones it supervenes on are.[199] Mapping the functional roles would be necessary as well, but would be even more complex, I doubt that 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[200]) we are talking about, not a genotype – so even completely understanding the human brain[201] in general would not be enough.

And if we would ever obtain such a complete description of all functional roles and neural correlates of a certain experiential content, the only way to "relive" it would be to have exactly the same neural correlates activated in the same way with exactly the same functional roles. In the end, this means that a brain that 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.

This goes so far that we can't fully appreciate tales about our childhood and, arguably, some childhood or otherly faint memories themselves, even if we don't have to be reminded of them by third parties. William James makes such an 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.[202]

3.4.2 Dennett's Multiple Drafts and Qualia

Dennett's multiple drafts model seems to be a pretty adequate description of what happens in the brain, and really is a replacement for cartesian theaters.

[198] Citation from [30], page 388.
[199] Note that due to the fact that the brain rarely does only one thing at once, even this task will be hard – although not impossible, probably best achievable through differential analysis of a series of tests.
[200] See [26].
[201] Assuming we are talking about the kinds of qualia that specific humans have.
[202] Citation from [42], as cited in [53], page 308.

It takes the massive distributed parallelism in our brains into account, and computational models based on it (or rather, based on the same thoughts and paradigms) do live up to what we expect from them – a great 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 (that, in his case, consists of visualization and probably logging).

Dennett does however also present his model of multiple drafts as an alternative to the traditional theories of qualia. Let us look at how exactly it is an alternative that can explain the same things however, and whether the theory of multiple drafts necessarily implies that there are no (non-eliminable) qualia.

In order to make the transition from qualia to the theory of multiple drafts, Dennett first observes that one of the prime motivations behind the concept of qualia was that the entities that qualia would have to be would only be "in the eye and brain of the beholder,"[204] not out there in the world ready for us to observe. But, if they are in our brains, there is no inner figment that they are made of either; Dennett straight out denies that things like "occurrent pink" would need to be included in popular science – because unlike Sellars,[205] Dennett (just like Metzinger) isn't opposed to the idea that us making judgements about a colour can be all there is to colour vision.

This implies that the colour is out there somehow after all. It is there ready to be presented to us, and our detectors that are specificially tailored to detecting them do present them, which then leads to representations within our brains.

This then is where the M-detectors and M-properties come into play – obviously, if entities (or properties) are out there ready to be observed, eventhough we might not be able to clearly define them by their physical description alone, there must be detectors that reliably find those entities. And how these detectors are built at their core is what the multiple drafts theory is about. It does perfectly explain 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.

But what Dennett doesn't face is the ineffability of qualia, the distinct way they look to us. He doesn't aim for appreciation, but merely description. More importantly, he also doesn't face 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.

So his theory of multiple drafts really 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 (but processed in parallel, except where one process has to wait for the results of another) procedure that finds "easy" properties first and "hard" properties later.[206]It is, however, not a theory that really explains

[203] See [69], and in particular [70], as referred to in [45].
[204] Adopted from [30], page 370.
[205] See [66], as referred to in [30], page 372.
[206] Note how "easy" and "hard" don't have to correspond to the perceived complexity of the processing steps involved, but rather with how optimized the system in question is for solving

qualia.

3.4.3 Qualia: Still Necessary?

So with Metzinger, we find that 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 shows up, and we are able to conceptualize volumes within that quality space transtemporally. 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.[207] Whether they are primary or secondary however, there are ontologically legitimate[208] qualia of "redness" that all kinds of color perception that are red (within distinct, certainly individual boundaries) fall into.

So indeed, qualia are at the same time reducible and still 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 elsewhere anyway though. And, they are what our memories are made of.

---

them. Things like detecting that the shape in front of us is a face that is looking to the right and belongs to your grandmother is a very complex and complicated thing, while playing chess is comparably easy. Just because something seems easy to us does not make it easy for another system – it's just that our bodies have adapted to needing such face detection mechanisms for millions of years, while the ability to play chess only has formed in the last couple of thousand years.

[207] Colour for example appears to be a secondary quality of things such as shape and motion – see [50], as referred to in [30], page 371.

[208] With the reservations that I made in chapter 1.1.

4 Subjectivity

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

For example, if you think about something you did as a 4 year old kid – be it recalled 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 4 year old kid 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 way more experiences we can relate to, different theories regarding both the world around us and ourselves. But it is absolutely certain that that 4 year old kid was still you, however many differences there might be.

Another example is intentionality (which we'll look at in chapter 5) – when thinking about directedness, there necessarily has to be an origin that the arrow of directedness comes from. That origin is always the same, an untranscendable "me" that stays constant throughout all our experiences.

This will be where we incorporate some 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 – and often, the models we have of others are (or at least seem to be) more accurate than the models they have of themselves. How can that be? Shouldn't I be the final authority when judging what it is that makes me myself? Don't I have priviledged access that should enable me to form a more accurate, more proper self than others, who merely form mental (empathical) models of me?

Interestingly, there are many deviant forms of subjectivity, and one of them is a recurring theme among Metzinger, Dennett, and many other philosophers of mind: The multiple personality disorder. Of course, because if we want to see what a self is, we best look at a place where it is already partially dissected, where its inner workings are more prone to lay bare because the protective layer around it has been removed and broken – metaphorically speaking, but often 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 driving forces behind me writing this thesis. As I admitted in [38], I had a hard time wrapping my head around what constitutes a self for quite a long time. It is one of the most fascinating aspects of the philosophy of mind to me to this day. However, I think that I am closer to understanding this now, and I hope that I can show up why that is.

4.1 Descartes: Dualism

We already talked about Dualism in chapter 1.2. 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 it is that creates subjectivity.

This is because the very distinguishing criterion for the *res cogitans* is that it is able to produce that very subjectivity. As such, subjectivity is built into the *res cogitans*. Of course, Descartes, being deeply religious, roots this 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.

So, since we are both unable to physically analyze a purely non-physical entity, and unable to introspectively dig deep enough to expose the metaphysical processes that would expose subjectivity itself, that is pretty much all there is to say on this matter:

Subjectivity is that which defines and creates the *res cogitans*.

4.2 Dennett: Center of Narrative Gravity

Dennet points out how there are two contradictory stances when it comes to selves:[209]

- 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 over and above our brains.

These stances are contradictory, but both intuitively plausible, so the most reasonable conclusion is that they must both be partially true. This will probably lead to a conclusion that is not as intuitive[210] as the two premises we just saw, however.

4.2.1 On Evolution and Absolutism

Selves must have been created through evolutionary processes: The first single-cell organisms couldn't have had selves, and today we do have them. So obviously organisms with selves evolved from organisms without selves. 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" – porously and with less clear boundaries than we would imagine, as there are synergies and symbiotic relations with some micro organisms in our stomach, for example.

[209] See [30], page 413.
[210] See chapter 1.5.2.

Dennett talks about the extended phenotype,[211] 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 us human beings include a self in that extended phenotype is fascinating:[212]

> 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 [...].[213]

The illusion that a soul is necessary for a self is as no more warranted than the notion that termite colonies have souls – the seemingly highly organized collaborative work that colony produces is nothing but the product of "a million of semi-independent little agents, each itself an automaton, doing its thing,"[214] as most scientists today agree.

What the self essentially is however is what Dennett calls a "center of narrative gravity for a narrative-spinning human body,"[215] 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't reasonably be an all-or-nothing phenomenon, how absolutism is misguided:

> Since selves and minds and even consciousness itself are biological products [...], 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. [...]
>
> 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.[216]

4.2.2 What a Self is

In the end, Dennett's proposal as to what a self exactly is is not that different from Metzinger's that we'll see in chapter 4.3 – although he explores the concept

[211] See [26].

[212] 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 [2]).

[213] Citation from [30], page 416, emphasis his.

[214] Citation from [30], page 416.

[215] Nomenclature by [30], page 418.

[216] Citation from [30], pages 421f, referring to [34], and [26], pages 101-109.

way less in-depth. Here is what Dennett has to say concerning the part of the hard problem we're 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.[217]

However, that is all he really 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 (and maybe teleological, although he doesn't use that word) point of view. He also explores what the moral implications are if a self really is only a fiction, concerning responsibility and agency. Since this moral tangent is not main subject of this thesis, I would like to leave it aside though.

4.3 Metzinger: Being No One

Metzinger makes a quite convincing argument, one that 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 that we create. However, as I said, he does go further than Dennett. Let's see how.

The self being merely a model does make sense from an evolutionary point of view. As Metzinger points out, such a self model can be a tool, and a weapon[218] – awareness of the continuity and existence of a self as opposed to the world is the very thing that enables self-directed action in a stricter sense, in two ways: It is necessary to have a self-model in order to be able to plan and execute directing action at what constitutes the space the self is perceived as being in, and 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 self.

A self-model also makes various internal parameters (like hunger, thirst, hormone and transmitter levels, intestinal states, and skin temparature) available for cognitive processing and thus reflected action. This reflected action then is not strictly stimulus-related, but probably better denoted as being stimulus-inspired – in that stimuli may end up 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, we decide on one such possible world and execute the actions that we hope will lead to it.

[217] Citation from [30], pages 426f.
[218] See [52], pages 344ff.

4.3.1 The Self is Not an Illusion

I want to get this out of the way first, as it is quite a tempting mistake to make: It is wrong to say that the self is an illusion, because if it would be, there would need to be somebody or something (a self?) having and owning the illusion. This would necessarily make the entire argumentation circular – and that exactly is an argument that has been brought up as something that invalidates the entire self-model theory. Metzinger is aware of this, and he also is aware of the fact that it certainly *seems* to us as if there were a self. In order to explain how that can possibly happen, he expands upon Plato's thought experiment of the cave[219] where we are pictured as being chained to the ground, unable to turn our heads, and merely seeing shadows of the things that actually are, 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 some of Metzinger's postulated entities: The cave is the entire organism, the shadows are low-dimensional phenomenal projections of high-dimensional objects, the fire is neural dynamics, information processing that is constantly perturbed and modulated by sensory and cognitive input, and the cave's wall finally is merely another aspect of the very same neural process.[220] All these things are very similar in that they are entities in both thought experiments, directly mappable to one another.

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

And 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, but experiences itself as being in the center. And shadows, phenomenal representations, as 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,"[221] and what we perceive as the shadow of ourselves is nothing but the shadow of the cave itself.

4.3.2 The Phenomenal Self-Model (PSM)

The PSM, the "phenomenal self-model", is the formal name Metzinger gives this part of his theory, and it is the more fundamental of the two main building blocks of it. The PSM really is the logical conclusion from what we have seen up to this point, and in particular, what follows from his multilevel constraints and the

[219] See [59], and Metzinger's referral in [52], pages 547ff. I will assume that you are familiar with Plato's thought experiment of the cave that he put forth in [59], so I will be very brief on its details here.

[220] See [52], pages 548f.

[221] Citation from [52], page 550.

fact that the system is mistaking the generated transparent self-representation in a naive-realistic self-misunderstanding[222] as a proper self.

This is his main argument for it:

> 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*.[223]

As a matter of fact, it seems to me that there need not be 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 the prereflexive proto self-model, that "effortless way of inner acquaintance,"[224] then there is the PSM, and then there is the PMIR that we'll look at in chapter 5.3. All those three parts together form the perspectivalness constraint, as we saw in chapter 2.3.8.

A PSM is formed naturally as soon as the prereflexive proto self-model is experienced phenomenally and transparently as an entity in the world by a system. All the 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 that is by some necessity (through genetic disposition) added to that prereflexive self-model during an organism's growth and learning phase.

4.3.3 Genesis of a Conscious Phenomenal Self-Model

A phenomenal self-model doesn't necessarily have to form even given many of Metzinger's multilevel constraints. He does specify the probable (and intuitively plausible) stance 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 selfhood. The representational process of mental self-modeling within a coherent world-model and

[222] This is a terminology that Metzinger uses in multiple places – see chapter 4.3.3.
[223] Citation from [52], pages 336f.
[224] Citation from [52], page 158.

> 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?[225]

The argument then is very similar to the one we already saw in chapter 2.3.15, where we ended up with ineffability as the defining characteristic of consciousness. It is exactly the phenomenal transparency (that is indeed constituted by satisfying the transparency, globality and presentationality constraints) of the contents of our mental self-model that is one of the things that are 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.*[226]

Metzinger calls transparency "a special form of inner darkness,"[227] and that darkness is constituted in the fact that all vehicle properties and many content properties (as inappropriate as this distinction between vehicle and content properties may be in this context) just like many causal connections between entities in either categories are hidden from introspective access. A self then is this:

> 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.*[228]

4.3.4 Switching off the Self

A self is hard work for our brains. Thus it makes perfect sense that the self is switched off from time to time, when it isn't needed – leading to times of partial or entire lack of conscious experience as well:

[225] Citation from [52], pages 330f, emphasis his.
[226] Citation from [52], page 331, emphasis his.
[227] For example, in [52], page 331.
[228] Citation from [52], page 332, emphasis his.

> 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 self-model, it simply turns it off. Together with the model the conscious experience of selfhood disappears. Sleep is the little brother of death.[229]

The reason why we haven't noticed this before is because of the transparent nature of our self-model – we cannot see it for what it is, merely an illusion that nobody has. This is the price we have paid for having a self-model in the first place however, for the increased autonomy and thus conscious self-control we have gained through it.

4.3.5 Why "Being No One"?

With the title of his book, Metzinger points to a problem that we encountered in chapter 1.5.2 already. It is not possible to be convinced of his theory of phenomenal self-modeling while remaining a subject. In order to even understand it, 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.[230]

That special stance is one of a thought experiment that tries to resolve the dilemma we are in.

- Either I attempt to understand that there really are no selves – however, I attempt that while I still perceive myself as "self" in that naive-realistic self-misunderstanding, so there is a logical contradiction 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.

[229] Citation from [52], 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]."
[230] Citation from [52], page 628.

Let's try to understand Minsky's theory about the self as it was now explained anyway. If we can't appreciate it or be truly convinced of it, we can still try and describe our own and our fellow agents' self-models as good as possible.

4.4 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.

4.4.1 The Six Levels of Mental Activities

According to Minsky, there are (at least) six levels of mental activities, all of which together then constitute consciousness – and produce the self-models. Those levels are:[231]

1. Instinctive Reactions: Things that we are genetically predisposed to do – and that often don't quite fit into the fast-paced modern world. There are many instinctive things that still make sense today though; regulating body temparature and heart beat rate, or squinting our eyes when we look into a bright light source.

2. Learned Reactions: This involves quite some brain power already, because we have to know which parts of a situation we want to learn something about have to be remembered in what ways, and what the steps that led to eventual success then were. Nevertheless, many simpler animals (like lab rats) do possess the ability to learn new things through positive and negative reinforcement.

3. Deliberative Thinking: At this level, we already have to be able to plan ahead. We have to be able to imagine things, establish counterfactuals, have some way of deciding between alternatives. With these things in place, we can decide between multiple possible ways we want to achieve a goal.

4. Reflective Thinking: This is a continuous monitoring of our own activities, that allows glimpses into both the past and the future within short-term memory. On this level, we are able to think about our own actions, and decide on better ways to handle things for a next time.

[231] See [53], pages 129ff.

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

6. Self-Conscious Reflection: The final level adds moral values and lets us compare who we are (and what we do) with who we want to be (and what who we want to be would do).

Interestingly, this bears many similarities with Freud's distinction between Id, Ego and Superego, with the levels roughly corresponding to these entities.[232]

Minsky's levels of mental activities require differently complex self-models. He doesn't explicitly list those, but I imagine they are roughly as follows:[233]

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

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

3. Deliberative Thinking does require a fairly complex phenomenal self-model that includes knowledge about what the body and mind bound to the self can accomplish, and it also requires the possibility for offline activation of that 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, in that different time streams (the one experienced as phenomenal now, 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, so they have to be part of the opaque component 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 for inspection (that is only introspection for the descriptive self-model) and comparison.

[232] See [53], page 148 for an illustration of the corresponding levels, and see [35] for Freud's introduction of these concepts.

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

Minsky does, by postulating so many levels that may seem to be very similar at first sight, consciously not follow Occam's razor[234] 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 now, without much success in resolving how consciousness works or what constitutes a self. 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.[235]

4.4.2 Multiple Self-Models

Similarly to other realms, like psychology or physics, just one model of a person often does not seem adequate. We certainly have multiple models of other persons in different contexts – their "business self" and their "private self", for example.[236] Minsky now suggests that the same applies to our models of ourselves; there isn't just one self-model, but rather, there is a multitude of those, activated as a selector for different ways to think[237] 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.[238]

[234] See [10].

[235] Citation from [53], page 147.

[236] See [53], 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."

[237] 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 [54].

[238] Citation from [53], pages 304f, emphasis his. This section comes with helpful illustrations that I can only recommend in his book.

Not only are different aspects of a situation bound to different self-models, but future and past selves can be modeled as well, among many others (social, athletic, mathematical, musical, political, loving, sexual, professional[239]). All of these can come either in normative or positive variants. What a switch between multiple such (usually still relatively close to each other) subpersonalities changes is 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.

Usually, such subpersonalities can somewhat easily be described by characteristics or character traits.[240] Minsky shows these different kinds of dispositions that make it even possible to neatly arrange a complex personality into one or a few categories:[241]

- Inborn Characteristics that we were born with (genetic predispositions).
- Learned Characteristics that correspond to the goals and priorities we try to fulfil, and that also influence emotive responses to sensory input.
- Investment Principle – if we learned a way that works, it's usually not efficient to learn another way that would also work, because the way we know works so well and we don't have to go through a learning period again.
- Archetypes and Self-Ideals, essentially the normative self-models we have, that shape positive ones by means of the higher levels of mental activities (the ones 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. It is easier for everybody like that, or to put it in Minsky's words: "It saves a great deal of effort and time to see people or things as stereotypes."[242]

4.4.3 Personal Identity

We certainly do all like the idea of a self, of being an individual. This has evolutionary origins, or at least is perfectly compatible with our current 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 stance of one self to a body. Social relations

[239] See [53], pages 306f.
[240] Minsky makes this example, in [54], 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."
[241] See [53], page 310.
[242] Citation from [53], page 310.

without selves would be awkward, just like it would be way harder to maintain attention and focus if we wouldn't have the illusion of a single, continuous stream of consciousness.[243] But how do we construct such a personal identity, all convenience of having it (and thus teleological explanations) aside?

Minsky completely follows Dennett's argumentation regarding the center of narrative gravity (see chapter 4.2) on this point, keeping in mind that regularly, we can't fully appreciate early childhood memories (see chapter 3.4.1):

> 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."[244]

4.5 Consolidation

Putting aside Descartes' seemingly rather naive view and other dualist tendencies that would involve at least a Cartesian theater (and thus exactly those reminescences of dualism that I followed Dennett in attempting to put away with in chapter 1.2.3), the other three views I presented might differ quite a bit in details, but agree on one central thing: The self is something that emerges from mental processes and states, something that supervenes on brain states and neural correlates.

Let's see how we can consolidate all three into one common view that incorporates all their strengths, hopefully without too many of their weaknesses.

4.5.1 Criticising Dennett

Eventhough Dennett's theory on what a self actually is is fairly thin, it is compatible with Metzinger's, which is way more extensive. He does however make a comparison I can't find myself agreeing with fully:

> 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.[245]

[243] See [53], pages 320f.
[244] Citation from [53], page 309.
[245] Citation from [30], page 430.

To continue to where Dennett's analogy leads us, a program cannot run on just any computer. At least not with extensive further emulatory work: A program that was created for a certain operating system, running on certain hardware, won't be able to run on another basis (unless they are explicitly made to be compatible).

To get back from our reductio ad absurdum of the analogy to consciousness and the fiction of selves: A self relies on a huge lot of representative and presentative processes that supply it with data. Data about both the environment and the inner state that the body the self belongs to currently is situated in. It also is "programmed" in a plethora of parallel processes with corresponding aligning (neural) pathways – pathways that themselves form the processes that data from sensory organs must pass.

Splitting the self from the body is thus not easily possible, as the body at least partially constitutes the self's current contents, and certainly formed the self in the past. A computer is incredibly different from a brain, so in order to emulate a human brain on one of today's computers, there would probably be more emulatory work to be done than actual software transfer.

And 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 fairly static (on a large scale, but changing on a small scale) neural areas,[246] all of which contribute to what ultimately constitutes the self.

The hardware-software distinction doesn't quite work when it comes to human selves, eventhough Dennett is fairly convinced otherwise.[247]

This stays true unless somebody sits down and goes to translate every possible low-level call into one that works on the hardware the program should be transferred to. Of course this is possible, it is today possible to run ancient (by computer standards) Commodore 64 programs on an emulator running on current-day hardware and operating systems, but the more different those hardware basis get, the more emulatory work has to be done. So when Dennett says 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 computer aka hardware, but also an important part of that "software."

4.5.2 Metzinger's Prereflexive Proto Self-Model

There are clues that make Metzinger's prereflexive proto self-model intuitively plausible. For example, phantom limb pains that exist for people who never even had limbs 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.[248] However, the functional components that are able to

[246] See [54] around 1:14:36.

[247] See [30], pages 210ff.

[248] Metzinger has an extensive section concerning phantom limbs in [52], pages 461-488. In particular, he writes on page 478, referring to physicians and neurologists who researched

form such a self-model are something that warrants more research. Because as we have seen, once there is phenomenal experience, and a pre-reflexive proto self-model, the rest that forms our phenomenal self-model (or, with Minsky, self-model*s* – see chapter 4.5.5) comes naturally.

That research would do well to answer these questions here – given Metzinger's teleofunctionalism:

- What kinds of teleological properties are necessary for the formation of a human-like phenomenal self-model, given the other 10 constraints 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, 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 and able to explain many things that remain mysteries in other belief systems. We know for example that there are brain regions that invariably are responsible for a certain body part's sensory inputs. But there are no fleshed-out theories yet that fit these things together. I do believe that this would be a worthy topic for future research.

4.5.3 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. The difference 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 (that then enable a cognitive first-person perspective).

---

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."

He does exemplify this transition by depicting Baker's very interesting analysis of Descartes' "cogito, ergo sum" ("I think, therefore I am"), or rather Baker's more specific reformulation, where I* is the own self perceived as thinker of first-person thoughts:

> I am certain that I* exist.[249]

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 2.3:

- "I*": The content of the transparent self-model, under the principle of autoepistemic closure (as in, it cannot discover its own representational nature by principle).
- "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."[250] In other words, the autoepistemic closure here 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.
- "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.[251]

In other words: The expression "I am certain that I* exist" denotes how exactly it is the case that or mind is fooling itself into believing that a self representing it 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.[252]

[249] Citation from [15], as cited in [52], page 398. Metzinger then goes beyond Baker and adds the distinction between [active phenomenal content] and <linguistic expressions> with corresponding symbols (square and angled brackets, respectively), which further clarifies the ways in which we refer to mental content – but for the purpose of this chapter here, we will skip over that distinction and merely look at phenomenal content.

[250] Citation from [52], page 400, emphasis his.

[251] See chapter 5.3.

[252] Citation from [52], page 401, emphasis his.

4.5.4 Empathy and Intersubjectivity

We now go even further, beyond mere subjectivity, with an outlook at 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, in the end, a phenomenal self-model, systems that 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 different things than modeling the own personhood – and that exactly is, according to Metzinger, what enables what he calls offline simulations of first-person perspectives even for other minds.[253] 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.[254]

This then leads to the possibility of modeling the phenomenal "you," 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 that include goal structures and predicted motor behaviour. It has been shown that some of the very neural structures that an observed human activates for, for example, moving a finger, are also being activated in observers' brains. 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.[255]

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 can realize that other persons are 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 and mental states

[253] See [52], pages 182ff.
[254] Citation from [52], page 365, emphasis his.
[255] Citation from [52], pages 367f, emphasis his.

that we would have if we were them) can we take those into account and act according to them.[256]

Notably, there are parallels to Dennett's intentional stance that we will briefly look at in chapter 5.2.1.

4.5.5 Multiple Phenomenal Self-Models

I do agree with most parts of Dennett's view. I agree with almost all parts of Metzinger's view, and I even also agree with almost all parts of Minsky's view. Considering 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. Allow me to attempt a crude version of just that.

Since Dennett's ideas on subjectivity are not as fleshed-out as these of Metzinger, and can mostly be expressed by a subset of Metzinger's ideas, I think we can for now drop Dennett. This leaves Metzinger's generation of phenomenal self-models, but 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 experts[257] 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 phenomenal self-model can be experienced.

When we are modeling normative selves, this happens in an opaque manner, so those are available for introspection – but the fact that our positive phenomenal self-model changes from one moment to the next is not necessarily available to us, since it is transparent (and some parts of it are ineffable). We can, as with other things, introspectively distance ourselves from our deeds and look at what we are doing; an example of that is when we take a step back and realize we are just acting so rudely because we are jealous when we really shouldn't be – but by default, we don't do that, and often don't realize how we're really not the same today as we were yesterday.

The common saying "I'm not quite myself today" obtains an entirely different and way deeper meaning with this theoretical background.

Those self-models can, to some extent, be stored transtemporally. This is always true for a description of an opaquely experienced phenomenal self-model. But we have to keep in mind that there is a reservation: For some of those stored self-models, it is no longer possible to fully appreciate them (in the sense that

[256] Do note that the PMS 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. We will look at Metzinger's way of forming intentional relations in chapter 5.3, introducing the PMIR. Metzinger says in [52], 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."

[257] As I mentioned before, unfortunately we do not have the space to look at Minsky's critics and experts – 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.

I constrained it in chapter 3.4.1) – 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 Intentionality

It is important to notice that "intentionality" is one of those (with Minsky) suitcase-like words[258] that can have many different meanings. What I intend to look at here is the peculiar way in which our mind can focus on a concept, or a group of concepts perceived as one entity, and seemingly forms a direct connection between us and this target entity – at least, that is the way we often phenomenally experience it.[259]

It is, as a probably naive but intuitively graspable model, the arrow of attention, pointing out of our heads at an object in the world (or at an imagined object inside our heads), while carrying a meaning under a certain aspect.[260]

5.1 Brentano: Immanent Objectivity

Without naming it "intentionality" as more recent philosophers do, Brentano already introduces and researches the concept of something that is intended with every mental phenomenon:

> 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 hpenomenon 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.[261]

The idea is very intriguing. There always is some kind of object of any mental phenomenon indeed, something that the mental phenomenon is about. As we

[258] See chapter 1.5.1.

[259] I do believe that it is a slight bit of a philosopher's delusion that everybody thinks about it with that intentional arrow pointing out of our heads – when speaking with 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.

[260] As Searle put it in [64], "All intentionality is aspectual."

[261] Citation from [17], page 88. 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."

saw in chapter 1.5.1, there are suitcase words that an entire class of entities can be described with; The fact that we can linguistically refer to different kinds of intentionality always as something of the same kind doesn't necessarily make that so from an ontological point of view. But 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.[262]

Brentano, in his work, already classifies mental phenomena into three kinds: presentations, judgements, and those "of love and hate" – by which he means all kinds of emotional phenomena. He is well aware that there can be other classifications, too, he explicitly mentions those of Aristotle (thought and appetite) and the apparently prevalent classification of his time (presentation, feeling and will).[263]

5.2 Dennett: Only Derived Intentionality

Dennett doesn't believe in intrinsic or original intentionality. That's the short version – the longer version starts with a thought experiment. Allow me to leave out many important passages in the following quote, to reduce it to its essence:

> 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 [...].[264]

Dennett goes even further: Not only does he doubt that intentionality can be intrinsically about something in particular, he does not believe that there is anything beyond derived intentionality (which has been called "observer-relative ascriptions of intentionality" by Searle in his Chinese Room discussion[265]) at all. That is what his argument for the intentional stance is all about.

[262] See [20], which is very interesting, goes beyond Brentano and also looks at Husserl and Ingarden.
[263] See [17], pages 194ff.
[264] Citation from [29], pages 294f. There, he also lists plenty of philosophers who either agree that there is intrinsic intentionality, or agree that there isn't.
[265] See [63].

5.2.1 The Intentional Stance

It is striking how often us humans ascribe intentional acts and beliefs to animals, or even machines who obviously don't have any feelings or even states that could properly classify as "mental" (or rather, phenomenal) states. We also ascribe intentional acts and beliefs to each other, with apparently more justification and less metaphorical.

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 though? Dennett makes out three principles:[266]

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

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.

So, essentially, evolutionary fitness and rationality are sole criterions for reasonable ascription of intentionality, and ascription of intentionality is all there is to true intentionality (or, in other words, there is no intrinsic intentionality). I will criticize this view in chapter 5.4.1, for now let's explore what this means: It means 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"*[267] and its biological needs), it possesses intentionality.

Strictly speaking, this definition also means that only biological systems (no artificial or postbiotic systems whatsoever) 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 substitute a similar concept for those "biological needs", 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 (biological systems simply die), this also includes reproduction.

5.2.2 Further Notions of Intentionality

It is important to note that Dennett talks about more than my narrowed-down definition of intentionality – he often talks about consciousness and mental states in general, intentionality as "possession of mental states." In particular, he is concerned with disproving that there is a meaning intrinsically attached to

[266] See [29], 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.""

[267] For example in [29], page 298.

individual such arrows of attention we direct at the world,[268] looks at higher-order mental states,[269] and even looks at the language of thought in the bigger context of aforementioned mental states.[270]

In no way do I want to disqualify or doubt the importance of those examinations by not looking at them here, but as I said in the introduction of chapter 5, I merely want to look at intentionality in a rather narrow sense. I do believe that if we want to understand consciousness, it makes sense to look at 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 6.

5.2.3 No Intrinsic Intentionality

As we've seen, Dennett does not believe that there is such a thing as original or intrinsic intentionality. Akin to Searle not finding anything that could give his Chinese Room's homunculus any intrinsic intentionality that he could pass on to the system he is in, Dennett does not find any part of the human mind that could give it intrinsic intentionality. He does refer to Dawkins[271] when he describes what us humans really are, from an evolutionary point of view.

He suggests a thought experiment where we would want to experience the year 2401, and then build a robot that could contain and protect our hibernating body for the time until then. Since the world is an inherently dynamic place and merely sitting always in the same spot could probably not ensure our survival, that robot would have to be able to distinguish things that are good or bad for its continued functionality, would require some forms of autonomous self-control with goals and subgoals, would have need for quick-and-dirty approximations in its calculative appraoches to problem solving. It would show equivalents of our mental states of wondering, seeing, deciding, and essentially exhibit a sophisticated form of derived intentionality, similar to what a machine passing the Turing Test would need to show. Now, what Dennett points out is this:

> 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,

[268] See chapter 5.2.3.

[269] See [29], pages 242ff.

[270] In particular, he points out (in [29], 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."

[271] See [25].

> 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![272]

This indeed does raise the question: Assuming we do not believe in magic or dualism, how does the derived intentionality (that we do 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.

5.3 Metzinger: Just a Phenomenal Model (PMIR)

PMIR stands for "phenomenal model of the intentionality relation", and it is the higher-level of the two main building blocks of Metzinger's theory of consciousness. It is also part of the perspectivalness constraint, just like the prereflexive proto-self and the PSM are. Metzinger sees that perspectivalness constraint as one of the three main building blocks 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.

Similarly to the PSM (see chapter 4.3), 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. Its content 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.

5.3.1 Kinds of Perceived Intentional Relations

The trivial description applies to a wide variety of cases, and a plethora of entities can be represented as object components. They usually fall into one of the following categories:[273]

- A perceptually deliberately attended object, given through sensory input.
- A consciously deliberately attended opaque kind of conscious content, like a cognitive self-model.

[272] Citation from [29], page 298, emphasis his.
[273] See [52], pages 411f.

- An object forcibly perceptually attended, given through sensory input.[274]
- An complex mental motor self-simulation, usually prior to embodying it by executing those motor behaviours.

Of course, all those categories can be represented with different kinds of relations, relations that would be called "attitude specificator" in propositional attitude psychology. Those include attending, thinking, willing, fearing, and many others. They also have a direction – the "arrow of attention" points outward (in regular perception) or inward (in introspection), from the first-person perspective, but always originates from the subject component – putting subject and object into the aforementioned asymmetrical relationship.

5.3.2 Phenomenalizing Intentionality

If we want to naturalize intentionality (as Metzinger wants to), phenomenalizing it is the necessary first step. Here's how he explains this:

> 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.[275]

Part of this is obvious: Meaning and the conscious experience of meaningfulness (which corresponds to intentionality) are not the same things, although they are experienced (thus the phenomenality) as being intuitively and unquestionably the same. Metzinger now makes the argument that while we do know that phenomenal content supervenes locally, 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 can thus potentially be naturalized.

Showing that intentional content is a special form of phenomenal content is of course what the theory about the PMIR is all about. Since all there is to intentionality is the phenomenal experience of an intentionality relation, it could exist completely without ontological counterparts to the entities represented by said intentionality relation's phenomenal contents. Intentionality could be entirely virtual, and we couldn't possibly notice. In fact, it is even probable

[274] Forcibly attended as in "finding yourself forced to automatically attend," as [52] specifies on page 412.
[275] Citation from [52], page 414, emphasis his.

that we misrepresent many things out there in the world, because that misrepresentation has proved to be more beneficial for our continued survival as a 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.[276]

5.3.3 PMS without PMIR?

There are pathological cases where humans have damaged brains that are still able to construct a PSM, but no PMIR anymore.[277] 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 that they could relate to, things they could will, 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.[278]

This is mainly because phenomenal volition is a form of phenomenal intentionality, and thus not possible without a PMIR. And without volition, even goal-driven movement cannot be experienced as agency, 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*.[279]

So while it is possible to have a PMS without a PMIR, a full-blown phenomenal first-person perspective requires both.[280]

[276] Citation from [52], page 416.
[277] See [52], pages 416f, where Metzinger also refers to case studies in [22].
[278] Citation from [52], page 419.
[279] Citation from [52], page 425, emphasis his.
[280] This will become clearer once we look at the bigger picture in chapter 6.

5.4 Consolidation

It is rare that a philosopher is talking about intentionality without also talking about consciousness or subjective experience in general. The two concepts are very close to each other – 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 this as follows:

> 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.[281]

What is interesting is how this functional coupling of the intentionality relation with consciousness lures many philosophers (even Dennett, as we saw in chapter 5.2.2) into looking at intentionality mainly in the bigger picture as well. Metzinger's analysis of the intentionality relation is among the most thorough that 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.[282]

With intentionality sitting at the core of consciousness, so much that it is often used as the defining characteristic of consciousness or more, it is very important that we have a clear picture of what it actually is. Metzinger, and to a lesser extent Dennett, has shown us a way to naturalize intentionality without losing the information 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. I do believe that this is a worthy approach to take.[283]

[281] Citation from [52], page 427, emphasis his.

[282] For example, when Dennett goes back to Brentano when undertaking a similar project to the one I had in [37], showing how the brain is merely a special kind of Turing Machine. In [29], 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 pretty much for "consciousness" here. Admittedly, I am guilty as well, I did mix up the two for the sake of better short-term clarity in chapter 1.2.

[283] See chapter 1.5.3.

5.4.1 Dennett's Intentionality Should Be Phenomenal

I am talking about Dennett's argumentation that intentionality is nothing but an ascription that happens through a stance that we take on when looking at entities that might exhibit inentionality. We remember, he says that it all comes down to whatever intentional states we ascribe to a target system has to be what its intentionality truly is.

This argumentation seems to be counter-intuitive to me (among many others), and the reason why that is appears to be two-fold:

- Less importantly, it 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. I guess this could be explained away by merely a different choice of words of course, nevertheless, I think it is important.[284]

- More importantly, his argumentation is strictly counter-intuitive because intentional acts for us humans have what Nagel might call a what-is-it-likeness to them. They are phenomenal in nature. This is implicitly included in the mere notion of intentionality – intentional acts ring a bell, because we all know what it is like to conduct an intentional act, and we also know that for a computer merely following a script, following that script can't possibly be anything like us conducting such an intentional act. In fact, arguably, for today's computers it's not anything like anything at all.

Now, Dennett does not ascribe any special meaning to phenomenality, as we know from chapter 2.2. I would argue however that for an observer-relative ascription of intentionality to be intuitively intentional, it has to be conducted by a system that has phenomenal states. It has to be a specimen of phenomenal intentionality, and it has to satisfy at least Metzinger's perspectivalness and transparency constraints.

The same applies for Dennett's heterophenomenology. Dennett defended heterophenomenology and with it, his form of intentionality, against the attacks from Nagel in [56]:

> 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

[284] So does Nagel. See [56].

> 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.[285]

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. But that is beside the point. And of course, even if they are similar, they won't be exactly the same for certain, so I do agree that they will not have the *exact* same dispositions as us humans do.

But, remember what I said about the difference between appreciation and description (see chapter 3.4.1). I believe this difference is both important and apparently ignored by Dennett.

Even assuming that description is all we can aim for (with which I disagree): The whole of Dennett's argumentation does build on on the assumption that there are no strictly private such experienced patterns. And Metzinger clearly does show that not only there are such private patterns,[286] but also why they are private: Because they're causally active (and thus can't just be ignored[287]), but not introspectively accessible – Metzinger's very definition of being transparent.

[285] Citation from [31].

[286] See chapter 2.3.15.

[287] Of course, there could be an even more fine-grained descriptory level that would also capture those, but this is where appreciation stops working – as soon as we merely have a description, we can't relive the experience, and thus it can't be fully appreciated anymore. 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.

Part III
Conclusions

We started out with some definitions and clarifications, putting epistemology, dualism, supervenience and evolution into a naturalist context, and having a look at some other conceptual subjects that frame this thesis' subjects. We then looked at all the individual fundamental parts that make phenomenal subjective consciousness what it is – phenomenology, qualia, subjectivity and intentionality – and have looked at various ways in which philosophers to this day have explained them.

What is still missing is that we put all these things together, into a bigger picture, and that we look out into what new achievements this can lead us to in the nearer future. This is what this part of my thesis is about.

6 Summary and the Bigger Picture

Let us take a step back from the four partial problems that I identified and analyzed in part II. With all the bits and pieces in place, we now know that full phenomenality partially relies on subjectivity, while subjectivity itself relies on some of the multilevel constraints that make up the rest of phenomenality as well. We have realized that intentionality relies on some basic form of self-model, while at the same time it is also what enables the higher levels of the self-model.[288] We also know that the most basic form of mental content are not qualia, but phenomenal representational content – and that qualia do have an ontological justification as well after all, because they group regions of presentational quality space for the symbolic *re*presentation that intentionality enables. With these additional bits of knowledge in place, we will need to structure our summary with some differences (with slightly more complexity) from the analysis itself.

I will now present again all the entities that I think are important for making up phenomenal subjective intentional experience. I will again follow the natural order of apparent increasing complexity, but with more confidence than I did so in part II. After all, we did have a thorough look at the pieces of the puzzle by now. This does mean however that I now have a different idea as to what is more complex than what else, and it also means that since so many things are interwoven, untangling them means that we require quite some intermediate steps between those four (too) easy levels.

6.1 Proto Self-Model

There is a requirement for the PSM and the PMIR that is actually part of Metzinger's perspectivalness constraint[289] as well. I am of course speaking of

[288] Don't worry if this isn't familiar to you, I didn not make this explicit so far.
[289] See chapter 2.3.8.

the still somewhat miraculous proto self-model that we saw in chapter 4.5.2. While I do believe that further research into what exactly constitutes this proto self-model is warranted, there must be at least the following two factors that importantly help in building it:

- Functional presentation centers in the brain that are responsible for receiving nerve cell input from various body parts (including inner organs). These are obviously genetically hardcoded, as the phantom limb experiments we spoke of in chapter 4.5.2 show.
- A single localizable body as region of maximal stability and invariance (that we already encountered in chapter 2.3.8) that allows the system to get accustomed to taking a first-person point of view stance and also assigning entities and properties to the first person this stance originates from.

Arguably, this prereflexive proto-self is not even 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 there is no way any higher concepts that refer to locations in any way can possibly form. 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 around us allows us to form a world model at all.

6.2 Basic Phenomenality

The level of consciousness out of those we saw in chapter 2.3.14 that is required for a very basic form of phenomenality[290] is probably the one Metzinger calls "Differentiated Consciousness."

This includes the presentationality constraint (2, the phenomenal Now), the globality constraint (3, integration into a coherent global state), their expansions in the convolved holism constraint (4, subdivisions within subdivisions that split this global state apart) and the dynamicity constraint (5, that allows a system to experience past and future, and the Now embedded in them). Also, we definitely need the transparency constraint (7, 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.

Since we realized in chapter 2.3.14 that the ultrasmoothness constraint (10, homogeneity of simple content and the non-existence of a grain problem) really is a prerequisite for convolved holism (as we get into an infinite regress otherwise when attempting to subdivide former wholes infinitely – which is exactly what the grain problem is about), it will have to be included in this step as well,

[290] At least if we eventually want to reach consciousness – of course, we could insert an intermediate step with Metzinger's Minimal Consciousness here as well.

although it will only be required for its true strength, concept formation, when we introduce qualia.

As we found in chapter 2.3.14 as well, we probably need the global availability constraint (1, mental contents are available for deliberately guided attention, cognitive reference and control of action) as well, with Minsky's reservations that we saw in chapter 2.4.2.

While the adaptivity constraint (11, everything has to make sense from an evolutionary perspective), at least in its weakened form from chapter 2.4.5, is certainly necessary for a conscious system that is also perceivable as such by us, it is not necessary for basic phenomenality. However, since my goal is to have a full analysis of a conscious system in this summary, I would like to include it in this part of the description as well.

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.

6.3 Phenomenal Presentational Content

As we saw in chapter 3.3.5, qualia are not the atomic entities that 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 about the most computationally complicated, and the one most of today's attempts at producing weak artificial intelligence are struggling with. It includes mostly pattern matching tasks: Speech recognition, facial recognition, three-dimensional object perception, orientation, memory formation, and other things that I will talk about in chapter A.2.

The framework and procedural possibility for these things has to be, in the case of biological systems, genetically hardcoded and built into the proto self-model, as it is not possible to build these from scratch, not even for such an efficiently self-learning system as our body and brain. But in order to be able to perceive the data from these perceptive and categorizing processes as presentational content, we also require basic phenomenality.

Arguably, the representation of intensities contraint (9, 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.

6.4 Basic Subjectivity

In the spirit of merging the concepts of Minsky and Metzinger (as I attempted to do in chapter 4.5.5), subjectivity here really is all 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.

Those self-models consist of the six levels of Minsky, as seen in chapter 4.4 (that somewhat correspond to Freud's distinction between Id, Ego and Super-Ego). However, for the later of them we require an intentionality relation, so only two such levels (the ones correllating to the Freudian Id) are part of the basic subjectivity we are about to describe.

1. Instinctive Reactions - requiring no self-model yet.

2. Learned Reactions - forming a simple self-model that has 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. Basic subjectivity however is more than the proto self-model in that it incorporates sensory input and allows processing it in various ways.

6.5 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 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.

Going with Dennett in dismissing intrinsic intentionality as per the argumentation we looked at in chapter 5.2.3, we have but one choice: There is no such intrinsic intentionality, so we need a replacement. And, handily, we have one in Metzinger's phenomenal model of the intentionality relation that we encountered in chapter 5.3.

That model is an entity that is continuously rebuilt in a dynamical process, hardwired in the brain, that allows us to *re*present entities that formerly were only 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 that is part of a huge selection of entities that would qualify for filling the "object component" slot in a system's intentional modeling framework.

That relation 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 such 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."

6.6 Qualia

While qualia are the true test for ultrasmoothness because it is now required for us forming proper categories of content, we did introduce the ultrasmoothness constraint (10) for basic phenomenality already and can thus merely apply it to phenomenal presentational content now. As we saw in chapter 3.4.3, eventhough traditional qualia as truly atomic 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 do have legitimation as the concept of volumes of qualitative space, borders of concept formative capabilities of a system.

As a matter of fact, for qualia we do need to be able to form concepts, and for that we require ways to relate to phenomenal presentational content. So we need both basic subjectivity (or we don't know what it is that relates to the presentational content) and intentionality (or we can't relate to it at all).

6.7 Full Subjectivity

There is more to full subjectivity than to basic subjectivity, because as soon as we also have an intentionality relation, we can potentially relate to our basic self-model (that in turn presents the proto self-model). I will list the first two levels again here as well, although only levels 3-6 are really what this chapter is about:

1. Instinctive Reactions - requiring no self-model yet.
2. Learned Reactions - forming a simple self-model that has the bipartition between self and world.
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[291] to be able to be activated independently by different streams of thought.[292]

[291] Or rather, the processes currently executing those process descriptions that enable the system to satisfy the dynamicity constraint.

[292] A stream of thought really is just such a process in the brain, initiated originally by some form of sensory input (that is experienced as presentational content and conceptualized as qualia). This sensory input can well be of system-internal origin – after all, our bodies never are truly immobile, at least the heart is beating all the time. And self-initiated motor action can well lead to further sensory inputs.

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, or there is no experiencing self – remember chapter 4.3.4 – and only one of them at a time is fully transparent in standard cases[293]), but as we saw in chapter 4.4.2 there is no complete all-encompassing entire-self-model because that would lead to too much computational complexity for our brain. Rather, only select memories and behaviours are available to each individual self, both for practical (simplification through information hiding) and continuity reasons; it makes sense to be able to rely on yourself.

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[294] that the agent itself can form – they can be entirely altruistic or entirely egoistic in nature. Human agents are a great example for this because they are really creative[295] when it comes to what goals their normative self-models serve.

These self-modeling routines are then also borrowed for intersubjectivity and empathy, the subject of chapter 4.5.4 – 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[296] discussions. As such, such a self-model is also the requirement for meaningful (in the same experiential sense) language generation, because only if you can model somebody as a person is it interesting to truly tell them something.

Looking at parallels to Metzinger's levels of consciousness in chapter 2.3.14: By this point we have added all of his constraints including the perspectivalness constraint (6, including the proto self-model, the PMS and the PMIR), the full transparency constraint (7, epistemic closure concerning inner workings of mental processes with the possibility of making some of these processes opaque) and the offline activation constraint (8, allowing representational content that is not directly driven by presentational content).

So together with the constraints we already added for basic phenomenality, we now have a fully conscious system that has, to say it with Searle, "causal powers equivalent to those of the brain."[297] Even better, we even have some slight idea what those causal powers could be!

[293] There certainly are several such self-models at once transparently active in the brain of a patient with a multiple personality disorder.
[294] See chapter A.2.4.
[295] See chapter A.2.5.
[296] 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 5.3.2.
[297] Citation from [63].

Part IV
Appendix

A Outlook

What do the conclusions from this thesis mean in the bigger context of philosophy of mind? What are paths we could now pursue to support or disprove my conclusions? How could these conclusions be used in other scientific fields?

A.1 Artificial Intelligence

When I set out to write this diploma thesis, I planned sections about artificial intelligence – where we stand, what positions and approaches there are, and how we can attempt to incorporate the findings from philosophy of mind into research from computer science. Unfortunately, I had to cut these parts eventually – there just wasn't enough place to talk about all of that as well.

Nevertheless, I do believe that artificial intelligence is one of the scientific fields of research that can help philosophy of mind, and vice versa. Computer science is where a philosopher of mind can experiment and put his theories to the test. And conversely, philosophy of mind is where a computer scientist can draw inspiration for experiments in artificial intelligence from.

A.1.1 Weak and Strong Artificial Intelligence

I may not agree with the point about strong artificial intelligence being impossible that John Searle makes in his Chinese Room thought experiment.[298] But in the discussion following the article, he does put forth a very important and necessary distinction: The one between weak and strong artificial intelligence.

Weak artificial intelligence is a mere simulation of intelligent behaviour. It might fool us, it might even pass the Turing Test,[299] or it might be different altogether from the way us humans act and think and reliably solve problems us humans can't solve. Weak artificial intelligence basically is any kind of systematic behaviour that seems as if it was intelligent to us.

Interestingly, all (or at least, most) of today's efforts that go towards producing artificial intelligence are purely behaviourist. We try to replicate human-like behaviour, and think that the machines that we produce will then automatically become intelligent and (more importantly) conscious. I think that this thesis shows up at least some hints as to why this approach will not necessarily lead to truly conscious systems.

Strong artificial intelligence on the other hand is true intelligence that includes subjectivity, phenomenality and intentionality, one that can feel qualia. Essentially, it is artificial intelligence after we've solved the hard problem of

[298] See [63] for his argument, and [37] for my discussion of it.
[299] See [12].

consciousness, and successfully applied the solution to an artificial system. As Metzinger rightly points out[300], that doesn't necessarily have to be a purely technological system – it can well be a postbiotic one, with artificially grown neurons, or maybe with quantum computers, or a technology we haven't discovered yet.

A.1.2 The Need for a New Approach

Minsky is a proponent of an older tradition of artificial intelligence, a more formal one that puts less focus on self-learning systems. Let's hear what he has to say about genetic algorithms and neural networks:

> I'd like to argue that if you're interested in artificial intelligence these days you're exposed to a lot of arguments that ... well, in the early days, people tried to do things with symbolic AI, sometimes contemptuously called old-fashioned AI, and that was too rigid and rule-based and mechanical, and it had to be programmed. What happened starting around 1980 was that most AI researchers tried to move in the direction of making machines smart without programming them, by using neural networks or genetic algorithms or statistical models. The value of that is that you're hoping that your machine can learn to be smart without your knowing how it does it.[301]

While neither him nor me want to diminish the value of the many achievements of artificial intelligence research in recent years, he does have a point: We cannot expect from our machines to learn to be smart without us knowing how not only it'll do that, but also how it could even learn that.[302] We have to set some guidelines, and for that we have to find out what those guidelines will have to be first.

A.1.3 How This Thesis Fits In

I do believe that this thesis here does help in going towards a new approach. Arguably, we need a framework, a base that allows all the other important parts of current research in artificial intelligence to work together, to form a bigger whole. They are the Ways to Think of a porential future postbiotic or artificial consciousness. They are what enables it. But in addition to these Ways to Think, we need something they are embedded in. A thing producing phenomenality, a transparent representation system.

To say it with Minsky, when he talks about how all of the currently employed strategies and methods of producing (weak) artificial intelligence have this weakness:

[300] See [52], pages 206f.
[301] Citation from [54], 8:40-9:45.
[302] Unless we happen to have a few million years at our disposal – see chapter 1.4.2.

> In fact, all of these methods are very useful for certain problems. What we don't know in general is: For what kind of problem is it good to use a statistical inference system? What kinds of problems are ones that can be learned and handled by neural networks or by genetic algorithms?
>
> What I'm proposing along with Gerry Sussman and Hal Abelson is to develop a new kind of AI system that has places in which we can insert all of the useful results that tens of thousands of AI researchers have made.[303]

What conclusions we reached from the theories of Dennett, Metzinger and Minsky could well be part of that foundation. If not, at least I hope they are part of our way there, even if I am completely mistaken with my conclusions. Because as Dennett says:

> Making mistakes is the key to making progress. [...] Instead of turning away in denial when you make a mistake, you should become a connoisseur of your own mistakes, turning them over in your mind as if they were works of art, which in a way they are. You should seek out opportunities to make grand mistakes, just so you can then recover from them.[304]

A.2 The Less Hard, But Still Hard Problems

While I attempted to look at what it is that enables phenomenality and subjectivity in conscious systems, there are plenty of other things that are required for a truly conscious system like us humans are. Most of these subjects are already topic of extensive research. Many of them are in the prime focus of today's attempts to produce artificial intelligence, while others are subject of psychological, neurophysiological, or even didactical fields of science.

I will only briefly touch all these subjects, while giving some pointers as to what research I found that helps enlighten our understanding of them, but keep in mind that there is much more that could be said about any of them. I would even go so far as to say that we do not fully understand any of them – we're not able to, exactly because they all are so interwoven with the hard problem of consciousness itself.

A.2.1 Presentational Content Generation

This is, as a matter of fact, one of the prime topics of today's research into (deliberately weak) artificial intelligence. Pattern recognition tasks are necessary for conceptualization of sensory input (in the case of artificial intelligence, often through cameras, scanners or in the form of text or other prepared data) – and exactly these pattern recognition tasks are what the modeled (trivial versions of) neurons in neural networks are particularly good at.

[303] Citation from [54], 30:08-30:55.
[304] Citation from [19].

Nevertheless, today's artificial algorithms are nowhere near as good as the ones evolution gave us humans. As it turns out, and as was not expected at first, it really is the case that those things that appear to be easiest to us humans seem to be really hard for a computer, while obviously complex things like playing chess are comparably trivial.[305]

These pattern matching tasks are so close to what we will need for forming presentational content for one reason: Our brain only works with concepts, not with sensory data itself. And in order to be able to work with these concepts, they have to be formed – they have to be grouped and categorized. This grouping and categorization *is* a kind of pattern matching, and it is exactly what optical character recognition (OCR) or similar tasks where pattern matching is employed are about.

A.2.2 Integration of Memory

The kind of memory that a mind like ours requires is pretty complex. It consists of way more than just the raw, abstract and atomic pieces of data we currently use for storing data in computers.[306] The resourcefulness capabilities of a human brain are vast and diverse – we do not directly access our knowledge, but often reason by analogies, and certainly associate by such analogies. Minsky calls the structures we use for this kind of reasoning "panalogies,"[307] for "parallel analogies."

- Analogies because every entity in our memory, every bit of knowledge is linked to multiple other bits – to similar objects, to properties, to a time and place if it is an episodic memory, to solutions if the entity is a known problem, and generally to other related entities and concepts.
- Parallel because every bit of knowledge is embedded in different realms (physical, dominion, mental, and more) with such analogies.

This kind of knowledge representation also requires quite a broad basis of "common sense" knowledge,[308] as for being able to form analogies like that we need something the new piece of knowledge could be analog to.

Furthermore, there are different kinds of memory – episodic and semantic memory[309] to name but one distinction, long-term and short-term to name another dimension, there is "knowing that" versus "knowing how."[310]

[305] At least until we attempt to model the half-intuitive way us humans play chess – a traditional computer program does it through raw computational power and predefined libraries.

[306] That isn't to say that it would be impossible in principle to store such semantic data on current-day computer hard drives, but there would need to be an abstraction layer between this semantic and a trivial raw representation of said data.

[307] See [53], pages 168f and 260ff.

[308] A very nice overview over current efforts to construct such a basis can be found at [1].

[309] As Metzinger points out in [52], page 415 (referring to [21]), "episodic memory, of course, is a process of *re*constructing what was here termed a PMIR, because one necessary constituent of memory retrieval is not smiply the simulation of a past event, but an association of this simulation with a *self*-representation."

[310] See [67].

A.2.3 Learning

The process of learning itself, the way memories are gained and obtained, is everything but trivial.[311] Learning essentially simplifies decision-making by storing context-relative data (patterns, knowledge, motor behaviour) in ways that are then retrievable later in the relevant (or relevantly similar) contexts.

There are quite some steps in memory formation – the process flow consists of at least encoding, storage, consolidation, recall and oblivion. "Learning" is a suitcase word[312] that encompasses all of the processes up to recall – encoding and storage are obvious, but consolidation is definitely an important part of it as well; it is embedding atomic facts in the panalogies that make them available for context-relative retrieval later on.[313]

The benefits of learning are many.[314] They all mean but one thing:[315] We need less time to reorient (in the sense of having to form new counterfactual self-models, often with motor simulations, that we then need to compare with the goal state) and thus can spend more time on executing behaviour that is efficient in regards to reaching a goal state. This can go so far that we are able to execute that behaviour while being consciously engaged with an entirely different subject. The behaviour itself is also more efficient, because learning allows us to anticipate common problems and avoid errors we made in earlier attempts of executing the action we improved through the learning process.

This makes it clear why teaching an action is often difficult for people who learned to do that action really well: They no longer need to think about the (motor or lower-level mental) behaviours they are performing, they internalized them to a large extent and thus are no longer necessarily able to make these deliberations and motor behaviours explicit.

A.2.4 Goal Representations

One kind of panalogy, and one that fundamentally influences at least our normative self-models, is what we call "goal." But, what are goals, with what are they linked, and how do we go about pursuing them?

Obviously, goals are in a hierarchical structure – leaving intuition aside, this directly follows from the convolved holism constraint (4). The nature of goals then is more of a subject for psychological and didactical research than for

[311] See for example [39], pages 738ff.

[312] See chapter 1.5.1.

[313] This means that I do not believe that there is such a thing as properly non-associative learning. Semantic memory may not be stored associated to the context it was learned in, but it certainly is associated with the contexts it is potentially useful in. Otherwise, I do not see how it could possibly be retreived later on. As such, there is probably a gradual and not a binary difference between episodic and semantic memory – a memory is episodic to the extent that it is associated (semantically, or at least experienced as such) with the circumstances it was obtained in.

[314] See [39], pages 742ff.

[315] My apologies for being horribly brief here and not giving Hacker's theory the space it deserves.

philosophy or neurophysiology. Arguably, goals are not much more than bits of memory (grouped in panalogies) that have normative force.

The bits of memory that can serve as goals are, so it seems, always (partial) self-models or (partial) world models – a lot of money, a happy state of mind, love, peace on earth with fluffy bunnies everywhere. While executing actions, we can then as per the offline activation constraint (8) access those goal models, run mental simulations and compare the results of the actions in world zero to the goal states. Minsky describes this as a constent process where we describe the current state, compare it with a description of the desired state, select a difference through built-in difference detectors[316] in the brain and then go and change the situation to reduce that difference.[317]

A.2.5 Creativity

Creativity is a subject that often comes up when the hard problem is being discussed, and indeed it is an important part of intelligence. As a matter of fact, Minsky goes even further and claims that resourcefulness (which he describes as having many ways to think available and thus being able to approach problems from many different angles – which is arguably the same thing as creativity) is one of the three defining characteristics of intelligence, together with persistence and aim.[318] He points out that a resourceful mind really has the following benefits:

- We can see things from many points of view.
- We have ways to rapidly switch among these.
- We have developed special ways to learn very quickly.
- We learn efficient ways to retrieve relevant knowledge.
- We keep extending the range of our Ways to Think.
- We have many different ways to represent things.
- We develop good ways to organize these representations.[319]

Since all these things rely on representations of memories and goals, on how many ways to think were obtained how and what goals and subgoals are active at the moment, this leads to a very unpredictable behaviour, particularly from a third-person point of view. And unpredictability is exactly the thing that makes creativity special, and that makes it intuitively impossible for machines to have it.

[316] See [53], pages 195ff.
[317] See [53], pages 267ff – and note how this is closely related to our learning process, that Minsky mentions there as well.
[318] See [53], pages 188f.
[319] List cited from [53], pages 256f.

Arguably however, given sufficiently chaotic inputs (and maybe some deeper forms of randomness than our current approaches handle, since quantum effects will most probably influence what's happening in the brain), it is not impossible for another resourceful system to exhibit creative behaviour.

A.3 The Hard Problem Demystified?

By no means do I mean to claim that we now have all the answers, or even that we have any more than first pointers as to what exactly it is that makes subjective consciousness possible and phenomenally interesting. But those first pointers we certainly do have.

From a naturalist or rational-causalist background, the traditional theories about the mind are not really satisfactory. It always did amaze me how dualist tendencies, 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, that other, artificial or postbiotic systems should be excluded from the club of the entities that can potentially be intelligent. There just is no rational reason for that, except of course the all too human fear that we could suddenly be superfluous.[320]

But if we don't try and see if we can find out if these other systems can be intelligent, and then informedly talk about what exactly the premises are that make it impossible for them to be intelligent, we can't possibly know, and we should admit that. In all honesty, Searle's "special causal powers" do not fit that bill very well, particularly not if they're defined as "the things that make it impossible for these other systems to be intelligent".

A.3.1 The Future of Philosophy of Mind

On the way towards such a better understanding of the hard problem of the philosophy of mind, one that allows us to informedly talk about the things we do not know about and narrow down what it is that we don't understand, I think that Dennett, Metzinger and Minsky deliver important signposts. By putting their theories together in one thesis like I did here, we did encounter some incompatibilities that had to be clarified, but we also found many areas where the fusion of their theories proved to be fruitful and able to explain more than their individual parts.

[320] I referred to [43] before, but will do so here again. To quote Bill Joy: "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."

I do believe that in particular the summary from chapter 6 is a somewhat novel approach to the mystery that is subjective consciousness. I do hope that you feel compelled to put it to the test.

References

[1] *Commonsense Computing Initiative*. 2008. http://xnet.media.mit.edu/.

[2] *Otto the octopus wreaks havoc*. 2008. http://www.telegraph.co.uk/news/newstopics/howaboutthat/3328480/Otto-the-octopus-wrecks-havoc.html.

[3] *Wikipedia: Cartesian Dualism*. 2009. http://en.wikipedia.org/wiki/Cartesian_dualism.

[4] *Wikipedia: Church-Turing thesis*. 2009. http://en.wikipedia.org/wiki/Church_Turing_Thesis.

[5] *Wikipedia: Emergence*. 2009. http://en.wikipedia.org/wiki/Emergence.

[6] *Wikipedia: Epistemology*. 2009. http://en.wikipedia.org/wiki/Epistemology.

[7] *Wikipedia: Evolution*. 2009. http://en.wikipedia.org/wiki/Evolution.

[8] *Wikipedia: Justified True Belief*. 2009. http://en.wikipedia.org/wiki/Justified_true_belief.

[9] *Wikipedia: Münchhausen Trilemma*. 2009. http://en.wikipedia.org/wiki/M

[10] *Wikipedia: Occam's razor*. 2009. http://en.wikipedia.org/wiki/Occam%27s_razor.

[11] *Wikipedia: Tractatus Logico-Philosophicus*. 2009. http://en.wikipedia.org/wiki/Tractatus_Logico-Philosophicus.

[12] *Wikipedia: Turing Test*. 2009. http://en.wikipedia.org/wiki/Turing_test.

[13] Hans Albert. *Traktat über kritische Vernunft*. Tübingen: Mohr Siebeck, 1968.

[14] Michael Bach. *82 Optical Illusions & Visual Phenomena*. 2009. http://www.michaelbach.de/ot/.

[15] Lynne R. Baker. The first-person perspective: A test for naturalism. *American Philosophical Quarterly*, 35:327–346, 1998.

[16] Susan Blackmore. *Memes and "temes"*. TED Talks, 2008. http://www.ted.com/index.php/talks/susan_blackmore_on_memes_and_temes.html.

[17] Franz 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.

[18] Roy John Britten. Divergence between samples of chimpanzee and human dna sequences is 5%, counting indels. 2002.

[19] J. Brockman and K. Matson, editors. *How Things Are*. New York: William Morrow and Company, 1995. The referred part from Daniel Dennett is available online at http://cogprints.org/288/0/howmista.htm.

[20] Arkadiusz Chrudzimski. Brentano, husserl und ingarden über die intentionalen gegenstände. *Existence, culture, and persons*, 2005.

[21] F.I.M. Craik, T.M. Moroz, M. Moscovitch, D.T. Stuss, G. Winocur, E. Tulving, and S. Kapur. In search of the self: A positron emission tomography study. *Psychological Science*, 10:26–34, 1999.

[22] Antonio R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace, 1999.

[23] Charles Darwin. *On the Origin of Species*. 1859. Available online at http://darwin-online.org.uk/EditorialIntroductions/Freeman_OntheOriginofSpecies.html.

[24] Donald Davidson. Mental events (1970). *Essays on Actions and Events*, pages 207–225, 1980.

[25] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[26] Richard Dawkins. *The Extended Phenotype*. San Francisco: Freeman, 1982.

[27] Richard Dawkins. *The God Delusion*. Boston, New York: Houghton Mifflin Company (2008), 2006.

[28] Daniel Dennett. *Consciousness in Modern Science*, chapter Quining Qualia. New York: Oxford University Press, 1988.

[29] Daniel C. Dennett. *The Intentional Stance*. The MIT Press, 1987.

[30] Daniel C. Dennett. *Consciousness Explained*. Penguin Science, 1991.

[31] Daniel 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.

[32] Daniel 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.

[33] David DeWitt. *Greater than 98common evolutionary argument gets reevaluated – by evolutionists themselves.* 2003. http://www.answersingenesis.org/tj/v17/i1/DNA.asp.

[34] Niles Eldredge and Stephen Jay Gould. Punctuated equilibria: An alternative to phyletic gradualism. *Models in Paleobiology*, pages 82–115, 1972.

[35] Sigmund Freud. *Das Ich und das Es*. Internationaler Psychoanalytischer Verlag, 1923.

[36] Hagar Gelbard-Sagiv, Roy Mukamel, Michal Harel, Rafael Malach, and Itzhak Fried. Internally generated reactivation of single neurons in human hippocampus during free recall. *Science*, pages 1164685+, September 2008. This article was referenced at http://www.nytimes.com/2008/09/05/science/05brain.html.

[37] Guido Gloor. *The Chinese Chatroom*. 2007. Available online at http://www.haslo.ch/philosophy/chinesechatroom.pdf.

[38] Guido Gloor Modjib. *The Legitimation of Traditional Phenomenology in a Rational-Causalist World*. 2008. Available online at http://www.haslo.ch/philosophy/phenomenology.pdf.

[39] Winfried Hacker. *Allgemeine Arbeitspsychologie*. Bern: Verlag Hans Huber, 2005.

[40] 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.

[41] Edmund Husserl. *Husserliana*, volume XXIV. Dordrecht/Boston/Lancaster: Mertinus Nijhoff Publishers, 1984.

[42] William James. *The Principles of Psychology*. New York: Simon and Schuster, 1890. Available online at http://psychclassics.yorku.ca/James/Principles/index.htm.

[43] Bill Joy. *Why the future doesn't need us.*, volume 8. Wired, 2000. http://www.wired.com/wired/archive/8.04/joy.html.

[44] Amy Kind. *Internet Encyclopedia of Philosophy: Introspection*. 2006. http://www.iep.utm.edu/i/introspe.htm.

[45] Ray Kurzweil. *The Law of Accelerating Returns*. 2001. Available online at http://www.kurzweilai.net/articles/art0134.html?printable=1.

[46] David A. Leopold and Nikos K. Logothetis. Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, 3:254–264, 1999.

[47] George H. Lewes. *Problems of Life and Mind (First Series)*, volume 2. Trübner, 1875.

[48] Carence Irving 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.

[49] Benjamin 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.

[50] John Locke. *Essay Concerning Human Understanding*. London: Basset, 1690.

[51] W. G. Lycan and K. Neander. *Scholarpedia: Teleofunctionalism*. 2008. http://www.scholarpedia.org/article/Teleofunctionalism.

[52] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. The MIT Press, 2003.

[53] Marvin 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/.

[54] Marvin Minsky. *Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. 2007. Available online at http://mitworld.mit.edu/video/484.

[55] Thomas 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.

[56] Thomas Nagel. What we have in mind when we say we're thinking. *Wall Street Journal*, November 7 1991.

[57] Jim 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.

[58] George Orwell. *Nineteen Eighty-Four*. New York: Harcourt, Brace & Co, 1949.

[59] Plato. *Republic*, volume VII.

[60] Plato. *Thaeatetus*. Available online at http://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=1726.

[61] William Van Orman Quine. Two dogmas of empiricism. *The Philosophical Review*, 60:20–43, 1951.

[62] Diana Raffman. *Conscious Experience*, chapter On the Persistence of Phenomenology. 1995.

[63] John 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.

[64] John R. Searle. *The Rediscovery of the Mind*. The MIT Press, 1992.

[65] Scott Sehon. *Teleological Realism*. The MIT Press, 2005.

[66] Wilfrid Sellars. *Science, Perception and Reality*. London: Routledge & Kegan Paul, 1963.

[67] L.R. Squire. Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82:171–177, 2004.

[68] Conrad Hal Waddington. The epigenotype. *Endeavour*, 1:18–20, 1942.

[69] Lloyd Watts. *Cochlear Mechanics: Analysis and Analog VLSI*. PhD thesis, California Institute of Technology, 1993. Available online at http://www.lloydwatts.com/thesis.html.

[70] Lloyd Watts. *Lloyd Watts: Neuroscience*. 2009. Available online at http://www.lloydwatts.com/neuroscience.shtml.

[71] Richard Wiseman. *Ghostly photographs from Hauntings*. 2009. http://scienceofghosts.wordpress.com/.

[72] Ludwig Wittgenstein. *Tractatus logico-philosophicus*. suhrkamp taschenbuch wissenschaft, 1984.