universität wien

DISSERTATION

Titel der Dissertation

"Living in a Natural World"

or "Keeping it Real"

Verfasser

Mag. iur. Bakk. techn. Günther Greindl

angestrebter akademischer Grad

Doktor der Philosophie

Wien, 2010

Department of Philosophy
University of Vienna
A-1010 Austria
guenther.greindl@gmail.com

Studienkennzahl lt. Studienblatt:	A 092 296
Dissertationsgebiet lt. Studienblatt:	Philosophy
Betreuer:	Prof. Franz Martin Wimmer and Prof. Karl Svozil

This Page Intentionally Left Blank

When superior students hear of the Way

They strive to practice it.

When middling students hear of the Way

They sometimes keep it and sometimes lose it.

When inferior students hear of the Way

They have a big laugh.

But "not laughing" in itself is not sufficient to be called the Way, and therefore it is said:

The sparkling Way seems dark

Advancing in the Way seems like regression.

Settling into the Way seems rough.

True virtue is like a valley.

The immaculate seems humble.

Extensive virtue seems insufficient.

Established virtue seems deceptive.

The face of reality seems to change.

The great square has no corners.

Great ability takes a long time to perfect.

Great sound is hard to hear.

The great form has no shape.

The Way is hidden and nameless.

This is exactly why the Way is good at developing and perfecting.

– Daode Jing, Verse 41[i]

i Translated by Charles Muller (LaoziMuller 2004)

This Page Intentionally Left Blank

Table of Contents

Preface

The following thesis is the result of a struggle to find an answer to the question of what it means to be a human being in this world. It is a first attempt at an answer, if answer is the right word in this context; I would prefer to call it a starting point for further investigation. This quest for meaning was the central motivation for my explorations. But why stress the motivational aspect at the beginning of a scientifico-philosophical thesis? Should not a work of this kind be devoid of all motivation and be purely "rational"?

Alas, these are but common misconceptions of what science and rationality are about. Science, as every human endeavor, is performed by living, breathing, in-the-flesh human beings, motivated by very human[1] traits such as curiosity, ambition, creativity, pleasure, spirituality. Thus, whatever rationality and science are about, they are certainly never without vibrant *life*. The task ahead is to take scientific results and try to integrate them into our understanding of life, meaning, and visions of the future. After all, inquiry into the nature of ultimate reality and the meaning of life are one of the main goals of philosophy.

With this set out, some more profane remarks are in order. In the thesis I will try to be very clear about core propositions, so that the reader may judge for herself[2] if she wants to explore the argumentation of the relevant proposition or is prepared to accept it and move on to other topics. These propositions will be highlighted in the following way (the example even contains content):

Thesis in One Sentence

The goal of my thesis is to show that, contrary to what many people believe, a rational, naturalistic and neutral-monist view of the universe is not detrimental to, nay, even *encourages* creative thought and compassionate action.

I do not want to elicit the impression that these propositions are taken to be dogmatic in any way; the emphasis is for the sake of clearness alone, not for the sake of unassailableness.

Concepts should not only be named but rather be conveyed. Only if the concept as such is accessible to all participants of a discussion can real exchange of opinion begin. Therefore I will illustrate some important ideas via graphics – seemingly complicated ideas often become obvious when visualized graphically; diagrams convey structural relationships more easily than words. Left-to-right/top-to-bottom text imposes a linearity on thought which does violence to the very essence of some concepts.

---

1 Nietzsche would have said "human, all too human."

2 I will switch between male and female pronouns for reasons of gender equality.

Acknowledgements

I am especially indebted to the work of Mario Bunge, who has developed a clear and comprehensive philosophical system; it is a sharp sword which serves well to cut through the veils of obfuscation erected by many a philosopher. I owe much to his thinking, though I have deviated in some aspects in the meantime. Concise introductions to the core topics of his work can be found in Bunge & Mahner (2004); Bunge (2006). Also of great influence to my thinking where Brian Cantwell Smith, John Heil, Galen Strawson and David Papineau. I thank Franz-Martin Wimmer for providing the intellectual environment in which this thesis was possible and for the intellectual stimulation his seminars on intercultural philosophy offer.

This Page Intentionally Left Blank

1 Introduction

1.1 One Magisterium

What is this place called the World? How does it work? What is our role in it? Since the dawn of humankind, these questions have troubled thinking minds, and rightly so.

Answers were given. They were most often mythical, magical or religious in nature. The stories, religions and mythologies thus concocted tried to explain and give meaning to the short lives of humans by populating the world with noble spirits, evil wizards and demons, and capricious gods, all controlling nature and the fate of men. Another development were philosophical-anthropocentric systems replacing the magical view of the world in certain places and times. This view placed humanity in the center of things, not spirits or gods. The ultimate truth was not sought in mystical domains of yonder, but in the hidden depths of the mind of man.

Today, there is a new view challenging to obsolete *all* others: the scientific world view. Rational thought and experiment together – neither alone suffices – are taken to give us knowledge – the ratio-empirical method can deliver knowledge, and it alone can deliver knowledge, regardless of the domain[3] of inquiry. Why this is so will be expounded below. That this bold claim should meet resistance from people representing the established mythic and religious views of the world should not come as a surprise.

But it is all the more astounding when adherents of the scientific method also join the ranks of the apologists. And so we must read with disbelieving eyes the the proclamation of non-overlapping magisteria (NOMA) by Gould, a late leading evolutionary biologist:

> The lack of conflict between science and religion arises from a lack of overlap between their respective domains of professional expertise — science in the empirical constitution of the universe, and religion in the search for proper ethical values and the spiritual meaning of our lives. (Gould 1997)

Gould seems to think that there is a separation of a factual domain, where the scientific method is applicable, and another domain where totally different standards apply, namely those of religion.

The problem is twofold. First of all, most religious people would not be happy with this fundamental split. They take their religion to be saying something about the way the world actually is: namely, to bring an example, that humans located in the world have souls; that is a claim about the factual nature of the world; it is an empirical claim. And while Gould is of the opinion that a

3 A domain is a problem-dependent section of reality chosen for closer scrutiny.

soul can't be scientifically refuted[4], part of this thesis will show that for a scientifically minded person, the concept of a soul is indeed very implausible due to purely empirical considerations. So we see that science and religion make contrary claims to a purely factual questions about the nature of the world.

The second problem concerns ethics: while the empirical aspect of science may not derive *immediate* ethical consequences, knowledge derived by science certainly has *indirect* ethical consequences[5]. Taking scientific knowledge and certain requirements of rationality seriously will lead to ethical claims conflicting with those of (Western) religion. Exempting scientific knowledge from applying to the domain of ethics only holds if one has a very reduced view of science in the first place.

An eloquent refutation of the claim to separate magisteria is given by Eliezer Yudkowsky, so I will not try to duplicate the effort but rather include a large quotation. Yudkowsky shows that religions, as originally conceived, saw themselves as quite empirical enterprises:

> The earliest account I know of a scientific experiment is, ironically, the story of Elijah and the priests of Baal.
>
> The people of Israel are wavering between Jehovah and Baal, so Elijah announces that he will conduct an experiment to settle it - quite a novel concept in those days! The priests of Baal will place their bull on an altar, and Elijah will place Jehovah's bull on an altar, but neither will be allowed to start the fire; whichever God is real will call down fire on His sacrifice. The priests of Baal serve as control group for Elijah - the same wooden fuel, the same bull, and the same priests making invocations, but to a false god. Then Elijah pours water on his altar - ruining the experimental symmetry, but this was back in the early days - to signify deliberate acceptance of the burden of proof, like needing a 0.05 significance level. The fire comes down on Elijah's altar, which is the experimental observation. The watching people of Israel shout "The Lord is God!" - peer review.

4 Gould:

> But I also know that souls represent a subject outside the magisterium of science. My world cannot prove or disprove such a notion, and the concept of souls cannot threaten or impact my domain. (Gould 1997)

5 A step in the right direction is taken by Massimo Pigliucci:

> The research program sketched [...] requires ethicists to become conversant in the language of science, especially of evolutionary biology, game theory, comparative anthropology, and neurobiology. Pigliucci (2003)

Haidt (2006) delivers a wonderful example with his book on how scientific evidence can inform thinking about age-old questions such as leading a life of *eudaimonia*, a core topic of ethics.

And then the people haul the 450 priests of Baal down to the river Kishon and slit their throats. This is stern, but necessary. You must firmly discard the falsified hypothesis, and do so swiftly, before it can generate excuses to protect itself. If the priests of Baal are allowed to survive, they will start babbling about how religion is a separate magisterium which can be neither proven nor disproven.

Back in the old days, people actually believed their religions instead of just believing in them. The biblical archaeologists who went in search of Noah's Ark did not think they were wasting their time; they anticipated they might become famous. Only after failing to find confirming evidence - and finding disconfirming evidence in its place - did religionists execute what William Bartley called *the retreat to commitment,* "I believe because I believe."

Back in the old days, there was no concept of religion being a separate magisterium. The Old Testament is a stream-of-consciousness culture dump: history, law, moral parables, and yes, models of how the universe works. In not one single passage of the Old Testament will you find anyone talking about a transcendent wonder at the complexity of the universe. But you will find plenty of scientific claims, like the universe being created in six days (which is a metaphor for the Big Bang), or rabbits chewing their cud. (Which is a metaphor for...)

Back in the old days, saying the local religion "could not be proven" would have gotten you burned at the stake. One of the core beliefs of Orthodox Judaism is that God appeared at Mount Sinai and said in a thundering voice, "Yeah, it's all true." From a Bayesian perspective that's some darned unambiguous evidence of a superhumanly powerful entity. (Albeit it doesn't prove that the entity is God per se, or that the entity is benevolent - it could be alien teenagers.) The vast majority of religions in human history - excepting only those invented *extremely* recently - tell stories of events that would constitute completely unmistakable evidence if they'd actually happened. The orthogonality of religion and factual questions is a *recent* and strictly *Western* concept. The people who wrote the original scriptures didn't even know the difference.

The Roman Empire inherited philosophy from the ancient Greeks; imposed law and order within its provinces; kept bureaucratic records;

and enforced religious tolerance. The New Testament, created during the time of the Roman Empire, bears some traces of modernity as a result. You couldn't invent a story about God completely obliterating the city of Rome (a la Sodom and Gomorrah), because the Roman historians would call you on it, and you couldn't just stone them.

In contrast, the people who invented the Old Testament stories could make up pretty much anything they liked. Early Egyptologists were genuinely shocked to find no trace whatsoever of Hebrew tribes having ever been in Egypt - they weren't expecting to find a record of the Ten Plagues, but they expected to find *something.* As it turned out, they did find something. They found out that, during the supposed time of the Exodus, Egypt ruled much of Canaan. That's one *huge* historical error, but if there are no libraries, nobody can call you on it.

The Roman Empire did have libraries. Thus, the New Testament doesn't claim big, showy, large-scale geopolitical miracles as the Old Testament routinely did. Instead the New Testament claims smaller miracles which nonetheless fit into the same framework of evidence. A boy falls down and froths at the mouth; the cause is an unclean spirit; an unclean spirit could reasonably be expected to flee from a true prophet, but not to flee from a charlatan; Jesus casts out the unclean spirit; therefore Jesus is a true prophet and not a charlatan. This is perfectly ordinary Bayesian reasoning, if you grant the basic premise that epilepsy is caused by demons (and that the end of an epileptic fit proves the demon fled).

Not only did religion used to make claims about factual and scientific matters, religion used to make claims about *everything.* Religion laid down a code of law - before legislative bodies; religion laid down history - before historians and archaeologists; religion laid down the sexual morals - before Women's Lib; religion described the forms of government - before constitutions; and religion answered scientific questions from biological taxonomy to the formation of stars. The Old Testament doesn't talk about a sense of wonder at the complexity of the universe - it was busy laying down the death penalty for women who wore men's clothing, which was solid and satisfying religious content of that era. The modern concept of religion as purely *ethical* derives from every other area having been taken over by better institutions. Ethics is what's *left.*

> Or rather, people *think* ethics is what's left. Take a culture dump from 2,500 years ago. Over time, humanity will progress immensely, and pieces of the ancient culture dump will become ever more glaringly obsolete. Ethics has not been immune to human progress - for example, we now frown upon such Bible-approved practices as keeping slaves. Why do people *think* that ethics is still fair game?
>
> Intrinsically, there's nothing small about the ethical problem with slaughtering thousands of innocent first-born male children to convince an unelected Pharaoh to release slaves who logically could have been teleported out of the country. It should be *more* glaring than the comparatively trivial scientific error of saying that grasshoppers have four legs. And yet, if you say the Earth is flat, people will look at you like you're crazy. But if you say the Bible is your source of ethics, women will not slap you. Most people's concept of rationality is determined by what they think they can get away with; they think they can get away with endorsing Bible ethics; and so it only requires a manageable effort of self-deception for them to overlook the Bible's moral problems. Everyone has agreed not to notice the elephant in the living room, and this state of affairs can sustain itself for a time. Yudkowsky (2007a)

Another simplified sketch of the course of events leading to the idea of separate magisteria could go like this: the most important questions pressing human beings bear no delay; the stone age warrior is as much afraid of death as the modern business man. But in the dark of night people appear who tell of other worlds where there is no pain and no death[6] – let us call them, for the moment, *priests*. The priests gained power because they claim to hold the answers to existential questions, and before the advent of the scientific method, there was no organized movement of critical thinkers to challenge the dogmatic views of the current – at the respective time and place – religion. Indeed, without the knowledge we have today, the religions where even quite plausible. But then the scientific method arose, and the priesthood sensed that they were threatened. The attempt to stifle science failed; a mutual agreement was reached, leaving power hierarchies in place: the scientific method should be used for things concerning this world, and religion concerns itself with the explication of another, more fundamental world: the separate magisteria had been created. Scientists agreed to this contract because their position was weak: they were confronted with a

6 Religion has various functions: emotional comfort, explanation of mysterious facts, and reinforcement of social order among them. For discussion see Boyer (2002).

powerful priesthood having steered the course of men and women for hundreds of years. But now the time has come to challenge this division. Power has shifted, and it is time to renegotiate old contracts.

People still demand answers to their questions regarding the meaning of life, the nature of reality; and many want answers that are not in obvious conflict with reason. Why should we not use our best method for gaining knowledge – the scientific method – for the most important questions we face as sentient beings? Why use traditional religious knowledge which came to power contingently? The religions some of us hold today simply reflect the historical genealogy of events, and not some divine plan. In historical times, where large scale conversions to other religions still "happened"[7];it may have been possible to believe that there is a "true" religion which everybody will sooner or later adopt. But no proponent of a major religion today, be it Christianity, Islam, Hinduism etc can seriously believe that we are converging to the "true" religion.

So it seems that it is time to challenge the view of "separate magisteria". Reality – the world, the universe – does not care about what divisions human beings bring to bear on it. Reality, I contend, is unified in a stringent way, which will be demonstrated in the following thesis with some examples. The scientific world view is essentially one of unification: one in which humans have truly arrived in the universe and are not split from it by an additional property such as a "soul" or "divine nature". There is only one world, and humans are a part of it; thus, science must address questions of meaning.

1.2 The End of Philosophy or Never Ending Philosophy?

What does philosophy have to say in addition to science? Should it not, as religion, be mute in this day and age? Well, not if it restricts itself to certain topics – though it should be stressed that these topics do not make up "separate magisteria", as we saw above. It is rather that philosophy should aim to take a broader perspective on the world than the special sciences, which have settled into *niches of exploration*. The philosopher John Heil markedly says what philosophy should not be:

> Philosophy today is often described as a profession. Philosophers have specialized interests and address one another in specialized journals. On the whole, what we do in philosophy is of little interest to anyone without a Ph.D. in the subject. Indeed, subdisciplines

7 Charlemange, for instance, waged a fierce war against the Saxons – pagans – starting 772 which lasted over thirty years, including mass executions (Blutgericht von Verden) and deportations (Bradbury 2004, p. 22). Christianity was brutally enforced.

> within philosophy are often intellectually isolated from one another. The same could be said for most academic specialities. Historians, literary theorists, anthropologists, and musicologists pursue topics the significance of which would elude outsiders. What distinguishes philosophy is the extent to which philosophical problems are anchored directly in concerns of non-philosophers. Philosophical questions arise in every domain of human endeavour. The issues have a kind of universality that resists their being turned over to specialists who could be expected to announce results after conducting the appropriate investigations.
>
> The professionalization of philosophy, together with a depressed academic job market, has led to the interesting idea that success in philosophy should be measured by appropriate professional standards. In practice, this has too often meant that cleverness and technical savvy trump depth. Positions and ideas are dismissed or left unconsidered because they are not *comme il faut*. Journals are filled with papers exhibiting an impressive level of professional competence, but little in the way of insight, originality, or abiding interest. Non-mainstream, even wildly non-mainstream, conclusions are allowed, even encouraged, provided they come with appropriate technical credentials. Heil (2003, p. vii f)

Unfortunately, browsing through the journals of relevance reveals this analysis to be correct. So, what *should* philosophy be? Ideally, philosophy should be the unifying science – not only philosophy of science, which is a specialized subtask of philosophy, but a synthetic enterprise generating knowledge that is then available for executive use in solving local and global problems which individual humans or the human species as a whole encounter. This work of integration needs to be done specifically. The specialist in the field has a focused skill-set essential for performing cutting-edge research in domains ever more remote from day-to-day life; he does not have the time nor the skill to perform integrative work.

Philosophy, then, could be called the discipline of applying the results[8] of scientific investigation to human beings in a reflexive and critical way. A division of this broad goal into more manageable subgoals may look like this:

8 Not the method: that would be sociology or psychology.

- Criticism of science: Criticizing the methods of the sciences, examining paradigms, improving the scientific process and exploring the nature of rationality. These activities are in the purview of philosophy of science and is where analytic philosophy can and does make excellent contributions.
- Synthesis: Unifying the knowledge of the scientific domains and incorporating philosophical reflections; making sense of this knowledge and what follows for our personal lives – performing the task of world view building (Aerts et al. 1994; Vidal 2007). Synthesis is the necessary counterpoint to analysis. Analysis is concerned with examination, dissection, making distinctions– if one carries this process far enough, at the end one will necessarily be left with nothing. This is what happens in some parts of analytic philosophy: the point of no return is crossed and one is left in a fractured world. The way to knowledge is the middle way, a dialectic, if you like, between analysis and synthesis.[9]
- Ethics: What is a good life? How should we act? The scientific method refined by constant criticism delivers the best knowledge we can hope fore; it is then synthesized in the next step into a coherent whole. Finally, in taking account of self-reflexive processes and asking questions about what we *want*, we arrive at – tentative – answers to ethical questions. If ethics wants to strive to be more than the enforcement of contingent power structures, science is essential, to know both about what is possible in the world and how we ourselves work as psychological human being – *gnothi seauthon*[10]. Of course, in this last part philosophy can't be completely eclipsed by science, because science is first of all descriptive, and in ethics we seek normative principles; needless to say, going further than the results of science does not mean that we will abandon rationality; in a sense, rationality can carry us the last steps where science alone does not suffice. A core task of philosophy then is to create a viable ethics – a domain often deemed the privilege of religion. Under the rubric of ethics philosophers should not shy away from criticizing society, culture and religion.

Some words on the aspect of synthesis are in order. Interdisciplinary dialog is difficult – if not impossible – not for lack of enthusiasm, but because there are no common concepts with which to communicate. An academic discipline introduces someone into a conceptual community – one learns a shared way of carving up a particular domain of the world. This is not to be misunderstood

9 The focus of my work is on the synthetic part.
10 "Know yourself"; allegedly inscribed in the forecourt at the Temple of Apollo in Delphi.

as embracing some form of constructivism or relativism; it is rather a recognition of the trivial but necessary fact that if one develops sophisticated scientific theories capturing lawful processes in the world, they have to be communicated via linguistic entities – shortcuts into concepts – between human beings.

Now, in some disciplines one might get away with merely learning the words without understanding the concepts; that is, talking the talk and not walking the walk. At least that is the impression one is left with after reading some postmodern work. With science, it is different. The difficulty of technically and mathematically oriented disciplines is that one can't eschew learning the concepts behind the linguistic entities, mathematical notation for example, without missing the essence of the science itself.

Now, when somebody has immersed himself with great effort and years of study in a discipline, why should she also "understand" the concepts of other disciplines without going through similar eduction? Of course, there may be overlap – the jump from theoretical physics to mathematics is smaller than from psychology to solid state physics. But nevertheless – there is no shortcut to doing the actual learning, similarities among certain disciplines notwithstanding. Kuhn (1962) calls these things the "disciplinary matrix" of a scientific community, which consists of symbolic generalizations, recognized as natural laws by the community, such as Newton's second law $f=ma$, shared models, values, and "exemplars", that is, paradigmatic solutions of the subdiscipline that serve as primitive examples for solving further puzzles. In this way, scientists with a shared paradigm constitute both a linguistic and a conceptual community.

Let us look at some eclectically chosen concepts from various disciplines:

Discipline	Concept
Physics	entropy
Information Theory	entropy (different meaning!)
Cognitive Science	connectionism
Philosophy	compatibilism
Psychology	affective state
Mathematics	analytic function
Logic	completeness
Biology	Lamarckian evolution

If one lectures in front of a mono-disciplinary setting, one can employ the respective words[11] and rely on the audience understanding what you say, because a concept was associated with the word in their education. The specialist in the field is not only trained to exist in a language community be regurgitating the appropriate words when paroled, but also knows which concepts stand behind them. But in an interdisciplinary setting, this approach does not work. It does not suffice to say that this or that concept exists and is well defined – if the concept is not cognitively present in the individual listener, merely uttering the words is as good as making random noises[12]. This situation demands a resolution, one which philosophers are predestined to deliver: with their training in conceptual analysis and the habit of tackling fundamental problems, philosophers are in a good position to break down scientific concepts in such a way as to quickly convey the essence of the matter in interdisciplinary or public contexts. And actual *understanding* is a crucial distinguishing feature of science versus dogmatic enterprises: if science is presented authoritatively, dogmatically, without elaborating on the conceptual backdrop, it is no wonder that some people think that science requires belief just as religion does.

Heil expresses similar sentiments concerning the philosophy of mind:

> I mention all this by way of calling attention to the absence of formal devices, appeals to purely modal notions like supervenience, and invocations of possible worlds in the chapters that follow. If it accomplishes nothing else, my decision to omit such technical trappings will certainly make the book more accessible to the non-specialist reader. In any case, the philosophy of mind, indeed metaphysics generally, is not — or ought not to be — a technical exercise. Philosophical theses should be expressible without reliance on specialized terminology; and I have tried my best to say what I have to say without resorting to such terminology. This strikes me as an important exercise for every philosopher. Too much can be smuggled in, too much left unexplained when we allow ourselves to fall back on philosophical jargon. (Heil 1998, p. xii f)

The tension between analysis and its jargon and synthesis, then, is also present in philosophy, albeit in a different guise. There are a myriad philosophical schools, asking deep questions, giving different answers, and often vehemently opposed to each other. But why should one be wrong and the other right? Is it not more likely that they simply concentrated on different aspects of a problem

11 In other words: inter-column row jumps in the table above are not permitted.
12 A good account of this problem is given by Yudkowsky (2007b).

or viewed a subject from different perspectives? Leibniz already found it more beneficial to look for synthesis than strife:

> "Die Erwägung dieses Systems zeigt auch, dass man, wenn man den Dingen auf den Grund geht, in den meisten philosophischen Sekten mehr Vernunft entdeckt, als man zuvor geglaubt hat. […] Der größte Fehler, den man begangen hat, besteht in dem einseitigen Sektengeist, vermöge dessen man sich selbst borniert hat, indem man alle andren Meinungen verwarf." (Leibniz 1698)

Maybe, when viewed from high enough, the perspectives begin to converge – when one can see why two seemingly contradicting positions are both right, new insight is acquired. What we should be able to perform is "rapid frame switching" – that is, adopting new robust[13] ontologies, frameworks, stratifications and perspectives as is required by circumstance. Seeming contradiction is then often resolved; some theories capture one aspect of a phenomenon better and vice versa; it is no shame to hold many theories tentatively at the same time. When in the history of philosophy and science did that insight get lost?

Definitions and words should not concern us in our journey: we will develop concepts and give them names, but the important thing is the concept, not the name. Paul Graham expresses my attitude vividly in his essay "How to do philosophy":

> Words seem to work, just as Newtonian physics seems to. But you can always make them break if you push them far enough. I would say that this has been, unfortunately for philosophy, the central fact of philosophy. Most philosophical debates are not merely afflicted by but driven by confusions over words. Do we have free will? Depends what you mean by "free." Do abstract ideas exist? Depends what you mean by "exist." Wittgenstein is popularly credited with the idea that most philosophical controversies are due to confusions over language. I'm not sure how much credit to give him. I suspect a lot of people realized this, but reacted simply by not studying philosophy, rather than becoming philosophy professors. (Graham 2007)

So, words are tools to designate facts, concepts and so on. I will try to make an effort to use these tools considerately and try to be very clear in my exposition, although it is never possible to eliminate all ambiguities for all readers.

13 What a robust ontology is will be explained later on.

But there *is* one sense in which the employed words are important: when we use words which are widely used, we necessarily always import all their connotations; and these connotations influence the way we subsequently reason about a problem. So while it is quite arbitrary if we call something "property A" or "property B", it is not arbitrary if we call something "clean" or "dirty", because the latter two words come with rich emotional associations.

I would like to mention another task of philosophy, which stands a bit on its own, as it is subservient to the goals above. It is the task of conveying the knowledge garnered in the philosophico-scientific process both to scientists and to the general public, at respective levels of sophistication. In the current institutional setup this is only done in a haphazard way – mostly by scientists and philosophers writing popular expositions at the end of their career, with the occasional science journalist chiming in. But the process of "pushing" information is far more important than its neglect would suggest; not only to a democratic public so that it may make informed decisions in the political process; but also to the scientific community, which, it seems, is separated into "non-overlapping magisteria" indeed.

This process of conceptual integration across disciplinary borders and the active propagation of results into the public sphere should, I think, be institutionalized as a core task of philosophy. This is a never-ending process which is at the heart of a critical and open society; never-ending because both the empirical knowledge and the methods of conceptual analysis grow and improve with time; and each time desires answers given in its time. Philosophers divest themselves of this responsibility to their own detriment.

So, what we need is scientific philosophy – that is, philosophy that respects well supported empirical evidence. A scientific philosophy will never be completed; firstly, because our knowledge of the world changes and new knowledge will modify our theories. Secondly, because our technology and environment, and, indeed, we ourselves, change so that different problems will be relevant at different times. A scientific philosophy will therefore never be a *philosophia perennis* – although some results can be expected to be more stable than others. The commitment to philosophy as a process – as a never ending journey of thought – is the most important guard against dogmatism:

> There is no completed philosophy. A closed system is as good as dead.

What must be ensured, of course, is that on this journey of change the possibility for change itself is not undermined. I will take a look at this important problem in section 2.5.4 .

This, then, is the answer to the question of the section title: philosophy, if it embraces the sciences instead of ignoring them, will not end, but is a never ending process. A "first philosophy" or a scientifically ignorant philosophy, on the other hand, has no place in today's science and technology driven global society[14].

1.3 A Scientific Philosophy of Life

> Mit Gedanken, die nicht aus der tätigen Natur entsprungen sind und nicht wieder aufs tätige Leben wohltätig hinwirken und so in einem mit dem jedesmaligen Lebenszustand übereinstimmenden mannigfaltigen Wechsel unaufhörlich entstehen und sich auflösen, ist der Welt wenig geholfen. (Goethe MR)
>
> Alles Gescheite ist schon gedacht worden, man muß nur versuchen, es noch einmal zu denken. (Goethe WMW)

It is time to be done with the general remarks and start to speak about the work at hand. The philosophy proposed here is one which can be lived; it should inform our every action, our every thought, our every living breath. The litmus test of a good philosophy is this: does it change the way we view the world, our life, and, ultimately and most importantly, the way we act? Beliefs we hold, even if they are not concerned with practical matters, should influence our actions in such a way that our actions better contribute to our goals. To deliver such a philosophy in today's complex world is, of course, beyond the ability of a single person. All the more so because a single person can only give a personal view from a certain time and place – a snapshot of his or her thinking, dependent on the idiosyncratic empirical knowledge available at that time; and while philosophy should also look beyond the current state of knowledge, an actionable philosophy should be more modest, lest it succumb to fantasy.

The core goal of my thesis will be to show that hard sciences like physics and neuroscience can integrate well with values important to human beings. When humans accept themselves as creatures of the universe, as parts of nature and not opposed to nature, and that consciousness, love, happiness

14 A remark as to the divide in contemporary philosophy between the analytic and continental tradition is in order. I do not see this distinction applying to my work, as I want to arrive at practical answers to important questions – frankly, I don't care from which tradition a philosopher comes from; either she has something important to contribute or not; unfortunately, in today's philosophical climate such a disclaimer is important, as there are many philosophers who simply refuse to venture into the other tradition "a priori". One could say that I follow the analytic tradition in it's appreciation of the scientific method and reason, and the continental tradition in my "essayistic" writing style and in addressing existential questions, or, put less dramatically, worldview questions. If I draw on mythology and archetypes, I do not pretend to be a scholar in them or imbue them with any mystical meaning – they are rather metaphors for a precise scientific view which may not yet fully be in our reach.

etc are constitutive of the material universe, and not an aspect of a god given *soul* tacked onto the world as an afterthought, humans will have finally arrived in the universe – they will have arrived home.

The thesis will have a close look at ontology – or meta-ontology – as I am convinced that to make sense of the world you have to get your ontology straight, otherwise confusion will not subside. High-level conceptual confusions have their roots in low-level ontological misconceptions. Again, Heil puts it best:

> Attempts to keep philosophy aloof from metaphysics are largely self-defeating. Whether we approve or not, the world has an ontology. Theorists and theories of the world are themselves parts of the world. This homely complication is too often forgotten or ignored by those who regard the world as a construct. If the world is theory dependent, what of theories themselves? Do these stand alone, or does their existence depend in some fashion on other theories ('theories all the way down')? Whatever the story turns out to be it will include an ontology measurable against competing ontologies. […] I want only to note the inescapability of ontology. We can suppress or repress ontological impulses. In so doing, however, we merely postpone the inevitable. Honest philosophy requires what the Australians call ontological seriousness. […] The test of the overall view is not its derivability from uncontroversial truisms, but its power: the extent to which it enables us to make sense of issues we should otherwise find perplexing. Heil (2003, p. 1f)

The reader should, after reading the thesis, be enabled to strive for a more rational, happy, tolerant and critical life. The present work could be seen as a guide book and stepping stone to achieve the above mentioned goals – ample references will aid in this quest, as this text alone can only mark the beginning of a journey. Indeed, what I attempt here is only to set the stage for further investigations; investigations in which others hopefully will join. I can't and won't present a finished solution. The views propounded here should be seen as a vantage point and fulcrum for further explorations. This conception also has a drawback, for which I must apologize to the reader: the thesis is a travel guide into the land of rational analysis of the meaning of life. The pace will be brisk, and much argumentation will be offloaded into the references. This presumes, on the part of the reader, either that she is well read, or that she is willing to read lots of books, or – but I do not

recommend taking this option – that she is willing to accept some assertions *just so*. I am unhappy with this state of affairs, but it is unavoidable at the current time.

Filling in the details into the picture painted here with broad strokes will be the task of a lifetime. Nevertheless, I thought it important to present the big picture first, before concentrating on minor questions. I follow Drescher who engages a similar project but with a more analytic slant than mine:

> Tackling one or another of the subjects examined here is respectable. But assembling so many of them may seem grandiose, except perhaps for my sincere acknowledgement of the tentativeness of the endeavor. In any event, the latitude is not capricious — matters of fundamental importance, whether about physics or about human consciousness, have ramifications that intertwine (or at least appear to), making it difficult to adequately address each such matter in isolation. And I am hopeful that at least some of the ideas I present along the way are novel, interesting, and even approximately correct. (Drescher 2006)

Whereas today one cannot hope to encompass or even oversee all of the available information, which is increasing at an exponential rate, what I will try to do in this thesis is to take a global look at some pertinent results of science and see what they mean for our lives as human beings. The danger of a broad approach is superficiality; the promise is a synthesis. One can never, I think, have one without the other.

Apart from the synthesis, do I expect to contribute something new? At the outset this was of course my intention – though extensive reading has robbed me of the illusion that this were possible. As Strawson says:

> My experience since I first lectured on the 'mind-body problem' in the late 1980s has been one of finding, piece by piece, through half-haphazard reading, that almost everything worthwhile that I have thought of has been thought of before, in some manner, by great philosophers in previous centuries (I am sure further reading would remove the 'almost'). It is very moving to discover agreement across the centuries, and I quote these philosophers freely, and take their agreement to be a powerful source of support. Almost everything worthwhile in philosophy has been thought of before, but this isn't in any way a depressing fact (see p. 200 below), and the local

> originality that consists in having an idea oneself and later finding that it has already been had by someone else is extremely common in philosophy, and crucial to philosophical understanding. Strawson & Others (2006, p. 184f)

And, on the referred page 200:

> ... to realize that there is really nothing radically new in the existing debate — nothing both new and true —, but this is a moment of illumination, not defeat. The fundamental positions in the mind-body debate have been marked out for a long time, and the quality of the present-day debate is embarrassingly lower than it was in the seventeenth century.
>
> It does not follow that there is nothing difficult and important left to do; nothing could be further from the truth. When Pascal imagined someone charging him with lack of originality, he replied:
>
> Let no one say that I have said nothing new: the organization of the subject matter is new. When we play tennis, we both play with the same ball, but one of us places it better. [Footnote 37: c1640-1662, §575. 'One might as well say that I've used old words', he continued. For 'just as the same words constitute different thoughts by being differently arranged, so too the same thoughts constitute a different body of work by being differently arranged'.]
>
> The point is of great importance and holds for all the discursive arts and sciences, even if it has special force in philosophy. The object of philosophy is not just to state the truth in a domain where matters are often so very difficult, but to make it shine out. To think that Pascal's dictum reflects badly on philosophy is comparable to thinking that the best science never produces new results; or like thinking that once someone had painted a picture of the Madonna and Child, or the Montagne Saint Victoire, there was no point in anyone else doing so.

Maybe, in philosophy, there is indeed nothing new under the sun; but it is important that every generation learns the ideas anew, in the vocabulary of the present age. Even if everything has already been thought up by someone sometime, that does not spare us the necessity of repeating the thinking by ourselves.

My intention, then, has reduced to placing the ball a little differently – only time will tell if it has also been placed better.

1.4 Outline of Thesis

1.4.1 Introduction

My thesis consists of three parts with regard to content: the first being on rationality and its import in tackling all questions facing us in our lives, not only a reduced domain of scientific investigation; the second, metaphysical in nature, forming an essay on the nature of the world, especially as informed by the principles of rationality sketched in the previous part; and the third, applying the findings of the previous sections to sentient agents (among them humans) in this universe.

I will argue that this takes nothing away from the richness of our lives, neither in the spiritual nor intellectual sense. The thesis wants to reduce resistance to science (Bloom & Weisberg 2007) by addressing some of the psychological concerns that arise when people are confronted with naturalism.

Another context in which you could view my thesis is "world view building" (Aerts et al. 1994). For a shorter overview, see Vidal (2007).

1.4.2 On Rationality

I will not attempt to lay down "definite" methods of rationality, simply because I don't think they exist – that is, at any given moment and for a given person – simply because there is probably always something more to discover which will also change our perception of what it means to be rational. But for a person who has to decide, at a specific moment in space and time, there will be a current set of principles of rationality. In this sense, I will attempt to lay down the principles of rationality that have helped me arrive at the consequences of the metaphysical essay.

What about logic, probability theory, falsifiability etc? These are methods of reasoning which have been developed over time and which all have their merit. Yet deductive logic is only a constraint – it says nothing about which premises to adopt (empirical science of course steps in at this moment to supply well-supported premises). Falsification is also only a constraint: it serves to weed out false theories, but underdetermination of theories by evidence sometimes leaves too many possibilities remaining. It also seems to hamper scientific progress as unorthodox theories are decried as "unfalsifiable" a priori. Probability theory in the form of Bayesianism offers an interesting approach, but, for every tool of reasoning one can find a problem for which it won't work.

The question now arises: which method do we employ to decide which methods are rational and which ones are not (the meta-method)? Why use this rule and not that one? We need some answer, but giving a definite answer will immediately violate the first desideratum of not laying

down final dogmatic rules. Thus, in this case, it is good to avoid premature formalization and stick with "fuzzy" words.

Our most helpful linguistic servants will be "open minded", "skeptic" and "coherent"[15]. The trick is to leave the definition of these words to the acting community of scientists and researchers, who have to solve the problems at their hands, problems of which we may yet not even know anything. At the end of the day, every individual is responsible to his intellectual honesty.

Delegating this responsibility to the community of scientists is not an excuse – it is unavoidable. "The price of freedom is eternal vigilance"[16], and permanent active critical thinking can't be dogmatized or institutionalized. It is the curse of the enlightenment project that it will never be completed. But the least we can do is to try to improve on our rationality. As such, the enlightenment will also never be defeated, as long as there exist beings who enjoy the method of reason.

The topic of core interest to any student of rationality should be machine learning – the way we mechanize learning on uncertain data; by this, we make explicit the rules we use ourselves when encountering new data. "Expert knowledge" that is gained by long years of training, and called intuition in human beings, is nothing other than a finely trained Bayesian Brain (Friston & Stephan 2007). The Copernican Principle (non specialness of one's own situation) and invariance principles are important, Occam's razor being a generalization of these ideas. Other topics of interest are embodiment and situatedness (Clark 2001) and cognitive biases (Kahneman, Slovic & Tversky 1982). I would also like to stress that emotionality is not opposed to rationality but rather an integral part of it (Sousa 2007).

I will argue that the rational approach is the best way to approach *all* questions facing us in our lives. Only by being rational can we extract as much information from our environments as possible; that is the domain of theoretical rationality. Our best way of gaining knowledge should quite naturally also influence our spirituality, our ethics, our view of the meaning of life and so on. It is important to apply the open, variable standards of rationality to all areas of interest to humans, not only to problems deemed "scientific" a priori. I will especially challenge Gould's theory of "nonoverlapping magisteria" (Gould 1997) for science and religion. While we may not be able to lay down all principles of rationality, we can at least try to be as rational as we can – there is no virtue in doing worse on purpose.

Knowing more is not only good for thinking; it is essential for acting: the agent centric approach is central for my view of rationality (Russell & Norvig 2003) (see also Hawkins (2004)). For individuals, knowledge means having a (hopefully) good mental model of the world: the closer to the actual effective factors in the world, the more potential there is for action leading to achievement of goals. Ignorance condemns one to inaction and passivity. Knowledge, and especially technology as "embodied" knowledge, increases the possibility for action. The litmus test for knowledge – and philosophy – is this: does it change the way we view the world, our life, and,

15 Consistency is a minimal criterion for philosophical thought (paraconsistent logics aside, their concern being better addressed by probability). Coherence is a stronger criterion than consistency.

16 Attributed to Thomas Jefferson; 3rd president of the USA (1743 - 1826)

ultimately and most importantly, the way we act? That is, do our world models influence our actions in the way they consistently indicate, or do we simply hold beliefs as beliefs (Dennett 2006)?

In the view presented in this thesis, rationality, as will be made clear, depends heavily on the physical world. The viewpoint of the thesis is strictly naturalistic and ontologically monistic (see more below). Naturalism and monism are adopted because of rational reasons themselves – this is not circular reasoning, but a necessary way of „bootstrapping“ into the world – a first move every cognitive system has to make, as meaning is built up in the inside of a system relative to its environment (Vollmer 1975). We see that rationality occupies a privileged position despite its heavy dependence on the existence of a natural world. In the end, the concepts are strongly entwined.

Will there be a last justification, a final foundation which everybody must accept? No – that would be reifying truth – the idea that somewhere there exists a magic box which contains a piece of paper, on which, in plain language for everybody to read are written the words: “The ultimate truth is...” I adopt the pancritical stance; in the end, there can be no justification – only the Münchhausen Trilemma (Albert 1991): dogmatism, infinite regress or the vicious circle. The most benign form is to embrace the infinite regress and the circle (evading viciousness as possible) in a positive way: the goal is to be, at least, metacircularly consistent (Drescher 2006) that is, that the methods of reasoning proposed by the theory support the theory at the metalevel (up to infinity) instead of contradicting it when applied to themselves. And to avoid dogmatism we will be radically self-critical – we will examine our beliefs again and again; and – again.

I do believe that when we seek knowledge, that this knowledge represents something outside of us correctly. That is what we call truth. But we admit that we are fallible, so we will never claim to have arrived at the final, absolute truth, but rather that not all stories are equally good, and that some stories – notably those at which we have arrived by critical inquiry – are better than others, that is, they represent outside circumstances in a better way.

One big question remains: is instrumental rationality enough (Nozick 1993)? That is, having fixed goals, being rational means optimizing your ways of attaining these goals. But we humans are often not sure which goals we should pursue – happiness, spirituality, knowledge? Can all three be attained at once? We must apply our knowledge and rationality recursively to define our goals and utility functions. What do we *really* want? This can best be found out by knowing what we are; our goals can then best be reached by also knowing how the world works. This leads to the concept of wisdom.

1.4.3 Metaphysics

I will make two commitments: to naturalism, and to monism (see above). Naturalism is the understanding that the world is not a supernatural place (Sukopp & Vollmer 2007). A demarcation towards religions and superstitions positing entities which are not subject to “natural” laws springs to mind, but this raises the very question of what a “natural law” is. In popular culture for instance, vampires, which classify as supernatural entities, also obey laws, but different ones than humans in the “natural” world. It is for instance not possible for a vampire to survive in sunlight, which is

quite lawful of and by itself.

The supernatural is actually never deemed lawless. The crucial difference between naturalism and supernatural thinking then seems to be that there are separate realms in a supernatural world view – entities belong to different realms, and they are governed (exclusively?) by the laws of their realm. So the central claim I make in this thesis is that naturalism, in any non-trivial or question-begging sense, must mean that there are no separate realms – neither for spirits, gods, ghosts, vampires or minds. So the naturalism I propose here is very closely related to monism.

Monism, the doctrine that there is certain sense of oneness about seemingly differing subject matters, is here proposed in a radical form: the oneness of the world, indifferent to categories of human inquiry and categorization. Central to my thesis is mind-body monism, the idea that mind and matter are not distinct, but different aspects of the same thing (Strawson 2006). What is easily dispensed with is a Cartesian substance dualism, but with non-reductive property-dualism, things are not so clear. The position I take is that of Russellian neutral monism.

The core metaphysical assertion of the thesis is that ontological stratifications can be variously drawn and always come at a price (Smith 1996); that ontic structural realism reduces this price (Ladyman et al. 2007; Esfeld 2009a) and that levelism is false and lies at the heart of many philosophical errors (Heil 2003). Dispositional and causal structure are presumed to be fundamental.

Considerations from philosophy of mind (Putnam 1975; Dretske 1995) lead us to the astonishing conclusion that neither meaning nor qualia are in the head. Meaning and qualia come from causally induced representations of the world. This insight is also bolstered by the concepts of embeddedness and situatedness from cognitive science; see especially Clark & Chalmers (1998).

How can matter configurations derive meaning from other matter configurations? A first tackling of the problem can be found in Smith (1996) – the relevant process is registering, which proceeds through connection and disconnection – abstraction is then seen as a violent but also creative act of information loss on which new things can be built.

One important point I would like to stress is that the ontological pluralism proposed in this thesis is not a relativism in the philosophical sense of "anything goes". The present position occupies a middle ground – it views the world as an open creative universe combined with lawful constraints, which are a precondition for stability and continuity, without which there would be nothing indeed.

Is the view proposed here reductionist? We should discern two levels: the ontological versus the epistemological (other distinctions are possible too). Epistemology is concerned with the question of knowledge, and is thus deeply related to understanding. We can only understand systems at the right level of abstraction – not too low, not too high. Thus, the special sciences will always have a place of their own, serving to understand high-level phenomena at a level which makes sense to us human beings. But understanding, happening in minds and thus part of our being, is only a small subset of all being (the ontological realm). In the ontological realm, we should be quite satisfied with being reductionist; but not in the sense of giving one description pride of place

(for instance particle physics instead of molecular chemistry), but rather in the sense of commitment to the ultimate unity of the world.

I think complexity science will shed much light on these issues, showing how complex systems result from simple laws. We can then call this phenomenon weak emergence (Bedau 1997), which is quite acceptable as it does not introduce additional ontological levels.

1.4.4 On Persons

This part will, surprisingly, be rather short, as the results will "pop out" for free after the work in previous parts. Oftentimes conclusions from science are not applied to human beings (for instance, to preserve "free" will, see below).

The approach I take is different: humans will not occupy any special pride of place. This for the simple reason that there is no clear cut distinction between humans and the world – humans are simply part of the world, not extraneous to it. This may be most easily seen by considering the development of technology, which will soon be used to change the very nature of who we are. Prosthetics are one current application, nanotechnology and molecular biology are contenders on the horizon promising a wide variety of further possibilities to interact with and change „human nature" (Berger & Glanzman 2005).

The core driving ideas of the preceding inquiry – rationalism, naturalism and monism – will lead to certain insights into the nature of persons and identity. Some of these consequences have already been pointed out in the previous literature (Parfit 1984; Noonan 2003), and I would like to add to the discussion, especially in relation to the qualia-problem also addressed by Chalmers (1996a).

Consciousness and qualia, in my view, are not a prerogative of humans, but result from certain matter configurations (Koch & Tononi 2008). A more basic form of qualia will be attributed to all configurations. We see immediately that humans are only a subset of possible persons and this forces us to rethink our place in the universe. Furthermore, ethics will not require persons as subjects, but qualia bearers.

The first ethical category are q-beings, qualia beings (including non-persons), which can be targets of ethical concern . Persons will be defined as q-beings satisfying certain criteria. P-beings (persons) are targets of ethical norms and enjoy full self-ownership (More 1997). There will also be a category of n-beings that are not full persons (children for instance). This is certainly the most delicate category, as these n-beings do not yet possess self-ownership but should be guided to mature self-hood, as opposed to q-beings which will never become a person (such as a toad, for instance). Thus the "n" in n-being stands for "nurture" – the correct attitude to adopt to not yet fully mature beings which have the potential to become persons. I would like to stress that n-beings should not be considered in any way less valuable as p-beings; they only do not yet enjoy full self-ownership due to obvious reasons (not having completed cognitive development, for instance).

If we assign ethical value to q-beings, there must be criteria for deciding what has qualia meriting ethical consideration and what not (Chalmers 1996b). An interesting approach can be

found in Tononi (2004). He purports to measure relative information integration as an empirical measure of consciousness. His assumptions conform especially well with the metaphysical positions adopted in the current work. This all leads to an ethical conception which avoids any form of discrimination according to non-relevant criteria.

Furthermore, it follows from the metaphysics that the ethics propounded here will only contain hypothetical imperatives, not categorial ones.

A core claim is that there is no essential self, no ghost in the machine; but that the self arises as a matter of representation (Metzinger 2003; Sloman & Chrisley 2003; Pollock 2008).

Finally, the concept of free will, where the word "free" actually means anything, is shown to be largely incoherent: it presupposes the mentioned disconnection of humans in regard to the universe. I will offer a different concept, *optimal will*, to replace "free will". Achieving optimal will is a difficult process; but one that is worth pursuing and which will achieve that what (in the limit) philosophers probably have in mind when speaking of "free" will (see Pereboom (2001a) for a view very related to mine).

1.4.5 Conclusion

The ideas which have been expounded above may seem strange to modern ears; but they find resonance in ancient texts such as the Dao De Jing – Smullyan (1977) offers a light-hearted and unconventional introduction. These ideas just go against deeply entrenched Western prejudices, stemming from Aristotelian and Judeo-Christian roots; prejudices which have come in conflict[17] with science and should be abandoned.

In the universe meaning arises only in relation to the environment. That does not make it relative in a nihilistic way, but rather in an exploratory way. We know that here and now, certain sets of values befit us and our situation. But it is also possible to transcend them, and the more we transcend our current limitations the different our values will become.

We should therefore proceed with *Erkenntnisoptimismus* – the willingness to know, to learn, to change, to understand, to help, and, in all that, never forgetting to be happy (Haidt 2006). We should strive to build ever more towers of abstraction (ideas, technology) which enable a richer experience of life. In closing, a deeply rational, naturalistic view of the universe leads not to despair, but to hope and to a vision for the future. It seems that we indeed live in the best of all worlds (Leibniz 1710).

17 It may be that these very roots gave rise to science in the first place. Nonetheless, now they have become a burden.

2 On Rationality

2.1 Introduction

The goal of the present chapter will not be to define or lay down definite rules for what it means to be rational. That is well beyond the scope of a thesis, and, I think, beyond the scope of current philosophy. The rationality advocated here should also not be viewed in the tradition of the rationalism[18] of Descartes, Leibniz and Spinoza, who are more aptly called intellectualists (Bartley 1984, p. xxvi). The approach advocated here may best be called ratio-empiricism (Bunge 2006); but the word is a bit awkward, so I will use rationality and ratio-empiricism interchangeably in the text. Rationality is a judicious mixture of empiricism and reason, inseparably intertwined. Disregard empiricism, and reason loses its grounding in the world and will drift off into the world of poets and fantasts; empiricism alone leads into the strange dream-landscapes of phenomenology and anti-realism; both approaches may lead to idealism[19]. The middle way – reason grounded in the real world – is the one I take in this thesis[20].

The goal of this chapter is firstly to lay down, in a heuristic fashion, what I do and don't mean with rationality; and, secondly, to connect with the relevant literature. "Rational" is one of those words which probably possesses different overtones and connotations for each reader. For this purpose, a clarification of what is meant by "rational" is in order.

I will try to both look at the prerequisites for rationality and to home in on core aspects of what it actually means to be rational. My thoughts are heavily influenced by artificial intelligence (AI) research, which has made more advances on the topic of cognition in recent years than much of the armchair philosophizing on the nature of reason and intelligence before. That is the case because it

18 Seventeenth century rationalism suffered from ascribing too much reliance on the power of introspection and a conception of disembodied reason, which allowed it to engage into flights of pure fantasy which were nevertheless presented as certainties. Funnily enough, pure empiricism can also degrade into anti-scientific irrationality quite quickly, because it puts too much emphasis on the idiosyncrasies of the human sense organs. The abyss into which empiricism is in danger of falling is skepticism. For a truly rational world view, it is absolutely essential to take *both* perception and reason into account and not to concentrate on one or the other aspect alone.

19 Idealisms are a form of madness. But that is for another day.

20 One could call it the scientistic stance:

> The scientistic stance […] incorporates elements of both empiricism and materialism. It retains some features of empiricism, namely the disdain for demands for explanation, and their satisfaction by posit, and the hostility to non-naturalistic metaphysics. These are just what make empiricism amenable to the secularist who regards the religious picture of reality as empirically ungrounded metaphysics posited only to explain the existence of the world, the meaning of human life and the persistent mythologies of certain cultures. From materialism, scientism takes the idea that we should have metaphysical picture of the world to discipline scientific methodology, and science and education policy, and it should be the one that incorporates and is most well-supported by the scientific image. […] According to the above, the true empiricist and materialist stances are compatible and their synthesis is a scientistic and secular worldview. (Ladyman Forthcoming)

is very easy to fool ourselves about our mental capabilities, introspection being a notoriously unreliable companion in the quest for self-knowledge. On the other hand, when trying to actually build something which shows intelligent behavior, we are forced to think clearly about the most subtle of issues.

To be able to talk of rationality in the sense of the present text, three conditions have to be met:

- Goals

  Without goals, there can be no rationality. This is the case both for theoretical as for practical rationality. Theoretical rationality concerns rational belief formation, and practical rationality concerns right action, given desires and beliefs. For instance, truth is a goal for those who seek knowledge; others will settle for empirical adequacy[21]. A general goal for people in everyday life is to acquire beliefs which give meaning to one's life, irrespective of their truth value. These goals will influence the rationality of acquiring *further* beliefs heavily. For someone interested in truth other items of knowledge will be relevant than for someone who wants to maximize other cognitive values.

  Similarly for practical rationality: what seems right for us to do in certain situations depends on the goals and desires that drive us. In human affairs, goals are often implicit, such as the goal of general happiness, and thus not noticed consciously by the actors, as opposed to more explicit goals such as earning money, eating when hungry etc. Negative goals are also possible goals, such as not wanting to live anymore. As human beings, we have the ability, at least sometimes, to reflect on ourselves and our goals. The most important question is of course which goals to adopt? The pursuit of this question – and the arrival at and adoption of answers – is usually called wisdom. Self-reflective agents and goal-reflective agents are special cases of agents, and ideally humans should strive to be of this sort. For the moment we can adopt a simplistic approach here, considering that goals such as happiness, freedom from pain, satisfying social relationships and so on are innate to humans. The first tenet of rationality is: without goals, there can be no rationality.

- Agency

  As there are no disembodied goals, beliefs and actions, contrary to idealist conceptions of the world, we posit a new abstraction, an entity which has goals and beliefs and which

21 Children are fascinated by the world, they want to become scientists because they want to know how the world works – they are looking for the truth. It takes a philosophy degree to lose this interest in the world and settle for empirical adequacy.

performs actions: an agent. Also, contrary to Aristotelian metaphysics, in which everything may be seen to strive towards its final cause, we will not speak of agents of little sophistication as trying to achieve goals. A stone, a bridge, a house – here there is no question of rationality. Our requirement on agents is that of material entities of sufficient complexity. We need not concentrate on the question of how much sophistication is sufficient; for the time being, we will be satisfied with the sophistication of humans and maybe higher mammals. This, then, is our second tenet of rationality: there is no rationality without an agent.

- Environment

  But now that we have introduced agents, the next question immediately arises: where is this agent located? The agent has to be situated in an environment – the alternate case, that the agent has no environment, could be fulfilled if the agent were a universe of its own. This would quickly destroy our project of defining rationality in a sensible way; as the universe does not have to react to anything outside itself, and can not acquire any beliefs about things outside itself – an agent which encompasses the entire universe can maximally introspect and draw rational inferences, but nothing much of our agent concept would be left. So, to be rational, we require an environment in which the agent is embedded. It is *about* the environment that the agent wants to form beliefs, it is about himself *in relation* to the environment that the agent has goals, and it is *on* the environment that the agent acts. Rationality, again, is not disembodied: it depends essentially on the environment. The environment for us humans is the physical world, the universe. What this environment is like will concern us in the metaphysical analysis; it is of utmost importance as this environment defines which goals are possible and desirable for us. I would like to raise an important point at this stage already: that the distinction between environment and agent is perspectival and does not represent a fundamental ontological truth about the world.

So, our preconditions for rationality are these:

- Goals
- Agents
- Environment

Firstly, we will turn to most simple of the three preconditions for rationality, agents. The environment and the goals both merit their own chapter.

2.2 The Agent Perspective

Here I will present a simplified model for what it means to be an agent, which will then be our working model for the rest of the thesis. It will be a *handle* to which we can attach our other concepts. I will adopt the model presented in Russell & Norvig (2003) which is quite modular and adaptable for numerous purposes. An agent, as far as the following discussion is concerned, is an entity which has some degree of "separation"[22] from its environment, which receives input from the environment via perceptors and is able to act on the environment via actuators.

The basic definition of rationality will be taken in spirit of the agent based approach:

> What is rational at any given time depends on four things:
>
> - The performance measure that defines the criterion of success
> - The agent's prior knowledge of the environment
> - The actions that the agent can perform
> - The agent's percept sequence to date
>
> This leads to a **definition of a rational agent:**
>
> For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has. (Russell & Norvig 2003, p. 35 f)[23]

Space does not permit a full discussion of the agent-model; the reader is kindly referred to the excellent text of Russell and Norvig for further details. For present purposes, a sketch of the most important points must suffice.

The performance measure can be either intrinsic to the agent or given by external circumstances. In the human case, we have intrinsic measures (e.g. happiness) and external measures (e.g. workplace evaluation, which then impacts intrinsic measures). We can equate the

22 Again: this generalization will be relativized in non-trivial way in the metaphysical part of the thesis.

23 As mentioned in the introduction of this section, this basic model is much more conducive to a clear analysis of human rationality than the wide ranging philosophical discourse on the topic, which is still largely preoccupied with fighting historically received misconceptions. This is not to say that the philosophical discourse is negligible – it will occupy us in due time; it is just that it is not as conceptually clear as the agent-based approach devolved from AI-research.

performance measure with the agent's utility function – an utility function represents the agent's internal evaluation, that is, intrinsic measure, and that is what we are primarily interested in. The utility function assigns numbers to the ordering of preferences of an agent. I will not commit to utility theory or decision theory in this thesis in any paradigmatic way; there lie pitfalls which would require a thesis of its own[24]. Nothing special hinges on the adoption of a *specific* model of rationality here.

The prior knowledge and the percepts represent the agent's knowledge. Prior knowledge is acquired by previous experiences, and is made possible by the agent architecture providing the preconditions for experience, to add a little Kantian flourish. In animals the preconditions for knowledge are embodied in the morphology of their bodies shaped by natural selection.

The percepts are due to subconscious preprocessing functions filtering the highly dynamic inflow of sensory data from the environment. Filtering occurs at many places: the sense organs themselves and their physical makeup, and, to take animals as examples, nerve excitations and brain processes performing evaluations higher up in the cognitive chain.

A rational agent, would then, *given* his preferences, and *given* the actions he can perform, act in such a way as to *maximize* his utility function given all the knowledge he *currently* possesses. Let us have a closer look at such a utility-based agent:

24 The literature is populated by paradoxes such as the St. Petersburg paradox, which expounds the problems of infinite expected payoff (Bernoulli 1738) (see Baron (2008) for a current exposition), *utility monsters* which monopolize utility (Nozick 1974) and repugnant conclusions which distribute utility so much that it is beyond recognition (Parfit 1984). These are problems for specific theories of rationality, not for a comprehensive take on rationality per se – after all, solutions to the problems will again have to conform to standards of rationality.

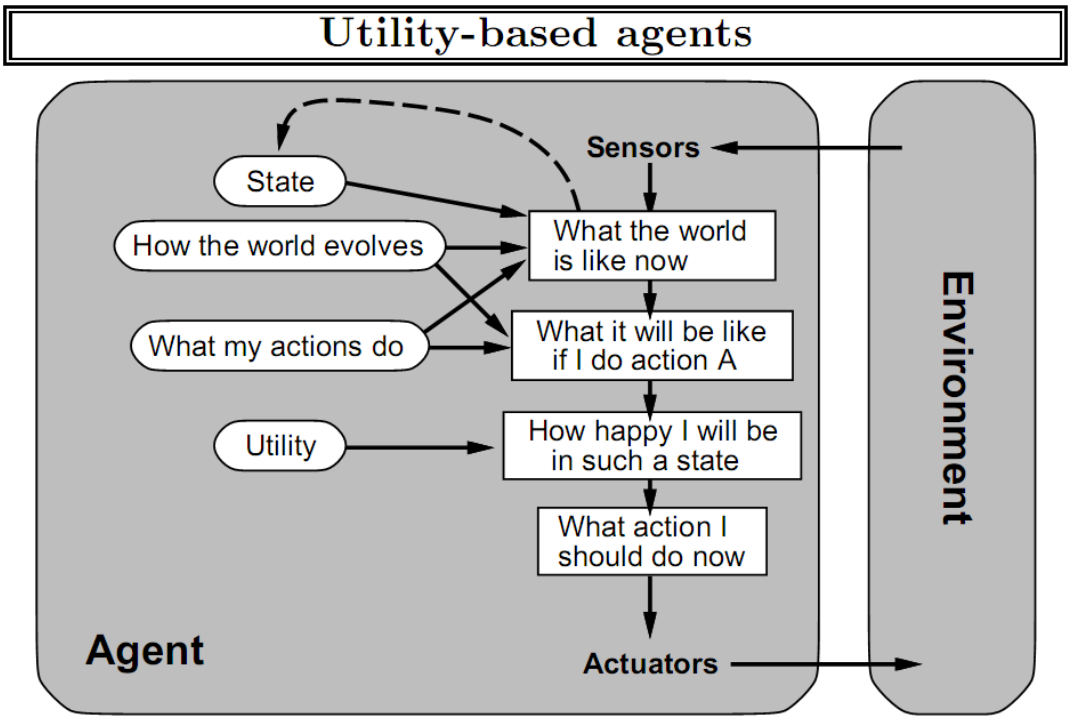


Illustration 1: Utility-based agents (Russell & Norvig 2003, p. 52)

The utility-based agent model is attractive because it comes at a level of abstraction that is ideal for further discussion. The agent, in addition to sensors for receiving percepts and the ability to perform actions via its actuators, has an *internal* model of the world, which does not only correspond to current percepts but also to background knowledge[25], and, which is, most importantly, not static[26]. The agent needs to have representational capacity of a certain sophistication to form complex internal models – that is, it must have physical subsystems which allow complex encodings (in humans this functionality is supplied by the brain, but things can't always be separated so easily – more on that below).

The agent has ideas of which laws govern the dynamics of the world, that is, its time evolution[27]. In its considerations of the world's time evolution, the agent need not only take into

25 That is, innate knowledge and prior percepts integrated into a model. The possibility of building models in our brain is due to our evolutionary heritage, more on that in section 2.3.1.

26 In the diagram, rectangles represent internal states in the *current* decision process and ovals represent background information used for this concrete process – they are, of course, also encoded in internal agent states.

27 This incidentally highlights the fact why scientific inquiry is so important: science supplies the laws of dynamic world evolution, and the better the agent's internal representation of the dynamic rules of world evolution, the

account the dynamic laws which evolve the current state into further states, but also how his actions would influence the world – an instance of counterfactual reasoning:

if I would do A, then, in conjunction with world evolution E, B would happen

if I would do C, then, in conjunction with world evolution E, D would happen

The agent can form various plans, and then needs to compute the possible outcomes of his plans *conditioned* on probabilities representing the uncertainty of the future.

if I would do A, then, in conjunction with world evolution E, B would happen with probability 10%

if I would do C, then, in conjunction with world evolution E, D would happen with probability 50%

The assignment of probabilities will most often involve guesswork – based on intuitions which are the result of previous experience (see more on intuition and how it relates to rationality in section 2.4.2). The agent then has to evaluate the possible outcome states for utility and implement those actions which maximize utility.

So, to sum up, the agent has the following at its disposal:

- an internal model of the world, encoding its current state; the current state is derived from percepts and background knowledge
- dynamic laws governing the evolution of the world state; ideally derived from (at least indirect) empirical observations
- the ability to form plans, that is, construct actions *sequences* that impact the world and have different outcomes
- evaluation of the outcomes according to the agent's internal measure of utility, taking into account the likelihood of the success of the plans; not every plan has equal likelihood of successful completion.

A simple example shows this framework in action:

Example1:

Percepts:	table, cup of water
World model:	tables and cups and water are objects behaving according to "folk" physics; where "folk" physics stands for the dynamic world evolution laws the agent has internalized

further ahead it can reliably evaluate action outcomes.

Utility:	the agent is thirsty, and would like to stop being thirsty, because not being thirsty is intrinsically valuable for the agent, thanks to evolution.
Planning:	Plan A: do nothing. That would not change situation.
	Plan B: If the agent would pick up the cup and drink the water it would not be thirsty anymore.
Decision:	implement Plan B, as it has greater utility.

An important extension of the utility based agent is the learning agent (Russell & Norvig 2003, p. 53). The details of that agent model need not concern us here. The reason why I mention it is that I would like to stress that no aspect of the agent – neither the utility function, the world model nor any other aspect need be static. All modules are available for *updating* via learning. Learning is simply the process of changing the inner structure of the respective module (i.e., adaptation) dependent on outer certain stimuli. More on this in section 2.6.1.

A human agent differs in many regards from the simplified utility based agent sketched above[28] – but unfortunately often to the worse. Humans, for instance, often do not know how to maximize their utility – we perform actions which make us unhappy in the long run because of short term benefits. Humans are often guided by unconscious patterns and rules, so that they seem more like agents which have fixed scripts and action potentials assembled in a haphazard way available for certain stimuli, rather than being full-fledged utility-based agents, this being due to our evolutionary brain architecture. Being rational – or, at least, trying to be as rational as possible – is the attempt to overcome this state of affairs and move into the direction of the utility-based agent.

Anyway, the simple agent model above supplies us with enough of a scaffolding to support the discussion of rationality in the rest of this chapter. And that is why we now move on.

2.3 Natural Rationality

2.3.1 Bootstrapping

The essential philosophical, ontological and epistemological problems which have concerned philosophers for the past millenia are all deeply connected to the fact that we have to bootstrap ourselves into *knowing*, also expressed by the platitude that "you have to start somewhere". Our goal will be a pragmatic one: to reason in this world, not in all possible worlds.

28 An illustrative attempt to model the human mind is the H-CogAff architecture developed by a team in Birmingham around Aaron Sloman. A diagram can be found on the web: http://www.cs.bham.ac.uk/~axs/fig/your.mind.png. An overview of the model is given in Sloman (2008).

In a sense, Descartes' "cogito ergo sum" is still our starting point in the bootstrap, although we may choose to use less contentious words such as "perceptio ergo sum". So, let us accept that our awareness evidently guarantees our existence; skeptics who deny even this[29] are beyond reach.

There is of course a little problem tucked away in our being thinking and perceiving beings: how it is possible that we can perceive and think in the first place. The underlying problem was clearly recognized by Kant, who discerned *time* and *space* as preconditions for perceiving and the *categories* as preconditions for concepts and higher thought (Kant 1781). With Darwin's theory of evolution and its modern synthesis (Huxley 1942; Mayr 1993; Mayr & Provine 1998) we can take the mystery out of the *a priori*[30]: the innate or a priori for the individual represents phylogenetic learning, which is of course *a posteriori* on the species level. The a priori of an individual represents the adaptation of a population tracking its environment. The process of evolution delivers the *initial bootstrap* – we do not come into the world without a history. We are all caused beings, forged since the beginning of time. For a detailed account of evolutionary epistemology see Vollmer (1975); Vollmer (2003). The problem of realism, which remains, as evolutionary fit need not be the kind of fit necessary for semantical truth, will be addressed in section 2.9.5.

A critic will simply point out that we can't avoid circularity by pointing to the process of evolution, which is itself known to us only a posteriori. This is true, but I do not consider it harmful. Not every circle is vicious, and however we construct our knowledge, we have to presuppose one thing or other. The main point of criticism launched against the phylogenetic a posteriori will certainly be the theoretical underpinnings of the theory of evolution; but there are theories which gain plausibility quickly simply due to logical reasons alone and others that need a wide range of empirical supporting evidence. The theory of evolution is of the former kind, the theory of relativity of the latter. Einstein's theory of relativity is generally accepted, despite being highly counterintuitive and needing experimental verification of the sort not easily accessible to the layman. To understand special relativity, one has to have mathematical concepts of a certain sophistication at one's disposal, those representing the non-trivial theoretical background of the theory. Contrariwise, the theory of evolution is heavily contested by certain groups, despite its being more easily grasped and with less theoretical assumptions than the theory of relativity. The moment we have heredity, variability and selection[31], evolution is a fact of logic. Everybody can check this

29 And I have met them.
30 Kant's *Critique* (1781) preceded Darwin's *On the Origin of Species* (1859) by close to eighty years.
31 And it is difficult to deny these three processes, even for the die-hard creationist.

with simple computer simulations[32]. So, let us for the moment accept evolution as the bootstrapper and look at the process of bootstrapping itself.

A short account of how a bootstrap works on a computer may illuminate the *analogous* problem for human beings; the account below is to be seen metaphorically. A computer is made out of physical parts, the hardware. But what makes it so useful as a tool is that it is programmable. Being programmable means that one can direct the causal flow of the machine without actually having to rewire it physically; the instruction of the machine is performed be inputting a certain sequence of symbols.

To this end, the computer need be prepared to be able to *accept* the input of symbols and subsequently use them to direct its own calculations. That is, we need a bridge from software (the symbols which a human enters) to hardware. An oversimplified sketch will do: computers today have a BIOS, the basic input output system, which are basic code instructions specific to the machine architecture and stored on a memory chip on the mainboard, a central part of a computer's hardware. When the computer is turned on, electric current begins to flow, at first following the physical pathways present in the machine; the machine is set up in such a way that the basic code of the BIOS is activated and get control of the machine, by "directing" electric flow through switching gates. The BIOS now "recognizes" the hardware, such as video card and hard disk, and from the latter further instructions, such as operating system code, can be read. The machine is escalating levels of complexity which it can process. When the operating system is loaded into active memory, the system is ready to interpret user input, produce output, operate on data etc.

Only the laws of physics, here especially electromagnetism are at work; the design of the computer hardware and the BIOS code are *structural* causes for the correct operation. The BIOS and the hardware were of course designed by human beings, who began with simple principles, gathered increasing know-how, and organized this know-how into structure visible as hard- and software. We can trace a causal history from every existing computer back to the first timid experiments of pioneers with vacuum tubes and electric wire. Nowhere does "magic" enter – causality rules supreme; causality which aggregates into ever more complex structures.

We humans, to make sense of the world, also have to undergo this bootstrapping process; but the work has already been done for us. Our hardware, our bodies, and our BIOS, the basic brain structure which has the disposition for perception and learning have been shaped by evolutionary

32 Tierra (http://life.ou.edu/tierra/) is one package out of many.

forces in the past 4 billion years. Those organismic "solutions" which receive input from the environment, process it and produce sensible output, that is, *live successfully* reproduce, and thus stay to leave their mark on the face of the Earth.

We are not always rational. Sometimes we behave plain stupid, contrary to our interests: the stupidity we witness in ourselves and in others is the result of old, violent and powerful energies: the unleashing of the fire of the Big Bang, and the tumult which followed thereafter. It is the raw and blind energy of primordial time encoded in our brains. Rationality – reason – is the product of a cooling universe, in which order and stability begins to prevail and brains learn to track this stability.

From the initial bootstrap we will now move on to the source of rationality – why is some behavior rational and other behavior not?

2.3.2 Instrumental and Non-instrumental Rationality

As we saw before, to be rational requires goals, and, in the other direction, if one has goals, instrumental rationality is the search for the optimal way to reach these goals. If one renounces rationality, one is renouncing one's goals – that position is incoherent or at least very nihilistic.

So, everybody who is less than instrumentally rational is content with not being prepared to satisfy his goals and desire. A skeptic could argue that wanting to reach one's goals has to be rationally justified itself, and in this way avoid entering the rational mindset, but that argument is a case for section 2.7.1 on the Münchhausen Trilemma; anyway, for the naturalist, this does not pose a problem, as this *wanting* is a given primitive through evolutionary selection pressures; again, one has to start somewhere.

If one says what rationality actually is – trying in the best way to achieve one's goals – it is quite clear that we need not talk anymore about if we should be rational, but rather, which strategy is rational in the current situation. For instance, sometimes in game-theoretic situations[33] it is good to randomize one's behavior – random behavior is then a rational prescription. Adopting random behavior or seemingly "irrational" strategies because rational deliberation has come to the conclusion that this would be the best strategy to adopt are not counterexamples to rationality but cases in point. The rational strategy is determined by the environment – it rests on physical facts

33 The "El farol" bar problem is a classic example (Arthur 1994).

about the world. That is why instrumental rationality is tightly connected to one's (hopefully correct) physical conception of the world.

The first credo of the rationalist is simple: I want to choose that action that lets me reach my goals; I will not follow normative constraints which will let me behave sub-optimally for the problem at hand[34]. To quote Miyamoto Musashi[35]:

> The primary thing when you take a sword in your hands is your intention to cut the enemy, whatever the means. Whenever you parry, hit, spring, strike or touch the enemy's cutting sword, you must cut the enemy in the same movement. It is essential to attain this. If you think only of hitting, springing, striking or touching the enemy, you will not be able actually to cut him. (Musashi BFR)

Is instrumental rationality – the linking of means with ends – enough? Nozick lets the instrumentalist quip: "Enough for what?" (Nozick 1993, p. 133)[36]. The difference between instrumental and non-instrumental rationality is one of human categorization and therefore not fundamentally important: some goals are construed as not being mere means; and a different kind of rationality is required to deal with those goals. For instance, it is commonly held that the instrumental rationalist can't be used to reflect on the goals itself, and for this, one needs either a more encompassing kind of rationality or the questions are seemed to lie beyond rationality altogether. But this presupposes a disembodied metaphysical picture: one where agents do not have any a priori desires. We humans are evolved and come prepackaged with a lot of goals and desires. When we follow our a priori goals with instrumental rationality, we can arrive at *values* via reflection and with insight into the nature of reality – we arrive at a value bootstrap, which is again delegated to evolution.

The problem of value rationality could be called, paraphrasing Chalmers[37], the hard problem of rationality. But it is when we have certain cognitive values – such as that of acquiring true belief – combined with our ability to feel pain and pleasure – that we can devise a rational ethics, one based

34 As two-boxers in Newcomb's problem would; see Drescher (2006) for a concise account.

35 A Book of Five Rings ("Go Rin No Sho"), english translation: http://www.miyamotomusashi.com/gorin.htm. Miyamoto Musashi was a superior Japanese swordsman. Eliezer Yudkowsky made the brilliant connection of samurai swordsmanship with the art of rationality (Yudkowsky 2008a).

36 He then proceeds to delineate some additional rational factors – symbolic and evidential considerations – contributing to utility weighting in addition to purely causative, that is, instrumental considerations. I will not follow Nozick in this; a detailed refutation must await another paper.

37 Chalmers speaks of the "hard" problem of consciousness (Chalmers 1995).

on instrumental rationality and given natural facts alone. Their need not enter any non-natural normative facts; more on this in the last chapter.

David Papineau places means-ends rationality on a firm analytical footing regarding the normativity of conceptual judgments; he highlights that the value of truth is a personal or moral value, and thus that judgments that prescribe certain forms of judgments (that is, are normative) are no different from other forms of means-end rationality. This lets epistemological norms also be seen as hypothetical imperatives. The essential point to understand for the analytically trained philosopher is that the account given here is not deficient if one takes a naturalist theory of truth as a starting point[38]; normativity is explained in terms of truth, not the other way round (Papineau 2003, p. 6f and p. 10f). And this footing can be supplied for all other domains too.

So, at the end of the day, who says what is rational and what not? The core proposition of this section is simple:

Normativity lies in nature (environment, other agents etc).

This is not a problematic conclusion, only difficult to accept, because it entails both an acknowledgement of the natural world as ultimately constitutive for our thought, and a burdening of responsibility: because there is no authority anymore who can ordain what is right or wrong – and nature is not an authority, it simply is the way it is – we can either accept that we are part of this universe and try to achieve our goals, or deny this. It is very much about responsibility and freedom. It is the simple insight that when one wants to act in this world and attain goals in this world, one has to submit to the natural laws that govern this world. To be clear: the laws to which one need submit are not the social norms or legal laws (which can nonetheless be important and mostly are) but rather those basic laws of nature which enable the world to be such as it is (that is, they ensure persistence and regularity etc).

Again, where does this normativity come from? Laws of nature lay the grounding for the laws of thought – as our thoughts track, through natural law, other natural laws, a harmony arises – because of this harmony I think one could call rationality the "Dao of thinking". The Dao is Chinese and can be roughly translated as "the way", and in Chinese thought, one has attained the Dao when one is in harmony and balance with the universe.

Normativity arises from reality: real goals, and working solutions to real problems. It is not a normativity as traditionally conceived: an authority commanding what should be done. It is a

38 Such as teleosemantics (Macdonald & Papineau 2006).

pragmatic normativity: either follow it, or fail. It simply is. Pragmatism should stop being a derisive word in philosophy.

What we can construct in this place is a schema for rational action:

Constraints: how the world is or can be considering physical knowledge: C_i

Goals: how we want the world to be: G_i

Rational action: If G_i is an attainable state according to C_i, bring current world states W_x into alignment with G_i through actions A_i, where all A_i[39] have to satisfy the constraints of C_i.

So, again to paraphrase Chalmers, goal rationality could then be called the easy problem of rationality. It is heavily constricted by the environment, which usually also consists of other agents and their beliefs and what exactly one wants to achieve. Philosophers who search for justification here will fail: there is none (see also section 2.7.1). To be more precise: there is no justification to be found in words or concepts. Words and concepts will always only find other words and concepts, while actions and beliefs find their "justification" in their success in the real world. For instrumental rationality, justification (and thus normativity) lies in the way the natural world simply is.

2.3.3 Reasons

We have looked above at the normativity of rational strategies. In the agent-based definition above, we can imagine rational agents that act without reasons. But we humans, having the capacity for higher-order cognition, usually have and are moved by reasons. For humans, debating about strategies is often done via reasoning, either internally with oneself or in a community.

But from where comes the power of reason? The answer is the same as above, but I would like to mention it explicitly. If being rational means to be prepared to be convinced by reasons[40], we may ask what reasons *are*: or rather from where does the normative power of reasons originate? We should distinguish a pragmatic aspect and the real normative aspect.

First the pragmatic aspect. Reasons for you are reasons for me only if I can make sense of them. Reasons are propositions where the reasoner says "aye, t'is true, and from this I know that something else, X, is also true". But from where does this sense derive? For a Polynesian hunter-gatherer, the conservation of energy will hardly be a reason he can accept to exclude some events from happening. It is not a reason for him because it does not fit his world model. Accepting

39 Now, the interesting question is of course how to find out the correct A_i; that is the job of science and technology.
40 As opposed to arguments, which may or may not instantiate reasons.

reasons depends heavily on our environment and the theories we have formed about it, that is, our experience and the way we organize our experience. Mutually intelligible reasons depend on a shared environment and a shared theoretical organization.

That is not to say in any way that reasons are arbitrary – in fact, one could in part describe the goal of science as the search for *good* reasons; as we can't change our environment, at least not at the fundamental level of natural law, that means science is about finding the best ways to organize the experiential substratum into knowledge – one that is consistent and coherent. So, as soon as we have a shared environment, the problem of someone willing to accept a reason or not dependent on his presupposed theoretical underpinnings is a pragmatic problem, not yet an epistemological problem.

Let us concentrate on the aspect of what makes a reason a good one, quite apart form the pragmatics of understanding, that is, we replace particular real-world agents with ideal epistemic agents. So, what makes reasons good ones or bad ones, ones that hold up to scrutiny and others that don't?

A reason is only good one if it relates in the right way to the environment, that is, it captures an environmental feature[41]. Good reasons track reality. For some this account may smack of heresy. But "relativizing" the power of reason to the environment is actually what again bridges the gap between the natural and the normative. Rationality – and the normativity it depends on – is not an ideal oozed out of an immaterial domain of *Geist*, something that exists outside of space and time, but something very much grounded in this world[42].

2.3.4 Logic and other Standards

What about traditional criteria for rationality, like logic or falsifiability? Deductive logic is only a constraint, it says nothing about which premises to adopt[43]. Falsification is also only a constraint, it serves to weed out false theories, but underdetermination of theories by evidence sometimes leaves too many possibilities remaining. Indeed, the Quine-Duhem thesis (Gillies 1998) actually sheds some dubious light on the principle of falsification; because falsification is also dependent on theories and concepts, such that a straightforward application of this principle is not possible. Applied rigidly, it can again all too quickly become dogma.

41 Humans are also part of the environment in the end; environment should be understood in a very broad sense here.

42 To be even more precise: thoughts (and reasons) have spatio-temporal locations; in Cartesian terminology, all res cogitans are res extensa – and have material causal histories (more on this in chapter three and four).

43 Put succinctly: One person's m*odus ponens* is another person's m*odus tollens*.

In epistemology one differentiates between defeasible and indefeasible reasons and reasoning. Insofar as one holds logic as a purely formal calculus, a matter of following rules, it is indefeasible: if one follows all rules correctly, then defeasibility is not a criterion[44]. Vexing questions crop up as soon as one takes into account what is applicable to real world situations. Is a domain properly captured by propositional logic? Or first-order logic, or rather some deviant logic? Here, defeasible principles come into play immediately.

The most stable and fundamental principles of logic – I will not argue which these should be, because all are contested somewhere or other in the philosophical literature – are those that seem to apply well in our everyday lives, such as the law of the excluded middle. These basic inferential skills are a product of evolution. Is this not reintroducing the psychologism Frege cast out of the halls of pristine logical thought? Logic is often seen as the highest principle of rationality, and something that must be the same in all possible worlds. It has always struck me strange as how a limited human mind, constrained by evolution in this world, can say something about all possible worlds[45].

Van Lambalgen advocates an evolutionary approach:

> An exaptationist account of the origin of logical reasoning might then run as follows. Planning is a capability shared by humans and nonhuman primates, even monkeys. If the above picture of the operation of working memory is correct, it requires the animal to represent goals and actions as nodes in declarative memory, and causal influences as links between those nodes. Humans have language in which to formulate goals and actions, but language also accesses the representations of goals and actions in declarative memory. Therefore one could suppose that the process subserving planning in animals also allows humans to draw quick conclusions, and to modify these conclusions if the need arises. Since the process is automatic, it need not be accessible to consciousness. That is, if for the moment we abstract from what we know about humans, it might have been the case that logical inference is more like a reflex, a form of low-level processing. (Lambalgen 2003)

Logic arises out of causal physical processes in the end: our bodies and our minds have been shaped by evolution, tracking reality for millenia. Our reasoning develops from experience in the

44 But it becomes a criterion as soon as a proof is checked for correctness in mathematical peer-review.
45 If we take the mathematical turn, this problem will become more pressing. But as for now, the statement will suffice.

physical world. Nowhere is a neuron untouched by prior physical processes. Minds reason correctly when they are in tune with reality. Logic represents inferential processes that capture basic modes of our linguistic perception of reality at the mesoscale; if we were quantum creatures we would probably deploy quantum logic, not classical logic.

Also, following Gigerenzer, I do not believe that we can bring helpful general principles to bear on all problems:

> Norms need to be *constructed* for a specific situation, not *imposed* upon it in a content-blind way. The reason is that content-blind norms disregard relevant structural properties of the given situation, including polysemy, reference classes, and sampling. I also show that content-blind norms can, unwittingly, lead to double standards: the norm in one problem is the fallacy in the next. The alternative to content-blind norms is not *no* norms, but rather carefully designed norms.
>
> […] Content-blind norms are of legitimate use within a formal system, such as for defining subjective probabilities in terms of certain rules like additivity. However, when we go beyond a formal system and want to find the best judgment or choice concerning a situation in the real world of human affairs, we have to construct norms for this situation, taking its characteristic structure and goals into account. In the real world, including the small world of textbook problems, the normative response depends on what we know or assume about the situation. The use of content-blind norms, in contrast, assumes that one does not have to take the situation into account. l

This is not to undervalue logic, probability theory, decision theory etc. But these are only *general* standards to apply to a problem; if one has a concrete problem, the rationalist can't refrain from taking into account the *specific* structure of the problem and *defeasibly* decide which standards he will bring to bear – every premature abstraction brings with it the danger of leaving out something important.

2.3.5 Coherence

Coherence is also an important criterion for the set of beliefs one entertains – incoherent beliefs may be entertained as conflicting models, but not simultaneously as a unified way of thinking about

the world. Coherence is more than consistency, but less than strict logical deducibility – it is a harmony of beliefs.

I follow Audi (2001) who holds that we need to distinguish *incoherence* from *coherence* as principles guiding our beliefs. Incoherence is needed to defeat the justification of (some) beliefs, whereas we need not strive for coherence as an active principle of justification. This is just as well, because coherence is notoriously difficult to pin down. Nevertheless: coherence, while not being necessary for justification, should be a guiding principle in the adoption of beliefs, in the sense that it is a measure – albeit a vague one – for the degree of unification a world view exhibits.

A concise definition of coherence – more precisely, a "coherence theory of inference" – is given by (Thagard 2002):

> 1. All inference is coherence-based. So-called rules of inference such as modus ponens do not by themselves license inferences, because their conclusions may contradict other accepted information. The only rule of inference is: Accept a conclusion if its acceptance maximizes coherence.
>
> 2. Coherence is a matter of constraint satisfaction, and can be computed by connectionist and other algorithms.
>
> 3. There are six kinds of coherence: analogical, conceptual, explanatory, deductive, perceptual, and deliberative.
>
> 4. Coherence is not just a matter of accepting or rejecting a conclusion, but can also involve attaching a positive or negative emotional assessment to a proposition, object, concept, or other representation.

2.4 The Body

2.4.1 On Emotions

Emotions result from sub-conscious evaluations of the environment, that is, they reflect prior experience; and so it is rational to take emotional states into account when evaluating situations consciously. Emotions are crucial both for the interaction with the environment, with other agents, and in self-analysis:

> … emotions play an instrumental role in relational activities. In other words, they are crucial for the successful interaction of autonomous individuals with their (subjective, individual) environment (which usually also comprises themselves as well as further similar individuals). Affective channels of communication provide continuously updated situational and contextual information of an individual's or a group's state, including clues about current tendencies to act (e.g., whether intending to approach or retreat, whether willing to pay attention, whether willing to give in or oppose). By nature, humans are highly attuned to the (in part unconditional) pick up of this kind of (partly unconditionally published) information, which forms an integral part of the coordination of social activities. (Petta & Trappl 2001)

There are a number of cognitive and neurophysiological models of why and how emotions work. A prominent neurophysiological theory is Damasio's somatic marker hypothesis (SMH) (Damasio 1994; Bechara, Damasio & Damasio 2000):

> Somatic markers are taken to function as a biasing device in that they substantially prune the search space of possible courses of actions by eliminating paths of likely unfavourable outcome. Jointly with subsequent deliberate rational decision making, somatic markers thus are likely to increase the accuracy and efficiency of the decision process, while their absence also has an explicit, negative, effect. An important function of somatic markers is seen in their support in overcoming short term horizon effects, as in the choice of actions whose immediate consequences are negative, but which generate positive future outcomes … (Petta & Trappl 2001)

Translated into a cognitive model, this assigns emotions to evaluative agent states selecting between more low-level *fixed action patterns*[46] (FAPs). Emotions would correspond to groups of FAPs and thus function as a pre-selector on FAPs appropriate for a certain situation (Petta & Trappl 2001).

While neurophysiological theories, such as the SMH stress the role of the body, cognitive theories jump a level up and look at the *cognitive value* of emotion. We need not go into detail here. It suffices for our purposes that

46 Our *scripts*.

> [a]cross virtually every current theory of emotion there is a consensus regarding the existence of fundamental, primitive and "pre-designed" mechanisms that already bring about the flexible reactive connectivity between internal or external events with low stimulus specificity and dynamically varying personal goals that is distinguishing of emotions. This characteristic complements and exceeds the capabilities of either reflexes (highly specific stimulus and low response flexibility) and physiological drives (specific stimulus and moderate response flexibility).(Petta & Trappl 2001)

Emotions are a further step in the bootstrap from simple organisms to organisms of higher sophistication. The hardwiring is provided, of course, by evolution, and connects higher level reasoning skills with lower level stimuli and their biased evaluation.

The decoupling of reason and emotion, in the extreme form of the separation of body and mind, was at the root of Descartes' philosophy and his greatest error. Damasio is explicit:

> This is Descartes' error: the abyssal separation between body and mind, between the sizeable, dimensioned, mechanically operated, infinitely divisible body stuff, on the one hand, and the unsizeable, undimensioned, un-pushpullable, nondivisible mind stuff; the suggestion that reasoning, and moral judgement, and the suffering that comes from physical pain or emotional upheaval might exist separately from the body. Specifically: the separation of the most refined operations of the mind from the structure and operation of a biological organism. (Damasio 1994, p. 249f)

Descartes, given his knowledge, is excused; today's scientists and philosophers are not; and while the mistake is not made by experts in the field – at least not consciously – the dualist intuition still underlies most of our thinking everywhere else (Papineau Forthcoming).

What about emotion in science? Emotion is relevant in two regards: firstly, the working scientist is a human being and has emotions; and secondly, emotions factor into the acceptance or rejection of scientific theories. The latter point is addressed in the whole thesis: I want to help reduce rejection of naturalistic theories due to averse emotional response.

So here I will address the former point, the working scientist. The picture of an emotionless, value-neutral machine is obviously false: the adoption of goals, such as epistemic virtue, or simply seeking coherent and empirically adequate theories, are already value decisions. Apart from that, it

would be ludicrous to try to eliminate emotion from our thinking processes, when, as we saw above, thinking is intrinsically emotional:

> ... we cannot insist that a person's thinking should be emotion-free when it is biologically impossible for people to think that way. (Thagard 2002)

Thagard analysed Watson's book "The Double Helix" (Watson 1969) for "emotion words", finding the following:

> Most of Watson's emotion words (163) occurred in the context of investigation. 15 words occurred in the context of discovery, 29 in the context of justification, and 28 emotion words occurred in other more personal contexts that had nothing to do with the development of scientific ideas.

It is important to note that Thagard does not follow Reichenbach's distinction of the contexts of discovery and justification distinguishing the "subjective" from the "rational" (a view completely orthogonal to the one proposed here anyway), but rather as different *stages of scientific inquiry* meriting their own label. Emotion and passion are common among successful scientists (McAllister 1996; Wolpert & Richards 1997). Without passion about a subject and the perseverance in the face of necessary setbacks in the course of scientific inquiry, no progress can be made – the emotional, the passionate scientist, is the one who enjoys success in the end.

Having said all the above, from where then comes the image of emotions being irrational? That is because, as we saw, they are low-level responses and thus very *prone to bias*. When reflecting rationally we need to be keenly aware of this; and while we can't be rational without emotion, we should always check if our emotions are appropriate in the current context, drawing on symbolic knowledge which has not yet been embodied and is thus not emotionally available. In an ideal rational agent who has propagated his knowledge through all associative nets, emotional response and rational verdict are in harmony.

I would like to draw up the following matrix:

	Emotional	Unemotional
Rational	OK	OK
Irrational	Not OK	Not OK

Irrational states are always bad, no matter if emotional or unemotional. Rational states can be held with emotion or without emotion – whereas the goal for a fully integrated individual is to move to the upper left corner of the matrix, emotional rationality[47].

2.4.2 On Intuitions

Damasio, in his book, sees intuition in close contact with emotions:

> As for the knowledge used in reasoning, it too could be fairly explicit or partially hidden, as when we intuit a solution. In other words, emotion had a role to play in intuition, the sort of rapid cognitive process in which we come to a particular conclusion without being aware of all the intermediate logical steps. … Intuition is simply rapid cognition with the required knowledge partially swept under the carpet, all courtesy of emotion and much past practice. (Damasio 1994, p. xix)

How much emotion is prevalent in an intuition can certainly vary – a certain amount is probably necessary, because an intuitive feeling that something is correct corresponds to past successful knowledge applications and the ensuing emotional reward, which is remembered. But there are certainly intuitions where very little emotion is involved apart from this reward aspect, and the past *experience* per se is at the fore. I think for instance of intuition in the mathematical domain, where one intuits solutions without immediately seeing all the deductive steps.

What I want to make clear in this section is that intuition is neither a spiritual force nor some kind of mystical access to a realm of direct and immediate true insight, even if it may seem so to the mind having the intuition. If someone has good intuition in a domain, this is because the person has either had much practice or (in single-case events) good luck.

In any case, it is rational to rely on intuition if one has to decide under time constraint, because it represents the sum weight of knowledge contained in an agent system – the sum weight of all neural connections, the synaptic chemical structure, the hormonal disposition etc. Your intuition is your life experience, everything you have ever heard, seen, thought of, smelled, touched and everything that has changed at least one molecular trajectory in your brain. So, you should use your intuition because that means harnessing more knowledge than if you only rely on those parts which have manifested themselves so strongly that they have trickled up into symbol space[48]. Our

47 For further discussion see Greenspan (2004).

48 I call symbol space that part of knowledge which has crystallized sufficiently that it is explicitly available for

intuitions are our background hypotheses in the Bayesian sense (see 2.6.3); we always have a probabilistic model of a situation, albeit one that is largely not verbalized/symbolized.

But this means that one should also beware of intuition: as it reflects past experience and not access to direct truth, it could be spurious. So it is imperative to always check if your intuitions serve you well or if they lead you to wrong answers[49], especially if one does not currently operate under a time constraint.

Only having intuitive knowledge on a subject clearly calls for more: trying to make the knowledge explicit so that it is available for reasoned criticism, versus being encoded in an indiscriminate neural blob, and engaging with the results of science to check if the intuitions are correct. Science guarantees best model building in the end. Having good intuition is no excuse for avoiding exposure to scientific evidence.

If one does not even have an intuition on a subject, that means that one has no experience in the domain: one should refrain from decision if possible and instead try to acquire knowledge, both theoretical and practical.

An excellent computational-algorithmic model of intuition is given by Yudkowsky (2008b).

2.4.3 Meaning, Understanding and Explanation

> I hear and I forget. I see and I remember. I do and I understand.
> Kǒng Fūzǐ (Confucius)

For cognition using representations, there is a problem of how *meaning* arises in such a system, how beliefs acquire their content. Symbols always need someone who interprets them, and it is the *interpreter* who supplies meaning[50]. But in a mind, there is no interpreter beyond the mind itself, no homunculus waiting to do the mind's work; for us to understand *how the mind understands* we need an account of the involved mechanisms.

First the backdrop which I will presuppose: all of our cognition is embodied[51], that is, it depends on our morphological/functional physical makeup, and its situatedness in an environment. It is then possible for meaning to arise in a representational system through its coupling to the environment

reasoning and reflection.

49 That is the difference between a well trained and a badly trained neural net.

50 Maybe one which has been agreed upon beforehand, in case of communication.

51 An excellent introduction to the concepts of embodiment and situatedness can be found in (Anderson 2003).

with its (also represented) body[52]. The key to understanding – which may occur through very different high-level cognitive processes, such as "empathy, metaphors, analogies or hypothesis" (Mahner & Bunge 1997) – is embodiment; everything that is encountered can ultimately be decoded into internal neural structures which are grounded in the body[53].

Second a terminological issue: when looking at theories that give an account of how beliefs can acquire content, we can differentiate between *output-based theories* and *input-based theories*. Output-based theories analyze the content of beliefs in terms of the actions they prompt, whereas input-based theories focus on the conditions which gave rise to the beliefs (Papineau 2003, p. 26).

Teleosemantics (Papineau 1984; Macdonald & Papineau 2006) is one such output-based theory, which I endorse with one caveat: I think that the truth of how meaning arises in representational systems lies somewhere in the middle – dependent on *both* input and output.

Inputs and outputs give positive feedback to certain representations and negative feedback to other representations. There is a causal linkage between physical events strictly outside the cognitive system, coupling to perceptual neural signals, and finally brain activity (input-side) which then leads to neuromuscular activations and actions in the world (output-side). Only if signals, brain internal feedback loops and actions form into a coherent whole does meaning – for this concrete entity – arise. That was quick. Again, the goal here is not to give an exhaustive treatment of the concerned subject, but only a hint that a naturalistic solution is possible[54].

Meaning is closely coupled to understanding. When we ask: what does this mean, we implicitly ask: I want to understand this. But there are different variants of understanding.

First, there is a kind of empathy, a relation to another human being: we know how it must feel like to be in their situation; we can relate to their emotions and motivations etc. That is not what I will speak about, although it is easily integrated into this account.

Another aspect of understanding is what I would like to call *internal* understanding, that is, a new fact integrates into previous knowledge. This kind of understanding is nothing other than integrating new neural input into antecedent neural structures[55]. Understanding derives from

52 The representation is also done in the body, mainly in the brain.
53 This also holds for abstract domains such as mathematics (Lakoff & Nunez 2000).
54 An excellent review article illustrating the grounding of words from an AI perspective is Roy (2005).
55 In a similar vein, sudden insight happens like this: insight is a spontaneous meshing of previously unrelated concepts in a neural system, leading to new embodied knowledge of a feature of the world. It is thus not surprising that scientists often report of ideas coming to them during sleep or in the morning after a dream-rich night: the neural system is highly active during sleep and performs a multitude of tasks during the different phases of sleep.

meaning present in the system beforehand, even if it may be combined into novel configurations. I am not sure if it is a good idea to call this process "understanding". I will refer to it as the "feeling of understanding".

Because there is a problem: just because we *feel* that we have understood something does not mean that we *actually* do. We accept something when it "feels right"; when it is in harmony with our prior knowledge, with our psychological makeup. But this feeling can be deceptive, for instance when our cognitive map of the world is already so wrong that an additional fact is integrated into this (dis-)harmony seamlessly without contributing to increased knowledge of going-ons in the world. While input-based accounts of meaning can account for the *feeling of understanding*, output-based approaches are needed to ground the agent in reality, so that this feeling does not begin to deceive her.

Real understanding – I will call it *external* understanding – is attained when you can perform successful actions in the world derived from your new knowledge (it presupposes internal understanding, that is, the neural integration of the knowledge).

External understanding is present when you can build; when you can prove; when you can transfer your knowledge to other domains. This is the output-based side of meaning. If no actions are performed to check if understanding has occurred, the agent may entertain the false belief that he understands while in truth he does not. The input-side accounts for our feeling that we have understood something, and the output-side guarantees that this feeling is correct[56].

Understanding (sometimes) requires explanation; the account of explanation endorsed here is the one given by Mahner & Bunge (1997); Bunge (2006); explanation, on their account, should be

> deductions from statements referring to regularities and circumstances, in particular law statements and data … the given fact to be accounted for is shown to be a particular case of such pattern [sic!] … in the case of explanation it [the pattern] is a mechanismic hypothesis or theory. (Mahner & Bunge 1997, p. 107)

The explanatory pattern, to be precise, is the following:

56 This account of understanding can also be transferred to communication via symbols: language, writing, cultural artifacts etc. The recipient of a message is now not only confronted with a direct impression (say, a fellow human standing before her making noises or drawing lines on a paper), but also with a message underlying the symbolic content. The message can either be decoded, not decoded, or partially decoded. Understanding depends on the receiver relating to the symbols in the correct way, that is, in the way intended by the sender. Understanding will never be complete, because the two people communicating don't have the same neural structure encoding their concepts; but complete understanding is not required to get along with one another.

> For all x: if Px then Mx; For all x: if Mx, then Qx, Pb ∴ Qb
> (Mahner & Bunge 1997, p. 107)

where *Qb* is the given fact (the explanandum), *Pb* the obtaining circumstance, and the logical relations of *Px, Mx and Qx* are the actual explanatory pattern (explanans). The crucial step where the account of Mahner and Bunge differs from others is the presence of M: M symbolizes a mechanism, where the word "mechanism" is used in a wide sense:

> The only condition for a mechanism to be taken seriously in modern science is that it be material, lawful and scrutable (rather than immaterial, miraculous, and occult). (Mahner & Bunge 1997, p. 108)

The reason why *these* explanations are truly explanatory while others are not[57], are; I contend our cognitive grounding in our bodies. Mechanismic explanations merge well with our sensorimotor neural apparatus and lead to genuine, embodied understanding.

Difficulties in understanding occur in when we transcend the familiarity of our everyday experiential world; when we enter the domains of non-linear phenomena, of multi-hierarchical complex systems; of micro- and macro-scale physical effects dominating the evolution of the universe at the ultra-small and the ultra-large; these phenomena transcend our cognitive arch; we were selected for mesoscale, local interactions at slow relative speeds and short time spans; it is in this region where chemically interesting things – life – happen.

If we have difficulty understanding scientific theories which transcend the domain of our evolutionary heritage, it means we must work harder to translate this knowledge into our cognitive purview; or enlarge our cognitive purview. It does not mean that reality is strange or mysterious; nor that our human, all-too human limitations are worthy of celebration.

2.4.4 Memes

But back to the symbols: the introduction of symbols also opened up a new avenue of spreading information via communication nodes which *do not possess understanding*. Let us consider words: we often read and hear them without thinking what they actually mean – indeed, it is often pleasing just to use or hear the expected words in a given context, because that smooths social relationships. Words seem to acquire a life of their own, quite independent of their meaning. This leads people to

57 Such as the Deductive-nomological model (D-N model) of explanation (Hempel & Oppenheim 1948). Bunge calls the D-N model "explanation by subsumption". It is not a full explanation, because it says how something happens, but not why, as in the full mechanismic explanation. The mechanismic explanation subsumes the D-N model.

recite whole sentences without knowing what they actually mean (meaningless metaphysics[58] would be one case in point; religious chanting of obscure passages another). But even the scientifically minded person will, when in a more reflective mood, discover words that she uses of which she has no clear conception of what they actually mean.

This aspect has been recognized by a number of scientists and philosophers such as Dennett and Dawkins who employ the term "meme"[59]. Dawkins not only proposes the thesis that humans are mere survival machines for their genes, but also proposes that memes, ideas, replicate under selective pressure and "inhabit" and "use" our brains. A concise definition is given by Bryson:

> Memetics refers to the theory that knowledge and ideas can evolve more or less independently of their human-agent substrates. While humans provide the medium for this evolution, memetics holds that ideas can be developed without human comprehension or deliberate interference. Bryson (2008)

The concept of memes has met with criticism[60]; mostly because the concept of the "meme" and its (pre-)theoretical elucidation are too simple to account for the rich variety of human culture and ideas and their varied expression in individuals. More sophisticated approaches to the concept of cultural evolution can be found in Boyd & Richerson (2005); Richerson & Boyd (2005).

But this is a dispute which goes too deep for our purposes; for the rationalist, it suffices to know that culture evolves (variation, heredity, selective pressure) and that memetics is a possible model – one that may be too simple for cultural and anthropological studies – but not too simple to construct a model which illustrates how ideas can infect populations like viruses; memetic viruses which have been bereft of meaning and may be detrimental to the host population.

The critical mind can now use the meme metaphor to screen incoming data and discard purely memetic information with no grounding; or, if she believes that the memetic information once had a grounding which got lost in the transmission process, try to recover the actual semantics of the meme.

58 For instance, Heidegger's famous "Das nichts nichtet"is a symbol concatenation which can't be grounded anymore.
59 Derived from Greek "mimesis" for imitation. The term was introduced in the book "The Selfish Gene" by Dawkins (1989).
60 See for instance Sperber (2000).

2.5 Things to Watch

2.5.1 Cognitive Biases

We are not evolved for reasoning but for survival; reasoning is a means for survival, not more, not less – at least it was in the beginning. Human beings are not rational animals *just so*. Rationality is something to we have to actively strive for. We have to exert *effort* to be rational; we have to *learn* to become rational – we are full of heuristics and biases which were evolutionarily sensible, but which may give wrong answers in the modern world. It is therefore necessary that we now have a short look at cognitive biases[61].

Without bias, we couldn't make sense of the world[62]: there is simply too much incoming data. A list of biases can be found in Baron (2008, p. 56-57). Another list of cognitive biases can be found on wikipedia[63]. The list is long. This all means that we must be extra careful in exercising our thinking skills.

Some classic biases are:

- **Bandwagon effect**: people believe things for the sole reason that others do.
- **Illusion of control**: the belief that one has influence of events which are clearly beyond one's control.
- **Hindsight bias**: the belief that past events could have been predicted (in the past)
- **Selection bias**: Collecting non representative data
- **False consensus effect**: agreement is assumed where none is had.
- **Illusion of asymmetric insight**: people believe that they know more about others than others know about them.

Science is one such reflective process which seeks to eliminate bias; the scientific method[64] is successful, among other factors, *because* it corrects for biases, whereas other worldviews do not systematically correct for biases.

61 The seminal work on heuristics and biases is Kahneman, Slovic & Tversky (1982).

62 This gains some explosiveness when it does not concern individuals but whole cultures; Wimmer captures this with his insight that it is always necessary to adopt some form of centrism; but there are more benign and less benign forms (Wimmer 2007).

63 http://en.wikipedia.org/wiki/Cognitive_biases

64 The individual scientist may well be prone to biases in areas foreign to her specialization or in in her normal life. That is of no import, as long as she follows methods that guarantee that no biases will contaminate her scientific results.

As individuals we should also have an interest in eliminating bias – simply because that rectifies our view of the world. Good questions to ask that may help us avoid falling into certain cognitive traps are: where could *I* be wrong? What are *alternative models* for the data in front of me? *Why* do I believe the things I do? How would my belief look to me if *somebody else* held it in a similar situation? In a *different* situation? And so on.

Richard Brandt (1979) suggests that an agent has reasons to do something if it is in accord with desires that survive *cognitive psychotherapy*:

> (a) Putting aside any of the agent's desires that are founded on nonempirical beliefs (such as normative beliefs).
>
> (b) Subjecting the agent's remaining desires to full empirical information, which may expunge some of the agent's desires and elicit some new ones.
>
> (c) Making sure the agent's reasoning is logically correct.
>
> (Hooker & Streumer 2004)

This procedure would be apt to remove many biases coloring an agents beliefs and desires, and thus ensure a more rational outlook. Of course, in the real world, we will have to settle for less.

So, biases exist, but they do not endanger the project of a rational worldview. On the contrary – biases have been discovered via scientific and rational investigation, and only with rationality can we hope to overcome them. In a similar vein also Nozick:

> In recent years, rationality has been an object of particular criticism. The claim has been put forth that rationality is biased because it is a class-based or male or Western or whatever notion. Yet it is part of rationality to be intent on noticing biases, including its own, and controlling and correcting these. (Might the attempt to correct for biases self be a bias? But if that is a criticism, from what quarter does it come? Is there a view that holds that bias is bad but that correcting it is bad too? If it is held to be impossible to eliminate bias, then in what sense does charging bias constitute a criticism? (Nozick 1993)

2.5.2 The Power of Words

> Tsze-lu said, "The ruler of Wei has been waiting for you, in order with you to administer the government. What will you consider the first thing to be done?"
>
> The Master replied, "What is necessary is to rectify names."
>
> "So! indeed!" said Tsze-lu. "You are wide of the mark! Why must there be such rectification?"
>
> The Master said, "How uncultivated you are, Yu! A superior man, in regard to what he does not know, shows a cautious reserve.
>
> "If names be not correct, language is not in accordance with the truth of things. If language be not in accordance with the truth of things, affairs cannot be carried on to success.
>
> "When affairs cannot be carried on to success, proprieties and music do not flourish. When proprieties and music do not flourish, punishments will not be properly awarded. When punishments are not properly awarded, the people do not know how to move hand or foot.
>
> "Therefore a superior man considers it necessary that the names he uses may be spoken appropriately, and also that what he speaks may be carried out appropriately. What the superior man requires is just that in his words there may be nothing incorrect." (Kǒng Fūzǐ[65], Lun Yü, Book 13(3))[66]
>
> "When I use a word," Humpty Dumpty said in a rather a scornful tone, "it means just what I choose it to mean – neither more nor less." (Carroll 1871)
>
> Transparent wording is the essence of a frame that fosters insight. (Gigerenzer 2003)

So, Kǒng Fūzǐ advises us to use the right words, whereas Humpty Dumpty is more lax in his standards. What does it mean to use words correctly?

65 More commonly known as Confucius.
66 Translation from: http://ebooks.adelaide.edu.au/c/confucius/c748a/part13.html

We can interpret this in two ways – either in an epistemological/ontological sense, meaning that we should carve the world up in the right way – that is, conceptualize it in the "correct" way, label the concepts via words and then use them when applicable; more on that in the chapter on metaphysics. Or – and this will concern us here – in a psychological sense. Words are not inert symbols with static meaning.

Words have a life of their own, in the sense that every word we learn, we learn by association with events in our life. Communication is possible because we humans share our environments, our physiology, our ancestral evolutionary history. The more different someones experiences are from ours, the more difficult communication will become. When we speak in "normal" words everybody has different connotations and associations bundled with them.

Words both have an intensional and an extensional definition. A short but sufficient explication is given in by Yudkowsky:

> To give an "intensional definition" is to define a word or phrase in terms of other words, as a dictionary does. To give an "extensional definition" is to point to examples, as adults do when teaching children. The preceding sentence gives an intensional definition of "extensional definition", which makes it an extensional example of "intensional definition".
>
> Intensional definitions don't capture entire intensions; extensional definitions don't capture entire extensions. If I point to just one tiger and say the word "tiger", the communication may fail if they think I mean "dangerous animal" or "male tiger" or "yellow thing". Similarly, if I say "dangerous yellow-black striped animal", without pointing to anything, the listener may visualize giant hornets.
>
> You can't capture in words all the details of the cognitive concept - as it exists in your mind - that lets you recognize things as tigers or nontigers. It's too large. And you can't point to all the tigers you've ever seen, let alone everything you *would* call a tiger.
>
> The strongest definitions use a crossfire of intensional and extensional communication to nail down a concept. Even so, you only communicate *maps to* concepts, or instructions for building concepts - you don't communicate the *actual* categories as they exist in your mind or in the world. (Yudkowsky 2008c)

This passage illustrates the difficulty of conveying concepts, of communicating. Concepts are not something you can give to someone like a piece of gold. They must be learned the hard way, and more often than not, due to slightly different conceptual learning, misunderstanding will ensue. The different learning histories of concepts will also lead to different inferential chains being traversed by cognitive entities when contemplating on these things.

For instance, if I learn concept A in association with B, and from B follows C, whenever I hear A I will also begin thinking about C. Not so for someone who has learned concept A in association with concept E, where no connection with D is present. The underlying theory is that of the prototype model of category structure in cognitive linguistics (Croft & Cruse 2004, p. 77-87)

That is why, in mature scientific theories, we use mathematics. In mathematics we try to avoid ambiguity by accepting only relationships made explicit in formalisms, computations or structures. That is the strength of both mathematics and conceptual analysis (mathematics being more rigorous of course); one declares ones frames explicitly and, in the case of mathematics, follows it strictly or, in the case of conceptual analysis, hunts for contaminations by other frames.

It should be stressed that by the maxim of "using words correctly" I do not intend some totalitarian declaration of what is right and wrong. I have in mind the respect for the partner of communication: the cooperative principle and the four Gricean maxims of communication come to mind, such as relevance, quality, quantity and manner (Grice 1975).

Even more dramatic is the usage not only of single words but of whole contexts: framing. One presents a narrative leading to certain expectations in people which influences their judgment. A classic example of this kind originating in the heuristics and biases tradition is the "conjunction fallacy". The example is presented thus:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?
>
> Linda is active in the feminist movement. (F)
>
> Linda is a bank teller. (T)
>
> Linda is a bank teller and is active in the feminist movement. (T&F) (Kahneman, Slovic & Tversky 1982; Tversky & Kahneman 1983).

The respondents were asked to rank the statements according to the degree Linda resembled a typical representative of the respective class; the result was that 85% of the respondents ranked in the following way: F > T&F > T. So far so good. The problem was that the conjunction (T&F) was, in a different test, also ranked as more *probable* (versus only as more representative, as in the former test). The psychological reasoning working with representativeness overrode the extensional logic of the probability calculus: the probability that P(A) >= P(A&B) <= P(B), for all A and B; adding additional features to a model can only make the model less likely, not more likely. The surprising result was that even *statistically trained* respondents did not outperform naive respondents; the fallacy was pervasive in all participants of the study.

Statistical fallacies can often be corrected when framed differently, for instance with natural frequencies or explicating reference classes (Gigerenzer 2003). These framings can have serious life or death consequences when patient decisions are influenced. An example of how framing can convey very different connotations:

> ...there are positive frames ("you have an 80% chance of surviving surgery") versus negative frames ("you have a 20% chance of dying from surgery"). My hypothesis is that they have an effect if patients can reasonably assume that the physician's choice of frame conveys additional information, such as dynamic information. For instance, the positive frame can imply that surgery will increase the survival chance from 0% to 80%, whereas the negative frame suggests that surgery increases the chance of dying from 0% to 20%. (Gigerenzer 2003)

Everything said is said by someone, and everything heard is heard by someone. There is a lot of leeway for interpretation, guesswork and misunderstanding in that simple chain. That is why we must be careful with words and we must be careful with frames, what they suggest, what they highlight, what they will trigger in the brains of the communicative partner. To choose the correct words and the correct frames is an important step in the way of the rationalist; not only in relation to others, but also in one's own internal dialogue.

2.5.3 Authority

> The fifth-century Chinese philosopher Xiaoguang Li observed that ancient civilizations are revered, and yet ancient civilizations are not wise like venerable human elders are wise. A civilization further

> back in time is younger, not older. The current civilization is always the senior, because the present enjoys a longer history than the past. Incidentally, does it change your opinion if I tell you that Xiaoguang "Mike" Li is actually a friend of mine who lives in the Bay Area? (Yudkowsky 2007c)

In human society – at least up until recently – elders have been revered as a source of wisdom and knowledge. It is natural, when thinking of your elders as wise ones, to transfer this attribute to their forefathers and so on, leading to the apparent wisdom of the peoples of old (one is reminded of the Yellow Emperor in Chinese mythology). But this is not so. With death, the wisdom of a person is lost save for those pieces of knowledge transmitted to progeny. The wise live on only in their pupils. But knowledge is accumulated from generation to generation; experience is aggregated, theories are formed, regularities observed: gradually, knowledge increases.

That is why the farther back in time we look, the more we should regard our ancestors as naive and unknowing children. We should not call the past "times of old", suggesting wisdom and knowledge; we should speak of the past as of the times when humans where still young. We are the grown ups – now – and we will be the children for the future generations looking back at us. But the moral is clear: we should have trust in today's knowledge. It is the best there is. There has never, ever, been more available, and it is increasing every day.

If Aristotle and Descartes would live now, we would not gather around *them* to listen, except maybe to listen to their tales of yore. No – they would go to university and start to learn: they would absorb the knowledge that undergraduates or high schools students take for granted today and exclaim in wonder and rejoice.

But what about the authority of the living? In science, authority is earned by publishing papers on one's original research work, the more prolific and exceptional the better; or by holding tenure at an esteemed institution, attending conferences and being invited as keynote speaker, by being a member in many scientific organizations and thereby exerting influence etc. Now, this authority is quite real in the sense that an expert on a topic will usually have a more informed opinion than someone new to the area. But one can never exclude that a novel approach captures some aspect in a better way than that propounded by the academic tradition. To this end, the rationalist must always bear in mind that all arguments must be weighed equally; adhering to academic mainstream is no excuse to ignore other positions.

Authority also acts as a filter against information overflow. We can't read all the articles being published in our field – the name of the author often functions as a guarantee of quality. On the other hand, one could miss an important idea this way, discarding an outlandish sounding but maybe correct idea just because it is not backed by authority. That is one reason why science progresses slower than it could. And it is also true that scientists in the field may not be ideal epistemic agents – that evidence contradicting a research tradition to which a whole life was dedicated is denigrated instead of embraced. But that is only to be expected where humans are at work. It does not call into question the truth-directedness of the scientific enterprise as seen from a global point of view; it just means that epistemic values have to be fought for every step of the way. And that is what young scientists are ideally trained to do.

2.5.4 A Cult of Rationality

Maybe the most intriguing turn of events in the history of rational thought is Ayn Rand's *Objectivism*, showing that even rationalist programs can degenerate. Ayn Rand was a novelist who outlined her philosophy of objectivism in the book "Atlas Shrugged" (Rand 1957). Shermer sums the philosophy up:

> Ringing throughout Rand's works is the philosophy of individualism, personal responsibility, the power of reason, and the importance of morality. One should think for one's self and never allow an authority to dictate truth, especially the authority of government, religion, and other such groups. Success, happiness, and unrestrained upward mobility will accrue to those who use reason to act in the highest moral fashion, and who never demand favors or handouts. (Shermer 1993)

But that, alas, was not the end of it. The movement progressively degraded into a cult; Shermer quotes Branden, an ardent follower of Ayn Rand, who described some of the beliefs the followers of Rand came to hold

- Ayn Rand is the greatest human being who has ever lived.
- Atlas Shrugged is the greatest human achievement in the history of the world.
- Ayn Rand, by virtue of her philosophical genius, is the supreme arbiter in any issue pertaining to what is rational, moral, or appropriate to man's life on earth.

(Branden 1989, p. 255f):

From a rational point of view, it is quite unclear how anyone could still follow this "philosophy of reason"; needless to say, many people abandoned the cause. But to guard against such failings in the future, it is more interesting to look at the fundamental flaw of Objectivism that could lead to such pronouncements.

The problem of Objectivism was that the philosophy proposed that through reason absolute, final, unassailable truths could be found. It did not have the corrective of fallibilism and criticism built into its basic structure.

> It is a lesson in what happens when the truth becomes more important than the search for truth, when final results of inquiry become more important than the process of inquiry, and especially when reason leads to an absolute certainty about one's beliefs such that those who are not for the group are against it. Shermer (1993)

Shermer characterizes a cult in this way:

> In this context, then, a cult may be characterized by:
>
> *Veneration of the Leader*: Excessive glorification to the point of virtual sainthood or divinity.
>
> *Inerrancy of the Leader*: Belief that he or she cannot be wrong.
>
> *Omniscience of the Leader*: Acceptance of beliefs and pronouncements on virtually all subjects, from the philosophical to the trivial.
>
> *Persuasive Techniques*: Methods used to recruit new followers and reinforce current beliefs.
>
> *Hidden Agendas*: Potential recruits and the public are not given a full disclosure of the true nature of the group's beliefs and plans.
>
> *Deceit*: Recruits and followers are not told everything about the leader and the group's inner circle, particularly flaws or potentially embarrassing events or circumstances.

> *Financial and/or Sexual Exploitation*: Recruits and followers are persuaded to invest in the group, and the leader may develop sexual relations with one or more of the followers.
>
> *Absolute Truth*: Belief that the leader and/or group has a method of discovering final knowledge on any number of subjects.
>
> *Absolute Morality*: Belief that the leader and/or the group have developed a system of right and wrong thought and action applicable to members and nonmembers alike. Those who strictly follow the moral code may become and remain members, those who do not are dismissed or punished.
>
> (Shermer 1993)

Objectivism was a cult. It was identical to religion except in name. We should take it as a warning: there never can be a final doctrine for reason and enlightenment. It is reasoning itself which must be taught to people, reasoning as a process. Everybody must work at improving himself or herself to attain this goal; it can't be delegated. Every individual is summoned on his and on her own.

If eternal doubt, the willingness to rethink your most fundamental results always anew, to assail your most cherished truths with new empirical evidence and new logical objections, is also *dogmatism,* then so be it. But it is unclear what more a thinking mind can offer; and unclear what should come in its stead. There are only heuristics for thinking critically, always evolving. That is not to say that there is no truth and objectivity, as some relativist thinkers would have it. Those invariances of the world that survive constant assault are our tentative truths. But we do not cherish them, we cherish the method.

An afterthought: Objectivism took itself too seriously; it could not laugh at itself; and in the same way, ideologies and religions can't laugh at themselves. Maybe our slogan should simply be: *With humor, but without dogma*. Where dogma prevails, thought is stifled; but where laughter is heard, the thoughtful mind is not far.

2.6 Updating

2.6.1 Learning

> Men are born ignorant, not stupid; they are made stupid by education. Attributed to Bertrand Russell

The most important feature of intelligent systems – maybe the defining feature – is the ability to learn. Shane Legg defines intelligence informally as:

> Intelligence measures an agent's ability to achieve goals in a wide range of environments. Legg (2008, p. 6)

To achieve goals, one has to acquire correct beliefs about the environment and correct means to achieve one's ends. Both are either innate, for instance through evolution in case of natural agents or they have to be learned by the individual.

Learning is the adaptation of internal representative structures of a sufficiently sophisticated agent to outside patterns. In this sense, learning has much to do with evolution: while evolution proceeds via *selection* on individuals and learning happens at the population level, learning in the normal sense of the word (that is, at the level of the individual) is *evolution of theories* in the individual agent[67]. Karl Popper is well known for articulating this position with the sentence: "let our theories die in our stead" (Popper 1972, p. 78).

In humans, learning proceeds via long term potentiation (LTP) of synapses, that is, synaptic change[68]. As a shortcut, I will henceforth use the term neural reweighting for "learning", to stress the structure-changing nature of the process.

This picture opens up another interesting way to view the relationship between power and learning. We have defined learning as the mirroring of outward processes in internal agent structure. Having power means that one does not need to adapt – need not change one's internal structure – instead, one changes the environment[69]. The danger, of course, is that this leads to inner rigidity; and that there will come a time and place when power has waned and the ability to learn too. Of course, we must not neglect the importance of power: if one learns and use this knowledge to

67 Vriend (2000) offers a nice introduction into the relevant literature as well as some criticism.
68 Nothing rests on a specific mechanism of learning: what is important is simply to realize that learning corresponds to structural change in the learning agent.
69 Karl Deutsch said that power is the ability not to have to learn anything (Deutsch 1963).

change the world, one must also acquire power. Power and learning, as all things, should be in balance.

2.6.2 Models and Kuhn

> There is nothing more practical than a good theory. Attributed to Kurt Lewin.

We get impressions from the world, and start thinking about what all that could possibly mean: we construct a *model*, a representation of reality which captures some aspects of the latter by similarity, partial isomorphism etc. Common-sense evaluations of worldly goings-on are models. Religions are models. Ideologies are models. Science is a *method* of arriving at good models. Good models let us predict occurrences in the world. That is, when we form beliefs about what we think is going to happen *and* these circumstances then come to pass, the confidence in our model grows. If our expectations are repeatedly frustrated, that is a good indication that our model of the world needs updating. Models are not only about prediction, but also about *understanding* the world – understanding is tightly connected to the visualization of *causal mechanisms* that operate to produce the phenomena in question. An excellent model should combine both prediction and understanding; if a model encompasses only one of the aspects, that is another indication that it is preliminary and will be supplanted by a better model in the future.

One of the most important tasks in good modeling is to make relevant variables *explicit*. In ordinary language so many assumptions are *implicit* that it is difficult to detect fundamental disagreements which are often masked by a shared vocabulary. A first step in a rational discussion is to make everything explicit. Only propositions or parts of models that are explicit can be recognized and criticized. Invisible shared or unshared assumptions will lead to false conclusions and failure of communication.

An adherent of the scientific method will acknowledge that every model fails to capture reality completely; and therefore, he will be careful in judgments on outcomes of complex physical processes. Outcomes are influenced by all of reality, not by the part we currently choose to model. Little physical differences can go a long way.

In the philosophy of science the word "model" also has a technical meaning, which comes to prominence in the *semantic view*[70] of scientific theories (Suppes 1960; French & Ladyman 1999). In

70 As opposed to the *received view* or syntactic view associated with the logical positivists, which views scientific theories as collections of linguistic propositions, with a partial empirical interpretation.

this view models are considered to be extra-linguistic entities, structures which satisfy certain linguistic descriptions but are not identical to them; also, they are not constituted by the syntactic aspects of a theory, but only described by them. Many structures can satisfy a certain linguistic description (axiomatization), entailing that a linguistic theory corresponds to a family of models. This is an important point, because misunderstanding this puts too much emphasis on linguistic issues, which, in science and the quest for finding good models of reality, are usually beside the point. The semantic view is an important move away from narrow conceptions of theories and a first move into relationalism, which will concern us in chapter 3.

How should we select between different models? Explanatory value and predictive success have already been mentioned. Coherence with models in other domains of science is another criterion. A nice formalization of some of the intuitions underlying model selection is Bayesian Theory.

But before we have a look at that, I want to deal with one obvious objection: that theories from different paradigms are incommensurable (Kuhn 1962). In the third edition of his book, Kuhn elaborates on this point in the postscript[71]. Kuhn details that he does not think that paradigm changes are irrational; only that in a paradigm change the scientist has to bridge linguistic and conceptual barriers; and that this requires certain techniques, but:

> the techniques required are not, however, either straightforward, or comfortable, or parts of the scientist's normal arsenal. Scientists rarely recognize them for quite what they are, and they seldom use them for longer than is required... (Kuhn 1962, p. 201f)

The techniques required to bridge paradigm clashes are those of translation, and:

> [f]or most people translation is a threatening process, and it is entirely foreign to normal science. (Kuhn 1962, p. 203)

These techniques of translation are *not* irrational; maybe they do indeed not yet belong to the tools of the ordinary scientist, but then it is high time that they should be made so; the ability to translate is the ability to switch ontological or theoretical frameworks, something of utmost import as we will see later on. Kuhn is somewhat negligent in his vocabulary, when he speaks of "persuasion" and "conversion" to a new paradigm, which suggest faith-based processes. On the other hand he acknowledges that paradigm changes are effected via reasons, possibly dependent on *values* the scientists have, such as simplicity and scope of theory.

71 Kuhn tries to distance himself in the postscript from the relativist reading of his book.

At this point it is enough to say that Kuhn's incommensurability only applies between linguistic and conceptual communities in regard to one another if they insist on *staying* withing their respective framework. If one has the courage to move beyond one's current framework, rational discussion of differences is again possible, because there is something we all share – our physical environment. That Kuhn's thesis of incommensurability was interpreted as making paradigm change irrational did not follow from Kuhn's thesis alone but rather from the logical-positivist environment which had equated scientific theorizing with linguistic constructions; then the switch between different linguistic constructions could indeed not be accounted for by a scientific process.

Incommensurability is not a problem for the rationalist, who is not interested in defending theoretical constructs, but in knowledge, and for knowledge the yardstick is reality and not some linguistic-conceptual entity. Everything is commensurable with reality. Commensurability is guaranteed by the world, which is the repository for referents; and the referents don't change when the theory changes[72].

Indeed, when we look at it this way, it seems that *commensurability* is at the forefront of paradigm changes, *not* incommensurability. Theory change is usually only accepted if the new theory explains more than the old one and more facts are unified in the same framework. We can say, contra Kuhn, that commensurability is the driving force behind paradigm changes: Maxwell unified electricity and magnetism; Newton sublunar and celestial mechanics; Einstein space and time. In all these cases, the paradigm change led to more phenomena of the world being commensurable than was possible before.

Given that I do not see any problem in principle with comparing theories – as to how they fare in prediction, accuracy, explanation and unification of empirical facts – it is time to honor the promise made above and look at a formal approach of how to perform this comparison – Bayesianism. The following section is a bit technical and may be skipped without much repercussions later on.

2.6.3 Bayesianism

So, what is the core of Bayesianism[73]? Bayesianism asserts the following:

- beliefs come in degrees

72 Ok, that was quick: a detailed discussion will follow in section 3.1.1.
73 For a technical overview, see Joyce (2004) or (Joyce 2008).

- all beliefs must be concurrently consistent with the laws of probability, as axiomatized by Kolmogorov (1933) and are represented by a probability function. This is a synchronic consistency criterion, violation of which leads to susceptibility to a Dutch Book[74].
- beliefs, held with prior knowledge at time instant t are updated[75] with *incoming new evidence* to form a new belief at time instant t+1 via the process of *conditionalization*[76], for instance *Jeffrey conditionalization.* If a person receives new evidence that sets her confidence in a proposition E to probability *q*, then she should change her *posterior* belief in the hypothesis *h* in the following way

$$P_{new}(h) = q \times P(h|E) + (1-q) \times P(h|\neg E).$$

where P is a probability function obeying Kolmogorov Axiomatization representing the agent's beliefs and $P(h|E)$ is the probability of the hypothesis relative to the proposition[77]. The notion of confirmation of a hypothesis by E can then simply be stated as:

$$E\ confirms\ h \Leftrightarrow P_{new}(h) > P_{old}(h)$$

An important constraint on "open-minded" probability functions is that they should be *regular,* which means that a probability of 1 is only assigned to logical tautologies and 0 only to outright contradictions; all other beliefs are never assigned a 1 or a 0, which would be tantamount to strong overconfidence which is never advocated in an uncertain world[78].

74 A Dutch book is a series of bets an agent is prepared to buy and sell at a "fair" price according to his beliefs in their probability, which will let the agent lose money because he violates a law of probability with his beliefs, say, additivity.

75 Belief updating is the process of learning.

76 Failure to condition properly makes one susceptible to a diachronic Dutch book.

77 This value, here with evidence as primitive instead of a proposition, can be calculated by Bayes' Theorem out of usually much less controversial *likelihoods*; as to the *priors*, see below. Bayes' Theorem (in one of its many guises):

$$P(h_i|e) = \frac{(P(e|h_i) \times P(h_i))}{\sum_{i=1}^{n} (P(e|h_i) \times P(h_i))}$$

where P is as as above, the h_i represent various hypotheses an agent entertains, where *i* is the index ranging over the number of theories, the $P(e|h_i)$ are the likelihoods of evidence occurring given the respective hypothesis (which should fall out of the theory in question), the $P(h_i)$ are the *prior* probabilities of the respective hypotheses, read: an agent's subjective confidence in these hypotheses being correct given her general background knowledge; and the $P(h_i|e)$ then are the posterior probabilities of the hypotheses given evidence *e*.

78 Two further possible constraints on one's credences are given by Lewis's Principal Principle, roughly saying that subjective degrees of belief should track objective chance (Lewis 1980) and Van Fraassen's Reflection Principle, saying that one should commit to one's future beliefs (Fraassen 1984). An elaboration would lead too far; the interested reader is kindly referred to the literature.

Bayesian learning, that is, belief revision and the adoption of a new posterior probability function, proceeds in two stages: a causal process impinging the epistemic system – for instance perception – which leads directly to altered beliefs; and a second, inferential step, propagating the impact of the new perception through the belief system.

What makes Bayesian epistemology attractive? Hajek & Hartmann (Forthcoming) list the following points in favor of Bayesian epistemology versus traditional epistemology[79]:

1) Bayesianism works with subjective degrees of belief and can therefore easily connect with decision theory.
2) Observations do not give definite evidence, but come with an associated degree of uncertainty. Jeffrey conditionalization takes this into account.
3) The criteria for knowledge in traditional epistemology are too strict, giving the skeptic too much foothold.
4) Degrees of belief seem more apt to represent complex mental states than the binary distinction of "belief" and "knowledge" in traditional epistemology.
5) Deductivism does not capture our ordinary reasoning processes which work with evidential support and not logical entailment.
6) Bayesianism has a vast formal apparatus at its disposal,
7) which is actively used in the social sciences, engineering, artificial intelligence, cosmology etc
8) There is a multitude of arguments, which – to employ the exact wording of Hayek and Hartmann – "collectively provide a kind of triangulation" to Bayesianism. The most prominent is certainly the Dutch book argument, which shows that an agent not conforming to probabilities in her beliefs will exhibit betting behavior that will guarantee her to lose.
9) Bayesianism is a simple theory with high explanatory value, for instance explaining how confirmation works.

Hajek and Hartmann (op. cit.) go on to list problems with Bayesianism, mostly as it fares in comparison with traditional epistemology, which is of little interest here. More severe are objections

79 Which stresses the arrival at justified true belief, and distinguishes in an all-or-nothing way between belief and knowledge.

concerning the *subjective* nature of degrees of belief. Does this mean that by adopting Bayesianism we are importing some kind of subjectivism or relativism into the halls of science?

Against the charge of subjectivity, one can only respond that *every* evaluation must be done by someone; it is only honest to factor this into epistemology. When a scientist holds a theory as more plausible as another, this of course depends on the sum of her knowledge and experience; the same holds for the scientific community as a whole, being constituted by human scientists. This is a fact about the world which is can't be eliminated by *avoiding* talk of prior beliefs and background knowledge in epistemology. On the contrary: avoiding explicit factorization of this circumstance would violate our desideratum of rationality to make everything as explicit as possible. Only the recognition that acceptance of theories is *necessarily* conditioned on the sum of one's prior knowledge and does result from some impersonal standard, a view from nowhere and nowhen, will make one more vigilant to possible errors in thinking. As Yudkowsky points out:

> Jaynes used to recommend that no one ever write out an unconditional probability: That you never, ever write simply P(A), but always write P(A|I), where I is your prior information. I'll use Q instead of I, for ease of reading, but Jaynes used I. Similarly, one would not write P(A|B) for the posterior probability of A given that we learn B, but rather P(A|B,Q), the probability of A given that we learn B and had background information Q. You can't unwind the Q. You can't ask "What is the unconditional probability of our background information being true, P(Q)?" To make that estimate, you would still need some kind of prior. No way to unwind back to an ideal ghost of perfect emptiness... […] But there's nothing wrong with that. It's not like you could judge using something other than yourself. It's not like you could have a probability assignment without any prior, a degree of uncertainty that isn't in any mind. Yudkowsky (2008d)

Is probability *itself* subjective[80]? That is, are probabilities always only degrees of belief, an epistemic state of an agent trying to predict what will happen next? Objective probabilities would be *propensities* of events, that is, of ontological significance: a propensity does not depend on agent beliefs. This is a difficult question, and I will refrain from tackling it here. All in all, even if probabilities were objective, that is, real propensities, our knowledge of them would have to be encoded in our brains and would thus be subjective, ideally tracking the objective propensities.

80 Subjective does not mean it is arbitrary – it only means that it is an epistemic aspect of the world – more on this in section 2.7.2.

The charge of subjectivism is especially glaring in standard Bayesianism, where the agent, while having to strictly conditionalize on incoming evidence, is perfectly free in her choice priors[81]. Convergence theorems are of not much help, because one can always chose sufficiently perverse priors such that convergence does not happen. As to the assignment of *concrete* numbers to degrees of belief, I think this would feign more knowledge than we actually have, giving a false sense of precision and exactness. What we can do is impose an ordering on the theories. In this we will have to operate with some concept of plausibility. As soon as we go outside the domain of mathematics and enter the physical world, so many uncertainties creep in that assigning an ordering is the best we can do; and usually we can classify some theories as very much more likely or less likely than others; that is often enough.

Does this make things look bleak? No – as we noticed at the outset, for every method of rationality we can find, if we look hard enough, counterexamples or weak points. *Actual* rationality – the process of thinking in the concrete instance – is not easily captured in a definite way. Where does that leave us now with Bayesianism? I would like to follow Horwich (2005) who suggests that we adopt Wittgenstein's conception of philosophy as therapeutic in this instance; that is, we take Bayesianism as therapeutic. According to Horwich, Bayesianism need neither address all criticisms leveled against it, nor deliver a true or complete theory of inference; it is enough if it illuminates some problems not tractable with traditional epistemology.

Another point speaking in favor of Bayesianism are developments in cognitive science, which suggest that low level brain operations, such as perception and sensorimotor control, are Bayesian in nature (Knill & Pouget 2004; Doya et al. 2007). If these results prove robust, it would constitute a nice symmetry that conscious reasoning processes are measured against the same standards as low-level unconscious brain operations.

To close the section with a word from Joyce:

> For all its shortcomings, Bayesianism remains without peer as a theory of epistemic reasons and reasoning. As long as we use it for this purpose it will serve us well. (Joyce 2004, p. 153)

81 Objective Bayesianism advocates restrictions on priors which have to be fulfilled before a belief may be deemed rational (Jaynes 2003).

2.7 Standard (Non-)Issues

2.7.1 Münchhausen Trilemma

The Münchhausen Trilemma is well described in Albert (1991, p. 13f). Classical epistemology asks why beliefs are justified. But this will lead, in the end, to the following trilemma:

- infinite regress: every proposition is in need of justification, and that justification is again in need of justification, and so on.
- logical circle: conclusions are used to justify premises.
- abortion of process of justification at some arbitrary point – say, where things become "evident" or "intuitive". This is actually a fallback into dogmatism, the principle of sufficient reason is suspended in its applicability.

The response of critical rationalism is to abandon the quest for ultimate justification and rather turn to the *principle of criticism*. There is never an instance of justification beyond criticism, everything is fallible. Of course, in criticism one also breaks off at some point, where the utility of further criticism does not contribute to solving any problem at hand. But there is nothing exempt from criticism a priori. Everything is regarded as fallible. The approach is pragmatic.

There is neither justification for rationality nor for the critical mindset. We simply observe the approach to be highly effective. We are also do not look for self-evidence or justification, but for grounding in reality. If we are aware of this grounding in principle, we are also free to question "evident" perceptions as possibly false.

2.7.2 Truth

> An error does not become truth by reason of multiplied propagation, nor does truth become error because nobody sees it. Truth stands, even if there be no public support. It is self sustained. (Gandhi 1927)

> You're entitled to your own opinions, but not your own facts. Anonymous

> The Ch'an Master went for a walk with Huang Shan-Ku, the Sung poet. When they walked past a mountain laurel in bloom, the Master asked,

"Do you smell it?" The poet answered, "Yes." The Master replied, "There! I have nothing to hide from you." Taoist Tale[82]

If one instrumental goal we adopt is the arrival at true beliefs[83] – we must at least look shortly at the concept of truth[84]. The question of truth arises at the moment where we have beliefs and representations. For a fact in the world, the question of truth does not arise – it simply is. We can encode beliefs, representations etc in propositions, and reduce the question to what makes these propositions true. The accounts given for truth are numerous and varied (Lynch 2001). The theoretical picture most naturally underlying any conception of truth is the correspondence theory of truth, that is:

$(t)\{t\ is\ true\ iff\ (\exists x)[(tRx)\text{and}(x\ obtains)]\}$ (Kirkham 1992, p. 132)

where t ranges over (unspecified) truth bearers, x is a fact or state of affairs[85], and R is a relation connecting the two, depending on the nature of t (say, beliefs, or sentences or whatever).

It is important to stress that the relation R is nothing mysterious. As we read in on in Kirkham:

> 'Correspondence' serves as nothing more than a handy summing up of a theory in which no such special relation makes any appearance. […] A. N. Prior has reached a similar conclusion. It is appropriate, he says, to use the word 'correspondence' in describing truth, although the word does not appear in his own definition of truth: "To say that X's belief that p is true is to say that X believes that p and (it is the case that) p. There seems no reason to see any more in 'correspondence with fact' than this" (Prior 1971, 21-22). [Prior, A.N. 1971. Objects of Thought. OUP.] (Kirkham 1992, p. 135)

Another criticism often leveled at correspondence theories is that "fact" (or states of affairs) is not well defined, or even if they were, while there were facts making beliefs true, one can also hold negative, conditional or disjunctive *beliefs*, and there are no *facts* of that kind. While the answer to the first question must wait for the metaphysical analysis, the second objection can be answered

82 Retrieved from http://davidlavery.net/Imaginative_Thinker/quotes/nquotes/noneed.htm; 27.08.2009

83 The goal of theoretical rationality as traditionally conceived is knowledge – "justified true belief". This account is both too strict (that is why we have adopted therapeutic Bayesianism above) and seems to be broken anyway (Gettier 1963; Floridi 2004). See the appendix for an alternative account by Bunge.

84 The wish to relativize the very concept of "truth" results from day-to-day power games and monopolizations of truth by certain groups. It is important to note that the adherence to a sensible conception of truth is not in any way intended to dogmatize certain propositions by ascribing truth to them. There is no danger because truth is always only a tentative ascription in scientific and rational discourse.

85 A state of affairs is a set of facts.

now: while the problem may pose itself if one wants to formalize the correspondence theory, it is not a problem for ordinary language usage, as here we can simply posit that truth in the relevant cases is derived from facts together with logical inferences. For instance, if I hold the belief that I am not seeing a pink elephant, the truth of this belief would derive from the sum of my perceptions and the additional assumption that these are exhaustive. This conglomerate would then indirectly correspond to the truth of the negative belief.

Truth is always evaluated by an agent. So truth is always "from a certain point of view", because it is in relation to the one doing the evaluation in accordance with his conceptual schemes. But that does not mean that truth is relative in a harmful way; because, if the truth bearer is indeed true, there will be *methods of translation*. – for instance, translations which respect structural invariances in the respective conceptual domains. If an object appears red to me, but green to you because you are under influence of a psychedelic drug, we can check how the drug affects visual perception and what mechanisms are responsible for changing color perception; the different truth values of color do not entail its arbitrariness; on a higher level, the truth values pop out the same.

That different agents will come to highly varied accounts of the world is not a challenge to truth: it is only a challenge to the scientifically minded problem-solver, who will immediately start the search for mechanisms entailing possible translations and unifications. There is often more than one good theory describing a phenomenon. Different theories simply capture different aspects of complex systems, and theoretical diversity more reflects our cognitive structure and capacity to think about things than the actual world in want of representation. There is no danger in recognizing the aptness of many *different* theories for describing reality. This will not lead to relativism, because there are always many more falsehoods.

The clearest account of truth is given by Bunge, where truth is defined via partial knowledge of epistemic agents tracking the world with cognitive processes. The truth of propositions is then parasitic on the underlying brain processes (see the Appendix for more).

2.7.3 Truth Maintenance Systems

It is time to move on to an important observation: humans are not "Truth Maintenance Systems". The concept of truth maintenance comes from AI systems that represent knowledge propositionally. Propositions stored in the system's knowledge base are called facts, and the systems

also exhibit inferential rules with which they can infer further facts from those presented to the system explicitly.

> Inevitably, some of these inferred facts will turn out to be wrong and will have to be retracted in the face of new information. This process is called belief revision. [...] Truth maintenance systems, or TMSs, are designed to handle exactly these kinds of complications. (Russell & Norvig 2003, p. 360)

But this task is not trivial – it is, in fact, at least NP-hard[86] (op. cit. p. 362).

We humans, in fact, nearly never make wide-scale belief revision in our knowledge base. When we are shown an inconsistency in our thinking, we only perform "local" updating, that is, revise those beliefs that come in immediate and actual conflict with the new belief. If we are in reflective mode, we may be inclined to think through some consequences of the new information and revise some further beliefs which have greater inferential distance. But generally, we don't sift through everything we have stored in our brains to weed out inconsistencies with the new information. This would not even be possible for us because knowledge recall in our brains works associatively[87]. Humans are an evolutionary bricolage – cobbled together to satisfice environmental constraints and optimize performance in naturally occurring situations. It is probably best to think of humans having "scripts"[88] for situations (see also above) which are executed on demand, and these scripts can be quite contradictory when scrutinized at a global level.

This failure of global updating is what Peter Strawson has in mind when he writes the following in regard to a difficult philosophical point with important ramifications[89]:

> It takes time to assimilate it fully. It cannot be simply read off the page. (Strawson 2006)

To know this – at a gut level – is very important for the aspiring rationalist[90]. Firstly, because it lifts into reflective state the meta-rule that only hearing about a logical or statistical relationship will not make the rationalist behave differently or take this new evidence into account everywhere where

86 Informally, problems of this class (or more difficult ones) have no efficiently computable exact solutions.
87 The illuminating Jets and Sharks model (McClelland & Rumelhart 1988) can be explored here: http://www.itee.uq.edu.au/~cogs2010/cmc/chapters/IAC/index4.html
88 What I call scripts here corresponds very closely to the concept of cached thoughts explicated in Yudkowsky (2007d).
89 The point will concern us in the metaphysical section concerning the structural nature of physical knowledge.
90 Important reading is also Gelder (2005) who illustrates the difficulty of teaching critical thinking which is closely related to the problem of global updating.

it *should* be taken into account. The really difficult task when hearing new knowledge is to embody it, that is, transfer it to all relevant scripts which are called up in appropriate situations. This requires practice, meditation, real-world and hands-on experience etc. Knowledge of the meta-rule can help achieve global rationality by prodding one to questions one's beliefs in every situation in light of the new evidence – although one has to make a preliminary[91] heuristic evaluation if this situation calls for reflecting on the new evidence in the first place, which is costly (the process applied indiscriminately would be taxing, time-consuming, and largely fruitless).

Secondly, this fact also explains how highly intelligent scientists can often hold very parochial views on religion or other domains foreign to their specialty. The reason for this is that parts of cognition which are affectively loaded in a high degree are not available for belief revision (Buggle 1992).

So, human beings, even those striving actively for global rationality, are, at most times in their lives (probably always) inconsistent in the sense that they hold *mutually incompatible beliefs*. Belief revision does not ensue the moment new beliefs enter the system, but only when inconsistencies can't be upheld any longer due to environmental pressure – in the case of the rationalist *intellectual pressure* upon discovery of the incompatibility. Some people do not even update in the case of environmental pressure[92].

All of the above illustrates the importance, in philosophy, to sometimes say trivial things. We have to say trivial things and then apply them in *other* contexts, where suddenly, they do not appear trivial at all. That is how we humans update our knowledge base. That is also why knowledge can not be simply read off of a paper. Take for instance the simple fact that you are made up of physical parts. That is quite a trivial thing to say – but what follows, alas, is painful to bear; at least on a first take. Many choose not to think this thought to the end. We will do so in this thesis.

Knowledge does not come easily – you have to think about it, meditate on it, act on it, embody it, in short: *become* it. Only then will understanding ensue. That is the difficult part. Concepts and ideas, as opposed to words, are not arbitrary. The distinction between memetic and embodied knowledge is relevant here (Bryson 2008). While the advent of pure memes (not grounded directly in experience but only semantically grounded) enriches cultural-evolutionary design space, it opens up all new kinds of possibilities for error; semantic grounding being one step removed from the

91 The preliminary evaluation can be wrong, but that is what fallibility is about.

92 Such as people continuing the practice of praying for personal boons despite prayer not showing any effects in clinical studies (Krucoff et al. 2005).

world, as opposed to embodied grounding. So a task of philosophy is to destroy scripts which are only semantically grounded but have no referents in the real world.

2.7.4 Objective and Subjective

An important distinction is that between the "objective" and the "subjective". Objective and subjective have nothing to do with truth. There are subjective truths and objective falsehoods[93].

Let us proceed in an orderly fashion. When I speak of ontological situations, I will speak of *facts*. Facts are the real thing, out in the world. Facts are what make propositions true.[94] Propositions are the statements with which sentient agents express their beliefs about the world. Only when we enter the domain of propositions and beliefs do question of objectivity and truth arise. Truth and falsity are then, as above, roughly construed as a correspondence of propositions with the facts, or, more accurately, a tracking of environmental patterns by brain states identical with the utterance of propositions.

The question of objectivity and subjectivity adds a new dimension to the analysis of propositions and beliefs. "Subjective" is every proposition concerning the internal side of cognitive states of agents, whereas "objective" is everything that is independent of such cognitive states. Note that objective is also different from intersubjective – there are belief systems, say, Christianity, which are shared intersubjectively but are not objective in the sense that the belief system is independent of human minds.

Bunge defines "objective" knowledge in this way (see also Appendix A: Terminology of Mario Bunge:

> Let p designate a piece of knowledge. Then p is *objective* if, and only if,
>
> (i) p is public (intersubjective) in some society, and
>
> (ii) p is testable either conceptually or empirically.
>
> Mahner & Bunge (1997, p. 67)

93 This is a nonstandard view propagated by Bunge, which I endorse. Nothing much rests on it though, as it is a simple quibble over words. The excellent analysis of Nozick (2001) of what constitutes objectivity can, for instance, be easily transferred to this system (it won't work for subjective truth though: which is why I follow Bunge and not Nozick).

94 Although they can't be defined as truth-makers of propositions, see our short discussion of negative, disjunctive and conditional propositions.

Subjective and objective are classifications of propositions; the ontological category is that of fact. There is nothing "less true" about the subjective than the objective, it is just that the subjective is not verifiable by *other* subjects. And verifiability is an epistemological concept.

Yudkowsky classifies a little differently:

> If you can change something by thinking differently, it's subjective; if you can't change it by anything you do strictly inside your head, it's objective. Yudkowsky (2008d)

It's a good rule of thumb, but does not discern between beliefs themselves and contents of beliefs. For instance, I can change my belief about whether the Earth is round or flat inside my head, clearly labeling my belief itself as subjective – but the content of my belief concerns something objective. So there are some subtle issues involved, explaining the confusion reigning in the subjective-objective issue.

Now, we *may* consider also separating facts into objective and subjective ones; I do not think that this is helpful. The distinction objective-subjective should not be applied to facts. Facts, as I wrote above, simply are. Thus, if I am currently happy, this is simply a fact. The proposition "I am happy" is a subjective claim though; and a true one if I really am happy.

There is something important to bear in mind: even a *subjective and false* belief, when held, is part of reality and can have causal effects. Science is the project of trying to attain true beliefs and avoid false beliefs; science is primarily concerned with *objective* propositions, as only those can be reliably communicated. But the rationalist must also concern himself with subjective propositions, as they refer to facts that are no less real.

A system of subjective and objective beliefs can be called a *perspective* on reality. What is a good perspective on reality? A perspective is good if it provides a consistent and enactable mapping of the world: that is, the inferential structure of the perspective tracks the causal structure of the world. The perspective will be good if it approximates truth, that is, it is in accordance with the facts.

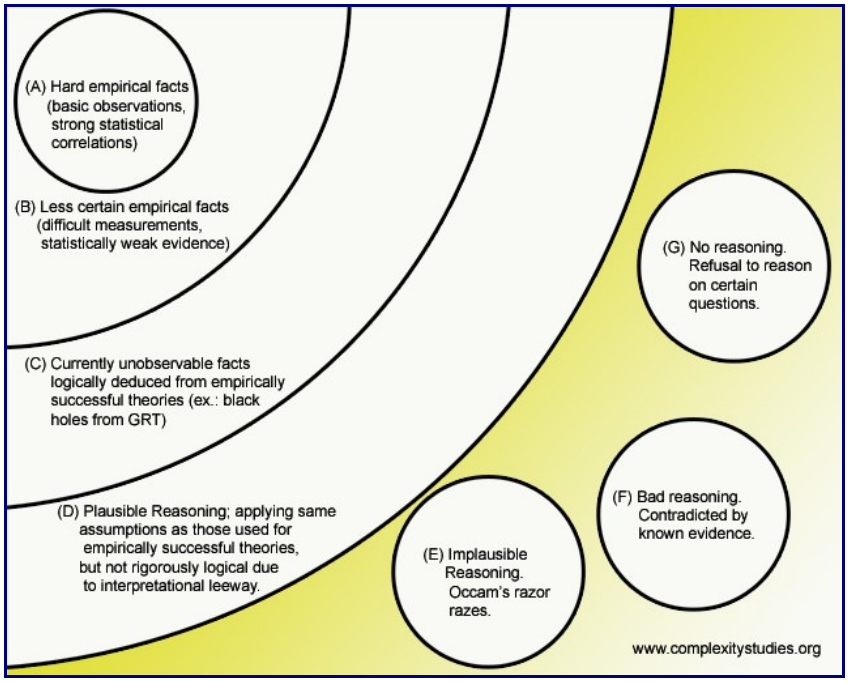
2.7.5 Science, Rationality and other Domains of Inquiry

Can we be rational in a stronger sense than the scientific *ideal* prescribes? I am not sure. But what we can do is try to do better than mere scientific *practice*, because that is nothing other than an evolutionary algorithm for ensuring that it works for the *population* of scientists. Describing the

scientific process in such a way could go like this: there is mutation (creativity of individual scientists), heredity (training in a scientific discipline) and selection (according to methodological rules and comparison with experiment).

Evolutionary algorithms, when operating on non-cognitive entities, explore blind alleys by necessity. But we humans are cognitive individuals with insight and planning abilities, so we can try to do *better* at a local level than the algorithm at the population level. The synthesis of knowledge of diverse disciplines in a single human may be conducive to weeding out theories that seem acceptable from *within* a discipline. A caveat: whenever we venture outside the domain of an established empirical discipline, extreme caution is urged. Paradigms and entrenched theories are usually there for good reasons, and, even if the rationalist may at times take the high perspective, she must always keep in mind that what may seem obvious from afar may be ruled out by some empirical detail unbeknownst to her.

And what we can do in any case is extend rational and scientific reasoning into other domains of inquiry. We can try to extrapolate from the empirical sciences and *structure* domains of inquiry which have up to present been the territory of unreflected, naive and historically grown discourse. The following diagram should clarify my point (it is not intended to be in any way canonical, only illustrative):



As we move outward from the inner core (A) things get less and less certain. But uncertainty does not exempt one from reasoning. Especially in the (D) section ideas for new experiments may arise; or even new conceptual structures for unification of theories.

(D) is the fecund field for discovering new knowledge and making sense of the world; it is the field of scientific philosophy. The problem is that most people who don't agree with the kind of reasoning performed in (D) don't argue from (A), (B), or (C); nor from (D). Criticism is raised coming from (E), (F) or (G) – the dominion of pre-scientific world-models fighting their last battle against reason. While science may not yet have enough empirical detail to analyze every aspect of the world, that does not mean that we are free to believe whatever we see fit in the white areas on the map; on the contrary, we are urged to exert even more care and restraint on belief in these areas than anywhere else.

2.7.6 Limits of Knowledge

That there are limits of scientific and rational knowledge does not mean that there are other methods to come by this knowledge. We have *defined* science and rationality such that these are the optimal ways of acquiring new knowledge. If new methods of discovery are found, they will be immediately integrated into the rationalist tool set.

Limits of knowledge are exactly that: limits[95]. You can't dream or intuit or make up answers[96]; humans are afraid of uncertainty, and that is why they prefer an imaginary answer over no answer at all. But an imaginary answer does not make the answer true or helpful beyond some initial psychological soothing.

And it should not go unappreciated that limiting results of science like the Halting Problem in computer science or no-go theorems in quantum mechanics also constitute a form of knowledge – we know what is *not possible*. Limiting results are knowledge about limits which *no one* can transcend.

Limiting results often have counterintuitive consequences. Take Gödel's incompleteness theorem for instance, which is often presented as a negative result. But there is a different reading:

> This positive implication of the incompleteness theorem is that we have a way of extending any system of axioms for mathematics that we recognize as correct to a logically stronger system of axioms that we

95 An interesting article exploring such limits is Svozil (2007).
96 What about revelation? Well – how to discern revelation from imagination? More on that in section 2.9.6.

> will also recognize as correct, namely by adding the statement that the old system is consistent as a new axiom. That the resulting system is logically stronger than the old system means that it proves everything that the old system proves, and more besides. Thus our mathematical knowledge would appear to be inexhaustible in the sense that it cannot be pinned down in any one formal axiomatic theory. (Franzen 2004)

So, mathematics definitely won't get boring – what starts out as a limiting theorem becomes a paragon of creativity.

An objection often raised when the rational exploration of the universe is proposed is that we *should* not know everything. Apart from the dubious origins of that "should"[97], we can answer in a very pragmatic way: exposing all questions to rational inquiry does not mean that *everything* will be solved at some future time and that life will become dull. The more we know, the more we can combine; knowledge is like a puzzle where each new piece enlarges the game. In the same way we may regard the open-ended nature of the universe as it presents itself to us in our current scientific inquiry – there always seem to be *more* problems; and when those have been solved, new ones will arise. Maybe people fear knowledge because they think it disenchants nature. But the true enchantment is that as knowledge grows, every new piece of knowledge can be combined with previous ones in an unending kaleidoscope of beauty[98].

A limit which is very baffling is the following: why *these* laws of nature, and not others: I think that at the moment this is still too difficult to answer[99]. There is no satisfying unified theory of quantum mechanics and general relativity yet; cosmological observations are getting better and better and will change our models of the universe in the future. We know very little. It should just be made clear that anybody who offers an explanation – such as religions do – doesn't have *more* knowledge than scientists, but usually *less*, so even less credence should be assigned to their proclaimed views.

---

97 The idea is that there is some magical border beyond which no inquiry is permitted; but who should delineate this border? It boils down to censorship and authoritarianism. The intention of course is to protect group biases (ethnic, religious etc) from science; we have seen this in action above in the section on "separate magisteria", a distinction which is, as argued, untenable.

98 And, as we will see in the metaphysical section, while reality is knowable, it is also inexhaustible – because one can "cut it up" in an infinite matter of ways and always discover knew interesting relations; without, to be clear, embracing relativism.

99 The anthropic principle aside, which should only be a last ditch explanation. Interesting ruminations on why something exists at all, instead of nothing, can be found in Tegmark (1996); Carlson & Olsson (2001); Witherall (2006).

2.8 Naturalism

> If you would learn to think like reality, then here is the Tao: Since the beginning not one unusual thing has ever happened. (Yudkowsky 2007e)

There are a number of definitions and variants of naturalism floating around (Sukopp & Vollmer 2007).

But at the core of naturalism, I think, are two ideas:

1) *The world is lawful.* There are no unlawful events. Maybe there are random events, but then appeal is made to stochastic lawfulness. Either way, no miracles are permitted. This of course raises the problematic issue of what a law of nature is. Criteria are given in (Vollmer 2003). I favor the mechanismic account of Bunge (see Appendix A).

2) *Humans do not occupy some special place in the order of things.* This is actually only a special mention of the fact that humans are part and parcel of the world, that there are no discontinuities involved; humans are self-aware matter configuration, nothing more, nothing less. It deserves special mention because humans are very prone to conceptualize themselves as in some way apart from the world.

Now, the commitment to naturalism can be adopted either a priori or arrived at a posteriori. A posteriori is of course the only serious method of arriving at naturalism. After careful personal observation of the world, diligent weighing of the scientific evidence and moderate knowledge of human psychology and its proneness to biases, a rationally minded person will sooner or later adopt naturalism[100]. Again, this is a kind of bootstrapping, but nothing objectionable lies therein from a methodological point of view; it is inevitable, and what is the use of objecting to the inevitable? This principle of naturalism is then a powerful heuristic for evaluating new incoming evidence. After one has "learned" naturalism from observation, one can use it to classify further observations.

Ants are natural. Their nests are natural. Humans are an integral part of the natural world. They build cities which invite analogies with ant nests. So why should we not call their cities, technologies and artifacts natural? The distinction natural/artificial is sometimes interesting – for instance, when we stumble upon an artifact on an alien world, this is astounding not because an artifact were something supernatural, but because an artifact suggests an artificer – an intelligence,

100That science is apt to test supernatural claims and rule them out is argued convincingly in Fishman (2009).

an agent, which made it – a natural being of some sophistication. But the distinction artificial/natural does not have *ontological* import. In the universe, everything is natural.

Since everything is natural, note that also the humanities form only a part of the inquiry into the natural world:

> The so-called moral sciences or Geisteswissenschaften (literally sciences of the spirit) are re-conceived. Anthropology, economics, political science, and sociology study the thinking and being of social animals, not collections of radically autonomous Cartesian agents, not of beings running on Geist — on spiritual fuel in the spooky sense. The unification of the sciences that study persons […] is made possible by the insight that all these sciences are all engaged in studying various aspects of the thinking and being of a certain very smart species of social mammal. (Flanagan 2007, p. 3)

The naturalistic study of music or literature, to give examples which are traditionally associated with the humanities, is well underway (Gottschall 2007; Ball 2008).

The thing with naturalism is this: the position is, once adopted, near obvious; all supernatural explanations would introduce additional, unneeded entities[101]. But naturalism is contested because there is significant emotional resistance to this position. Bloom and Weisberg have investigated childhood origins for adult resistance to scientific thought:

> ...both adults and children resist acquiring scientific information that clashes with common-sense intuitions about the physical and psychological domains. Additionally, when learning information from other people, both adults and children are sensitive to the trustworthiness of the source of that information. Resistance to science, then, is particularly exaggerated in societies where nonscientific ideologies have the advantages of being both grounded in common sense and transmitted by trustworthy sources. Bloom & Weisberg (2007)

The results apply equally to naturalism, which is simply consequent application of scientific thought to *all* domains of life. Rationality and naturalism challenge us at a very deep level as human

101If one does not adhere to some principle of simplicity in theory selection, one can, of course, believe everything; for instance, that every event, no matter how insignificant, is caused by a god manipulating matter.

beings and the way we view ourselves as persons in the world. Embarking on the journey of complete naturalization is not undertaken lightly, and even highly intelligent people may recoil[102].

2.9 Opposition to Rationality

2.9.1 Variants of Opposition

Reason – so difficult has it come to this universe and to humanity, so great have been its benefits – and yet … the moment one pronounces the attitude that all domains of inquiry are and should be open to rational discussion and scientific investigation, vehement opposition is not long in the waiting.

The reasons for opposition are varied, and sometimes one is not sure if there are reasons at all. But one can classify some kinds of opposition:

1. image of the self
2. misconceptions about rationality
3. charges of overconfidence
4. anti-authoritarianism
5. religion
6. ludditism

That scientific results challenge our self-image is the backdrop of my whole thesis; whereas for instance the *big bang* as a theory about the origin of the universe is uncritically accepted by the populace and many physicists because it merges well with traditional Western religious cosmogeny[103], other results suffer widespread rejection because they challenge our self-conception as special creatures of the universe. Underlying this resistance is oftentimes a kind of human-species self-love which would disgust us if displayed in an individual: human beings are seen as "creation's crowning glory", as a teleological goal of a guided universe – that is narcissism taken to extremes.

102 A fascinating account of this phenomenon is given in Buggle (1992, p. 325f), who shows how people of such caliber as the physicist Carl Friedrich von Weizäcker use sub par reasoning to justify their religious beliefs due to psychological constraints on their thinking and the need to "protect" deep-rooted emotional indoctrinations.

103 The big bang is, as so much in modern physics, heavily theory dependent. There are far more interesting cosmological models, for instance the cyclic universe (Steinhardt & Turok 2002), in line with Hindu mythology.

Science is a path with a goal – knowledge – but no destination, and indeed, even the path may seem to fade before one's eyes when the journey is long and arduous. It is an attitude of not wanting to relent in the quest for knowledge; not to be content save with the best we can do. A will to think, and always think anew. Adopting the scientific world view primarily means relinquishing any claim to possessing truth; relinquishing above all the dearly held truths which stand between oneself and the void. This makes people fearful – and that is the origin of opposition to science. Cherished thoughts, traditional world views, power structures – all crumble before the advent of the rational mind. But we can conceive of reality in ways which is more in accord in science than traditional pictures of the world. I deal with this in chapters 4 and 5, so not much more need be said here.

Points 2 and 3 are easily shown to be wrongheaded; they have been partly addressed above. Point 4, anti-authoritarianism will be looked at closely, especially in its guise of radical constructivism.

Point 5 is about a vested interest in not accepting rationality, religion, which will also be looked at in a little more detail.

Point 6 is, contrary to the other points, logically coherent, but dangerous in the universe we live in.

Rationality is about taking the consequences of one's beliefs into account. All irrational positions suffer from the malady of not doing this. It begins that their very starting point, that is, "arguing" against rationality, is already incoherent. If one decries reason, why bother arguing? The only honest objection against the rational world view is a shotgun.

2.9.2 Misconceptions

A quote in case is George Bernard Shaw's[104] phrase:

> The reasonable man adapts himself to the conditions which surround him. The unreasonable man persists in trying to adapt surrounding conditions to himself. ... All progress depends on the unreasonable man.

This is the kind of reasonableness the rationalist definitely does *not* have in mind. The phrase "be reasonable" may be a rhetoric tactic in everyday life to make other people conform to one's

104It is attributed to George Bernard Shaw in Bartley (1984), and diverse sources on the Internet. I was not able to track down the original source. In any case, the quote is illuminating even if it's not from G.B. Shaw.

wishes; but it has nothing to do with the art of rationality; indeed, the rationalist will often act very "unreasonably", that is, contrary to convention, in day to day situations. But this encounter with the word "reasonable" may well predispose people to react negatively when confronted with the appeal to exercise reason at all times.

Other misconceptions, such as "lack of emotionality" or not taking intuition into account have been dealt with above.

Some people seem to mistake rationality for sophistry. Maybe, for clarity, we should distinguish *sincere rationalists* – henceforth simply called rationalists – and *manipulators*. The difference would be that the former are interested in knowledge, understanding, explanation and ethics etc – while the latter are simply trying to push a (most probably hidden) agenda, a doctrine of their in-group for example. But it should not be too difficult to distinguish between the two in practice; the simplest criterion being if the person in question considers new evidence or not.

But the danger of falling into the trap of a manipulator, can't be side-stepped by avoiding a rational attitude. Even in a social environment favoring imprecise speech, there will come a time when problems will have to be solved in the real world (remember: as long as there is no need to act there is in principle no need to be rational; rationality is a requirement imposed upon us by the physical environment and the desire to reach goals in this environment).

And problem-solving requires commitment to some mental model of the world, which may be more or less accurate, and which has to be communicated and justified. But what now? How to avoid a rational attitude?

Even by renouncing reason, it could still be that influence is spread via other channels (manipulation need not go by way of argument); maybe a manipulator is saturating the environment with pheromones that inspire trust. One can never exclude that one is being manipulated. But the best one can do – and indeed, that is all the rational method demands, that one does the best one can do – is to start thinking rationally yourself.

2.9.3 Overconfidence

Another challenge of often raised is that the rationalist is overconfident in his application of reason; that the answer is probably unknown. Again, this criticism quite misses the point, because, when we are doing science or scientific philosophy, fallibilism is always presupposed – the understanding that we are constructing models that are most probably in some, maybe even all,

aspects, but nevertheless better than competing models. In science we always attempt to weigh theories with respect to one another, *given what we know,* not more, not less.

That is actually the difference between science and rationality versus other undertakings of the human species like ideology or religion: the people participating in the former are always aware that we are just in a state of approximation, and that there are facts of which we are currently quite ignorant.[105]

An example of how rational reasoning may look from the perspective of someone decrying the approach, and ways the rationalist could respond might go like this:

Imagine a rationalist weighing certain propositions, A, B and C assigning them a probability ordering, and coming to the (conditionalized) conclusion D. The skeptic is patiently (more often, impatiently) listening and thinking that this is all to no avail because the rationalist has not considered propositions E and F and apart from that, many unknown relationships and facts may enter the picture in the future.

But, again, the skeptic is being inconsistent if he denies the rational approach to questions: because if he has thought of propositions E and F, he must reason why they apply, and the rationalist will happily accommodate. As to the unknown relationships and facts invoking them is an unfair move because they are per definitionem unknown; the rationalist takes these possibilities into account with the fallibilist stance.

A rationalist who does not act in the above way – that is, flexibly adjusting to new propositions ventured or being *too* sure about his conclusions, is not a true rationalist. Is this a "no true Scotsman"[106] argument? No, because being a rationalist is not like being a member of a political party or a religious movement or having certain ancestry, where those criteria are constitutive of the

105In Bayesian terms, we would ascribe a large portion of prior probability to an "unknown true theory" and make sure that our updating rule would guarantee that this slice never shrinks to zero. Of course, in practice, as we use Bayesianism to compare existing hypotheses, we might as well drop this procedure altogether and just calculate the relative posteriors. But it is good to keep the above procedure in mind when thinking of our brains as Bayesian inference machines, and how the very brain states of a rationalist would incorporate the fallible nature of our knowledge.

106This fallacy is described in Flew (1975)

> Imagine Hamish McDonald, a Scotsman, sitting down with his Glasgow Morning Herald and seeing an article about how the "Brighton Sex Maniac Strikes Again." Hamish is shocked and declares that "No Scotsman would do such a thing." The next day he sits down to read his Glasgow Morning Herald again and this time finds an article about an Aberdeen man whose brutal actions make the Brighton sex maniac seem almost gentlemanly. This fact shows that Hamish was wrong in his opinion but is he going to admit this? Not likely. This time he says, "No true Scotsman would do such a thing."

respective label. Being rational is a method to which you can adhere to or not (in varying degrees); solely the application of this method makes you a rationalist or not.

When a scientist or a rationalist are very certain about something, it is usually because they have considered a lot of available evidence and pondered long and carefully over possible alternative hypotheses. They will not repeat the long chain of argument every time they present some theory, simply because this is not possible. From this may come the false impression of certainty without sufficient backing; the skeptic who is of a different opinion than the scientific establishment is well advised to embark equally upon a quest of learning and thinking. Simply shouting "overconfidence" will not to the job.

2.9.4 Anti-authoritarian

Another variant of resistance to science, one unfortunately meeting with much sympathy in some academic circles, are those of the deconstructivist, postmodernist and relativist slant. But while the original intention of these movements may have been noble – to uncover oppressive structures and pave the way for new avenues of thinking – the approach does not work.

While it may be feasible to be a relativist in an academic seminar, at the very moment where one has to solve real-world problems, where one has to take action and make decisions, these "philosophies" do not provide any guidance for resolving conflict. Such situations will immediately degenerate into a power struggle.

What does it actually mean to place a subject outside of science and rational discourse? People who are not willing to be rational on certain subjects but have opinions nonetheless are actually playing a game of power.

The answer to the problem of entrenched opinion, dominance and authority is not to advocate relativism, but to advocate reason. Problems never arise because of too much reason, always because of too little reason[107]. It is time to look more closely at one of these "philosophies" (as a proxy for all of them), namely, radical constructivism.

107This is also my answer to Horkheimer & Adorno (1947) ("Dialektik der Aufklärung") – what they attempt to criticize, albeit in unclear words, is lack of reason, not abundance of it.

2.9.5 Radical Constructivism or Reason gone Wrong

> "Reality is that which, when you stop believing in it, doesn't go away." Philip K. Dick[108]
>
> Before I had studied Zen for thirty years, I saw mountains as mountains, and waters as waters. When I arrived at a more intimate knowledge, I came to the point where I saw that mountains are not mountains, and waters are not waters. But now that I have got its very substance I am at rest. For it's just that I see mountains once again as mountains, and waters once again as waters. - Ch'uan Teng Lu (22)

The Zen quote above illuminates the transformations an epistemic agent undergoes. The first stage, when mountains are mountains and waters are waters, is naive realism. It is the child's view as soon as it learns concepts; a car, a cat, those things are real in virtue of the way they are conceptualized; there is no process of critical reflection.

Adults usually transcend this stage at some stage of their life, and reflect on their concepts and come to realize that a concept must not necessarily correspond to a thing in reality[109]. One sees through language concepts and societal concepts; many things learned at school and the conventions of human society are seen to be arbitrary categorizations; reality becomes richer, more open to opportunity, for creative action, but also less anthropocentric, less homely. Few people attain complete dissolution of all concepts, including their very personal "essence" – but only complete dissolution opens the way to stage three. Incomplete dissolution, that is, seeing through the fabrication and construction of some constructs but leaving others (unconsciously) intact leads to problems.

The third stage is attained only after painful processes of personal transformation. Things are now again seen as they are, in immediate experience. Concepts are seen as sometimes helpful and benign, but are dropped when recognized as harmful to progress. Reality becomes malleable to the mind without becoming in any way relative or subjective: rather, one attains the skill of drawing up conceptual borders at will around more or less stable features of objective and invariant reality. Achieving this third stage is very difficult – it does not suffice to read about theoretical results of

108"How to Build a Universe That Doesn't Fall Apart Two Days Later" by Philip K. Dick, 1978. Online: http://deoxy.org/pkd_how2build.htm

109Of course, many concepts are protected from critical reflection, as delving too deep causes fear: concepts relating to one's identity, for instance. Some central tenets of religion, like the possession of a soul, are often too painful too drop. Here, conceptualization remains naive.

science and philosophy of science. The hard part, as always, is integrating this knowledge into the deep neural structure of one's brain. Sometimes, one can't attain deep insight alone; one must seek teachers, mentors or therapists.

A philosophical position which as a whole failed to attain the third stage is *radical constructivism (RC)*. Radical constructivism asserts that nothing in the world is knowable – epistemological solipsism. All is only construction of the mind. Reality in itself is not denied though, as in ontological solipsism, but it is unclear what is gained by that move.

Radical Constructivism is not only an old hat – it is extreme scepticism in a new guise – but also dangerous, because it is a straightforward leap into relativism. It evokes concern in the literature:

> Der sogenannte 'Konstruktivismus' ist … die gefährlichste moderne geistige Tendenz, und, so darf man wohl sagen, eine der am weitesten verbreiteten Auffassungen. Er verbindet zwei Kantsche Ideen mit dem modernen Relativismus, nämlich die Idee, dass wir die uns bekannte Welt mit Hilfe unserer Begriffe herstellen, und die, dass wir eine von uns unabhängige Welt durch unsere Erkenntnis nicht erreichen können. (Albert 1996, p. 15f)

And:

> Constructivism has recoiled from Realism, that is, only to succumb to its equally untenable opposite. … a failure of imagination: an inability to talk about the world except in terms subsequent to the registrational achievement. And that in turn has led to three insurmountable problems. First, it has meant that the constructivist's only option in describing the world prior to the act of registration is to deny that what was only then registered ever existed (this is the source of the shock). Second, even more seriously, it supplied no apparatus with which to describe the true nature of the registrational achievement, thereby somewhat ironically denying the constructivist any way to describe the metaphysically constitutive act of construction. Third – this was perhaps the most obvious of all, to the outsider – it failed normatively, giving one no handle on what was important or significant. (Smith 1996, p. 353)

It should be stressed that Smith does find positions of merit in constructivism – as I do; but radical constructivism simply goes one step too far; or, staying in the imagery of the Zen quote above, one step too few.

The philosophy only works if there is a radical cut between knower and known, and, despite proclamations to the opposite, assuming this cut essentially underlies radical constructivism[110]. Maturana & Varela (1987) develop the metaphor of a person sitting in a submarine submerged in an ocean, and reading values off of controls. The person always only sees control readings, never the ocean outside; and this example servers to "illuminate" the epistemic cut between knower and known. I think it is no coincidence that this metaphor is reminiscent of the homunculus so pervasive in naive conceptions of the self. The position of RC is usually presented as absolutely revolutionary. But the assumption of closed cognitive entities which have no direct access to an external world is actually a very deep Cartesian intuition which underlies much of Western philosophy; it is not revolutionary, but quite mainstream.

Among the core tenets of the constructivist approach, as taken from Alex Riegler's guidelines[111] for submissions to the Journal "Constructivist Foundations", are the following:

- According to constructivist approaches, it is futile to claim that knowledge approaches reality; reality is brought forth by the subject rather than passively received;

- Constructivist approaches entertain an agnostic relationship with reality, which is considered beyond our cognitive horizon; any reference to it should be refrained from;

- Therefore, the focus of research moves from the world that consists of matter to the world that consists of what matters;[112]

The first point above shows clearly that, despite assertions to the contrary, the subject/object dichotomy is upheld. The traditional view of knowledge being passively received is supplanted by the complementary and dialectically linked view that the subject brings forth what is perceived. Both positions are false. There are only relative configurations, but more on that below. *Knowing* is a *mode of reality*, not something separate from it. And while of course everything we know quite trivially is dependent on the state of our mind, the distinguishing feature between knowledge and fantasy is that the former exhibits a certain harmony with external affairs, whereas the latter do not. I do not know what is intended by blurring the distinction between the two. That the world has no knowable structure is the naive assumption of some structured entities in the structured world.

110There is no such cut. More on that below.
111 http://www.univie.ac.at/constructivism/journal/faq.html#denominators
112 Riegler (2001) offers a clear and concise overview of core constructivist positions.

Structure is necessary for memory, thought, identity, everything what makes persistence of cognitive entities possible. Without structure there is no doubt, making the doubting of structure self-refuting.

Points two and three are even more revealing: given those, it is not clear why we should continue to practice scientific research. Why not leave the construction of knowledge to artists and poets? While this may be appealing to some people, the drawback is of course a purely practical one: poets will not help it getting real problems solved, such as building bridges that don't collapse, nourishing people that are starving, battling diseases and what have you.

If we dispense with an ontological scratching post (that is, reality), we can believe anything we want. Science is not necessary anymore, and any thought of Enlightenment must be relinquished, a prospect welcomed by some, for whatever reasons and agendas.

The reason for positing points two and three above is the purported cognitive closure of epistemic systems derived from biological considerations. The objection is obvious: if cognitive closure follows from empirical science, but this empirical science is actually only a construction due to cognitive closure, it can't be epistemically relevant: radical constructivism, once more, refutes itself.

Schmidt (1987, p. 39f) recognizes this; the rebuttal he offers is not an argument though, but mere restatement of ideology: the status of empirical theories is declassified as mere "intersubjectively shared operational knowledge"; but the question immediately arises: *shared by whom*? As radical constructivism is construed as epistemological solipsism, why take other human beings seriously? They are, after all, also "only" constructions; why offer other human beings special privilege in the epistemological landscape? The sleight of hand usually goes unnoticed because we conceptually weigh humans higher than other entities of the world. Radical constructivists should undertake the project of delineating the difference between epistemological and ontological solipsism. For all practical purposes, as mentioned above, the two should collapse.

The grounding in the world needed to share experience and ensuring a correct grip on reality was provided by the blind algorithm of evolution. An account for how this leads to epistemic success is given, as mentioned above, by evolutionary epistemology (Vollmer 1975; Vollmer 2003). The Kantian conditions for the possibility of experience would be given a simple explanation in evolutionary epistemology. Sense organs and neural structures would evolve in such a way, as to be able to know the very reality of which they are a part of.

An illuminating example of how organs adapt to external situations in the world is the *eye:* we perceive those frequencies of the light spectrum of the sun which penetrate the atmosphere and thus are *available* for tracking the environment.

The light frequency on Earth attains its maximum at wavelengths between 400 and 800nm – this depends on the elemental composition of the sun, its age and its size, the distance of the Earth to the sun and the composition and thickness of the Earth's atmosphere etc. Our eyes perceive in the spectrum of 380-760nm (Vollmer 1975, p. 98). The connection is obvious. And there seems to be no evolutionary advantage of being able to see the minimal amounts of UV-light or gamma radiation present in our environment at the current geological development of the Earth and stellar development of the sun/solar system – although evidently it would be great fun to be able to do that.

Evolution works with what it is has – new structures avail themselves of older ones, leading to exaptation and other evolutionary phenomena such as suboptimal design of organs[113]. A constructivist could argue that maybe evolutionary suboptimal branches selected for at the beginning and which track reality wrongly would rule out "correct" knowledge of reality. One thing speaks against this: analogous development, that is, convergent evolution, for which constructivism necessarily can give no satisfactory account. The independent development of the eye in vertebrates and cephalopods is a case in point (Futuyma 1998, p. 110f). The existence of analog development of organismic features is a powerful indicator that there are states of affairs "out there" which are attractors for evolution.

And even *if* an initial "fit" were a simple instance of satisficing, and not a mapping of real structures onto internal models, nothing would prevent evolution from discovering, with time, species which were better adapted to exploit features of reality and dominate species which only exhibited an adequate "fit". In fact, this did happen: humans developed the most sophisticated mind on the planet, an organ which lets them plastically map real world structures in near real time, as opposed to the gradual generational adaptation of evolution; the result is near complete domination of the planet[114].

Evolution also enters from another angle. Even RC can't do without a basic notion of consistency. But from where does this criterion of consistency come from? From an immaterial

113 A well known suboptimal design is the optic nerve leaving the eye, leading to an absence of retina at that location which then leads to a blind spot in vision. (The brain fills the blind spot in, so that it is not noticed in ordinary vision tasks.)

114 This remark is to be read in an ethically neutral way. It is simply a statement of fact.

mindspace independent of corporeality – that would be a strange commitment for someone denying basic reality.

The ability to recognize consistency and inconsistency in the first place is due to our having implicit models[115] of what is consistent – these models having their grounding in basic perceptual skills which we owe, again, to our evolutionary heritage. The empirical evidence for this can be extracted from number skills present in infants and even animals: *innate arithmetic* (Lakoff & Nunez 2000). Basic numerical skills are present in many living beings before any mathematical training is performed – not only in humans but also in animals. As animals can't talk this requires some ingenuity in experimental setup and of course contains some interpretative leeway, but the description of the experiment is suggestive:

> Rhesus monkeys in the wild have arithmetic abilities similar to those of infants, as revealed by studies with the violation-of-expectation paradigm. For example, a monkey was first presented with one eggplant place in an open box. Then a partition was placed in front of the eggplant, blocking the monkey's view. Then as second eggplant was placed in the box, in such a way that the monkey could see it being put there. The partition was then removed to reveal either one or two eggplants in the box. The monkey looked significantly longer at the "impossible" one-eggplant case, reacting even more strongly than the babies. (Lakoff & Nunez 2000)

Even the Rhesus monkey, it seems, has some sort of consistency constraint on his worldview: the multiplication of an eggplant without any kind of visible cause was surprising for the animal; we may speculate if this surprise requires *some* empirical input or if it is entirely innate, but that is beside the point: both are anchored in the physical world.

If we would live in a world where things popped out of and vanished back into thin air, we would not be surprised by the eggplant scenario. I contend then that the very criterion of consistency, also as it finally appears in logic (such as the law of excluded middle), is an abstraction derived, ultimately, from the physical world we live in. The underlying pattern for the abstraction of consistency is nothing less than the stability and structure of the physical world. The only criterion to which the constructivist may appeal to defend himself against the objection that his philosophy can't even distinguish between stable constructions and pure fantasy – consistency – finds its roots in the very external world he wished to abolish. Constructivism operates with concepts such as

115 And from which more complex forms may bootstrap.

viability and usefulness and consistency: but what are these if not impingements of reality on constructivist fantasy? Enough of evolutionary considerations and on to motivational issues.

A motivation advanced by philosophers advocating radical constructivism is the apparent liberating effect of the doctrine – a liberation from a "constraining reality" which grounds truth and where "claims to absolute truth necessarily lead[ing] to oppression"[116] (Schmidt 1987, p. 47). But this, firstly, misconstrues the aims of science, which, while aiming at truth never gives up the postulate of the *tentativeness* of its knowledge. That is the essential difference to claiming to be in possession of absolute truth, as we saw in section 2.5.4. Also, the claim hat there can be no knowledge/truth is already a dogma: it's formal structure – it being an indefeasible paradigm – makes it a dogma, even if it asserts it's own contradiction. Whenever we take RC seriously, it refutes itself. That is the sign of a defective philosophy.

Most importantly, the fear constructivists have of real knowledge – that is, that propositions that correspond to external reality necessarily lead to oppression – is entirely misguided. Our physical environment and it's exploration provides a common space for agents to build a *community*; knowledge is the basis for communication, for being able to relate to the other being and take her concerns seriously. The constraint reality provides is not be lamented but to be acclaimed: it is what makes interaction possible. An unreal world open to arbitrary construction is one where every agent is damned to live in solitude and isolation. To put it harshly: radical constructivism is inimical *to love of the other*.

We may draw up the following diagram:

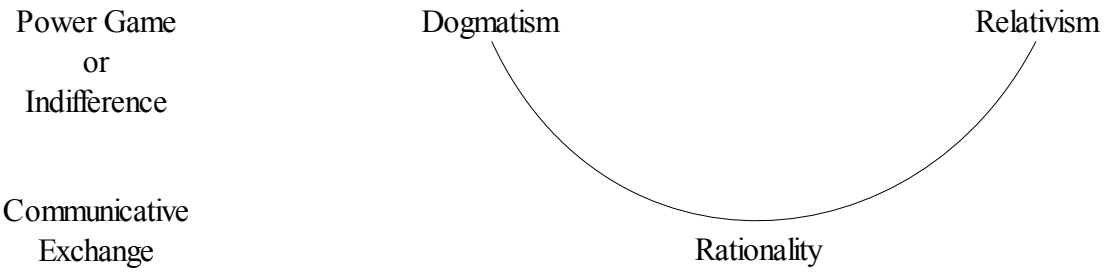


Illustration 2: The Middle Way of Rationality

One the one hand, there is communicative exchange – interest in what the other has to say; on the other side there is total disinterest in the other sentient being, which can express itself as either

116Translation from German by the author.

indifference, which ultimately underlies all forms of relativism (the other is even denied the right to be wrong); or as power struggle, that is, the imposition of the will of the stronger agent on that of the weaker agent.

The area of communicative exchange is maximal where rationality – which only makes sense if it is grounded in realism, as we saw above – prevails. The refusal to apply rationality is the proclamation of a willingness to exert power[117]. Rationality, on the other hand, is the commitment not to apply power, but reason.

A qualification is in order: rationality must ensure, if it wants to endure, that it is not swept away by more primitive systems exerting power. Rationality is a way of *structuring* discourse in such a way as to lead to maximal amount of sustainable diversity in intellectual and social life. So, to be precise, one can't evade power games completely, but one can structure them and reduce their domain of applicability. But the rational way is not just any structure. The fundamental asymmetry of the rational way that makes if different from other ways is this: the rationalist can be quite comfortable with other opinions, as long as they do not seek complete domination of discourse; the dogmatist, on the other hand, is never content with variety. And the relativist (like the radical constructivist) has no arguments and no persuasive power against the dogmatist or fundamentalist; relativism is not sustainable: it is powerless in the face of an aggressive doctrine. Being rational means being able to sustain diversity *robustly*.

Another attraction to constructivism may come from the simple fact that social norms and conventions are constructed. That human laws, norms, traditions and mores are constructions of the human mind is trivially true (that they may not be arbitrary is another question, one whose answer may be found in evolutionary game theory). Most power exerted today via hierarchies is based on constructions; but their power is no less real than the kinetic energy of an incoming meteorite: whereas the latter can be modeled by Newtonian physics, the former may be approximated via complex systems modeling. The power immanent in social hierarchies is historically grown via human interactions over generations and the patterns of these interactions being present in human brains; their power results from intentions of agents in accordance with those constructions.

Scientific realism is the intellectually more honest and satisfying response to the issues raised above; scientific realism rejects all our *naive preconceptions* of reality but does not rest in

117 Indeed, I was myself a bit surprised at the atmosphere prevalent at a conference concerning radical constructivism: discussions where heated and dogmatic; the difference to a physics conference in the same year, where appeal to experiment and reason were made, couldn't have been more different. Such is the price one has to pay if one is prepared to relinquish the objective and the rational.

complacency; it is a call to inquiry, not a lulling ideology. For scientific realist, academic respectability is not enough for taking a theory seriously; it must be applicable to the real world and help solve real problems. Every model remains fallible; will be criticized, changed, replaced. But models will not be replaced arbitrarily. They will be replaced because better models are found, and the criterion for "better" is a more precise tracking relation to the external world.

Constructions have their place in this view. As acknowledged above, every thought is trivially a construction in the sense that brain states of the cognitive entity are necessarily involved in every cognitive act performed by a material being. Where humans tell stories about the world, there is construction. Science is about building robust constructions, robust in the sense that they correspond to actual patterns in the world, and the scientific realist can account why some constructions are better than others, namely by appeal to external reality. And thus the history of science is not only about *construction*, but also about *deconstruction* of naive world models. Have not the best achievements of science radically overturned the way we view ourselves and our place in this universe?– science is first of all a *destroyer* of constructions. Science has driven traditional world views away wherever it has ventured (geocentric cosmology, absolute time and space, divine origin of humankind etc etc).

The lesson, then, to draw from the misguided philosophy of radical constructivism is this:

> The alternative to rationality – the willingness to engage in reasoned communication, reason deriving from a shareable environment – is unstructured power struggle and domination.

2.9.6 Religion

> Science can destroy religion by ignoring it as well as by disproving its tenets. No one ever demonstrated, so far as I am aware, the non-existence of Zeus or Thor - but they have few followers now. (Clarke 1953)

> Some of you say religion makes people happy. So does laughing gas. So does whiskey. Clarence Darrow[118]

What about religion? Religious people often claim that their knowledge is of a different kind and not amenable to rational criticism. From the rational point of view, taking into account

118 http://en.wikiquote.org/wiki/Clarence_Darrow; Quoted in an eulogy for Darrow by Emanuel Haldeman-Julius (1938). Retrieved 22.08.2009.

everything we know about evolution, society, in-group ethics, human psychology, biases etc this claim is obviously a claim designed to protect propositions from revision, and would merit no further discussion.

But due to practical considerations, that is, the real force this kind of argument has on people not trained in the art of rationality, some words are in order[119]. The following exposition is a reduced version of Gowder who asks the question if the pope should be Catholic[120] and argues as follows:

> Basic 1: Theological claims are beliefs -- they are statements with propositional content that refer to the world.
>
> Basic 2: Beliefs should be held for reasons. ... (normative, rather than explanatory reasons).
>
> Basic 3: Beliefs have objective and universal truth-values. It is flat-out incoherent to utter sentences like "God exists for me but not for you." This might not be true about everything, but it is manifestly true about ontological claims about the world, historical claims, etc. (Gowder 2008)

The pope certainly believes that he has reasons for his beliefs; also, he should acknowledge that whatever reasons he has, these reasons are probably available to people of other denominations as well; or he could believe that the evidence leaves the possibilities open, but then again, he should take others seriously. All these points merit more discussion, but are not of the essence. The most interesting point raised by Gowder is this:

> [The pope] believes he has access to reasons that are fundamentally incommunicable. Gnosis. There are two worries about this claim. First, is truly incommunicable belief really impossible? Agents ought to be able to communicate the fact of their gnosis: the pope ought to be able to turn to the mullah and say "I experienced a gnosis, and so did these millions of other people," and that ought to count for the mullah, if the mullah holds his beliefs for reasons ... Second, if it doesn't count for the mullah, maybe it's because the mullah

119 Of course, for those who renounce reason altogether – at least on this subject – there is no path out of the religious worldview, except maybe in-depth personal discussions of what is actually at stake.

120 The post was inspired by the Pope Benedict's statement that an interreligious dialogue is not possible (http://www.nytimes.com/2008/11/24/world/europe/24pope.html?_r=2&hp). A fortiori, a dialogue with scientists who grant even less commonalities with the Pope than adherents of other religions among is equally not possible.

> experienced a gnosis too, and that possibility, of course, ought to count for the pope[121]. (Gowder 2008)

This is the fundamental dilemma of the religious person: that even if he claims that his knowledge has some special epistemic status (implanted directly by God), the very possibility of this act (of God) must allow him to speculate that another person could also have received his beliefs directly, through an act of God.

Now when he encounters different persons, A, B, and C, of which only A agrees with his beliefs, but B, and C voice different beliefs, but also justifying them with a gnostic event, there is an impasse. The rationalist, not believing in such epistemic miracles, can readily deal with such a situation: he can give naturalistic accounts of how the people have arrived at their beliefs. The religious person would require the following argumentation:

1. Gnosis is possible.
2. I have experienced Gnosis, and I know which parts of my knowledge base is gnostic (that is, the gnostic knowledge includes meta-knowledge about *what* is gnostic).
3. The persons with different religious beliefs only think that they have experienced a gnostic event. In fact, they are deluded, and are suffering from a fantasy.

The problematic proposition is of course 3: it encodes the belief that one is in a privileged position, namely, the only person having had a real gnostic event and knowing it as such. Of course, one is free to believe such a thing. But everybody must decide for himself if such a strategy is satisfying. Stephen F. Roberts put it simply[122]:

> I contend that we are both atheists. I just believe in one fewer god than you do. When you understand why you dismiss all the other possible gods, you will understand why I dismiss yours.

Worrall puts the point another way:

> Finally, I have made it clear that my whole argument rests on the assumption that a rational, scientific person needs good evidence before admitting god into her worldview, just as she would before admitting, say, electrons into it. Alvin Plantinga has mounted a well-known defence of the striking claim that belief in god can be

121These are actually two points. I will concentrate on the latter, the possibility of multiple gnostic events.
122 http://freelink.wildlink.com/quote_history.php. Retrieved 01.08.2009

> 'properly basic' – that is, taken to require no evidence. [Footnote 12: See, for example, Plantinga (1981)] Although again it requires detailed treatment which I cannot give here, I should at least indicate my response. This is that, on analysis, Plantinga's view amounts to no more than the obviously true descriptive claim, that some people as a matter of fact take belief in god as basic. But this is no news, the question of course is whether or not they are justified in doing so; and, in so far as Plantinga has anything to say about this issue, it seems to rest on the sort of simple-minded relativism that I have throughout taken to be eschewed. His response, for example, to the obvious question of why in that case one couldn't take belief in a flat earth (or come to that, the innate superiority of the 'Aryan' race) as 'properly basic' seems to be simply that no Christian would in fact take – or is under any obligation to take - such beliefs as 'properly basic'. This, however, is plainly not the issue – the question is what such a Christian would say to someone who did assert as 'properly basic' (that is on no basis at all) a claim that s/he, the Christian, found abhorrent – and, assuming that she would want to challenge that claim how s/he would deal with the tu quoque objection. Long live evidentialism! ...(Worrall 2004)

2.9.7 Ludditism and Existential Risks

The current technological-global situation is very unsafe: there are a multitude of different nations, religions, ethnicities, and tribal thinking with in-group and out-group ethics is still very much alive. In bygone ages tribal thinking discharged itself into local wars. Bad enough as these were, the risk is incomparable to today, where the deployment of weapons of mass destruction looms.

Renouncement of science, technology, the progressive-rational view of the world and a return to simpler ways of living – ludditism[123] – may seem like a valid response, if only theoretical in nature. But it is a *possible* argument against human achievement and further progress and that is why it shall be shortly addressed.

Every new bit of knowledge, every new useful conceptualization makes the world an agent lives in bigger. Knowledge expands the possibilities an agent has at her disposal, be it potential for action or pursuit of the arts. But what holds for a single agent holds all the more so for humanity as a

123 There is a neo-luddite movement (Sale 1996).

whole: knowledge expands our possibilities. An obvious case in point is medical research, where each new discovery leads to ease of suffering and new hope for patients: already in this "local" sense, scientific research is a moral duty (Harris 2005)., where the luddite is hard pressed arguing for a return to simpler ways of living.

But there is also a bigger, more dramatic sense in which scientific research is a moral duty: existential risks. There is an increasing awareness building up in academia regarding these problems (Bostrom 2002; Posner 2004; Hanson 2007; Bostrom & Cirkovic 2008). Existential risks, while including man-made risks such as threat of nuclear war, also concern purely natural events. For example, over geological time periods in Earth's history, there have been several mass extinction events (Raup & Sepkoski 1982)[124]. Nature is more inimical to life than today's environmentalist romanticism would have it.

Simple reasoning shows us that everybody who values human civilization at all can't advocate a turn away from technological development.

Life as evolved on earth depends on earth-like conditions persisting. While gradual shifts can be compensated, a rapid change in the environment can't be evolutionarily countered. Also, there does not seem to be a jumping point from water-based life on earth to life in a vacuum; *evolutionary* conquest of space remains, as seen from today's knowledge of biology, the purview of science fiction. That is, there is no natural solution which guarantees ultimate survival of life when environmental changes on the planet occur that are too harsh and rapid.

The conditions on earth *will* change radically and inevitably sometime in the future; at the latest when our sun dies[125]. Call this event *event x*. This is a purely abstract event not corresponding to a *single* physical event to which one could easily point. A possible exception would be a large-scale meteorite impact.

So, if we want life to persist past *event x* (not only human life, but all life including plants, animals, bacteria etc), we have to develop technology at some *time point y temporally prior to event x* which enables us to avoid extinction.

Now, why postpone time point y to some indefinite future? We have surplus existential risks through certain technologies *at the moment*, so it is highly advised that we transit through this state

124 Bambach, Knoll & Wang (2004) present a modern view.
125 See Bounama, von Bloh & Franck (2004) for an analysis of why life on Earth will have failed much earlier than previously supposed.

rapidly[126]. Given the above reasoning, everybody who wishes life to continue indefinitely and not end at a far future but definite point can't advocate a return to an anti-technological age. Therefore only one alternative remains: the way forward.

Following through on such a project of countering existential risks – for instance in the form of increased funding of space exploration and colonization[127] – would also provide the additional benefit of bringing a grand vision to a global society. And humanity needs a new frontier.

2.10 Rationality, the Dao of Thinking

> If you want to understand, really understand the way things are in this world, you've got to die at least once. And as that's the law, it's better to die while you're young, when you've still got time to pull yourself up and start again. (Bassani 1962) [128]

Sapere aude![129]

Becoming a rationalist is a bit like dying – that is, the former person you were dies, and a new, stronger person is born in its stead. The rationalist eliminates all his beliefs that do not hold up to scrutiny and starts looking at the world without fear. The algorithm of inquiry never halts. There will never be a time when all problems are solved and the enlightened society is completed; when somebody from the Outer Heavens arrives with chariots and trumpets and proclaims: "well done, humanity!".

The enlightenment – the will to think – is a process which demands commitment, again and again. That is the true challenge posed to beings wishing to follow the path of rationalism: to

126 Modernity and technology have caused a lot of problems, be it environmental pollution or social upheavals. But every problem created by technology is transitory – either because the technology will be abandoned or because, as the problems are being recognized, new technologies are created to cancel out negative effects of the previous technology. But the transition needs to be done via technological development, aided by scientific inquiry which proceeds rationally.

127 In a similar vein Lamb:

> Recognition of the precarious status of life on Earth may herald a renewed interest in colonization. Repeated warnings of asteroid collisions with the Earth and similar global catastrophes which, in the past, have decimated life on Earth, may drive home the point that the human race has too many eggs in one fragile basket. For if we are alone it might be argued that we have a duty to export life. Lamb (2001, p. 197)

128 Inspired by Riemen (2008).

129 From Epistle II of Horace's Epistularum liber primus: "Dimidium facti qui coepit habet: sapere aude: incipe." ("He who has begun is half done: dare to know: begin." Translation based on http://en.wikipedia.org/wiki/Sapere_aude, retrieved 24.06.2009). The phrase was popularized by Kant (1784).

engage actively in uncertainty; to seek uncertainty, not for the sake of it, but as a method for gaining knowledge – "the price of freedom is eternal vigilance"[130].

A short summary of the principles sketched above – also giving them a somewhat different slant – to which a rationalist should adhere is given below:

- Take into regard the fallibility of our senses, but also their adequacy in principle due to evolutionary pressure.
- Be aware that all knowledge of reality comes in the form of models – not only scientific models; religions and ideologies also only have models at their disposals.
- Know about biases and common inferential errors; acquire at least the basics of statistical reasoning and pitfalls that lie in that domain.
- Be prepared to face reality as it is, not as you would like to have it. Take heed of empirical results, especially if they contradict your own position. The rational actor can change the world in a second step, but the first step is always to know how it actually *is*. Constantly ask: what could be wrong with my way of seeing things? Be prepared to revise beliefs relentlessly; update models with incoming evidence; use plausible reasoning, where plausibility means integrating (cohering) all knowledge and not deliberately ignoring knowledge. Consistency is a minimal criterion, coherence a desirable criterion.
- Update everything, but not everything at once (scepticism without dogmatic adherence to scepticism)
- Be clear and precise in what you think and what you say.
- Criticize and receive criticism; the ability to criticize presupposes broad readings in a large number of disciplines to get the "big picture"; the ability to handle criticism includes decoupling personal value from beliefs held.
- Employ diverse scientific heuristics, never adhere to only one of them: verification (confirmation), falsification, unification, invariance, explanation; and even more contentious principles such as simplicity, elegance, symmetry and beauty. While in science the use of formal methods, especially logic and mathematics, is indispensable for being precise and to figure out what inferentially follows from certain theories, these are only *tools* for being rational. Observing logic rules of inference, the probability calculus, and heeding

130 Attributed to Thomas Jefferson; 3rd president of the USA (1743 - 1826)

empirically quantifiable relationships are prerequisite; this does not exhaust the rational mindset. Rationality is more than adhering to content-blind norms or to an allotment of norms depending on situation: rationality is about clear thinking and observation; intellectual honesty and epistemic virtue; rationality is a skill which takes a lifetime of practice. Rationality is about recognizing one's limitations and still bringing the best to bear on the problem one can.

- Construct causal models which lead to explanation, understanding and the ability to intervene in physical processes. Guiding principles in theory construction: Where does it come from? Why? Which purpose does it have? Always consider evolution.
- The goal is the search for truth not finding it – that would be stagnation and dogma.
- Do not believe something because of authority or tradition (only use it as a very weak heuristic).
- Train introspection.
- Be creative. Elimination of theories is one aspect of rationality, conjecture is another[131].
- Apply everything recursively and iteratively.

Note that we always only deal with hypothetical imperatives, as there is no obligation to pursue truth in the first place. Imperatives of rationality and critical thinking only follow if one *adopts the wish to know* the world in the first place. If one does not want to do this, one is free to follow arbitrary epistemic principles, or none at all.

If one adheres to the above principles one can hope that one's beliefs will track the world ever more accurately. Rationality, in short, is the Dao of Thinking: inner thoughts being in harmony with outer events. Let the following be our ultimate principle; the Dao of Rationality:

> If nothing within you stays rigid, outward things will disclose themselves.

131 A fruitful way of arguing inspired by Chalmers philosophical humor page (subsection "Proofs that P" at http://consc.net/misc/proofs.html)goes like this:

```
If not P, then what? Q_1? Q_2? q_N? therefor P.
```

The heuristic of presenting "insensible alternatives" is designed to foster the creative process to find alternative sensible models, thus contributing to the multitude of hypotheses. The mode of argumentation presented highlights the fact that we must look at all possible alternatives, and try to argue why some are absurd and others are not, and encourage the search for other hypotheses. People not content with P will try to come up with other Q's. This process is rational – looking for alternatives in hypothesis space is a basic requirement of rationality.
The cited page is hilariously funny and well worth a visit. The inspiration for the "creative argument" above was taken from the following parody on the argumentation style of Morgenbesser proving that p: "If not *p*, what? *q* maybe?".

I have now laid out some principles of thinking and rationality. I will not systematically refer to these principles of rationality in the following chapters, but the investigations below where guided by them.

3 Metaphysics

3.1 The Ontological View

3.1.1 Dereferencing Reality

When the wild bird cries its melodies

from the treetops,

Its voice carries the message of the

patriarch.

When the mountain flowers are in bloom,

Their full meaning comes along

with their scent.

(Smullyan 1977, p. 15)[132]

One of the central questions in the philosophy of science is the attitude one takes to *realism*. The question is not intended in the sense of "is there reality at all?", whatever that should mean[133] – but in the sense of *in what degree one should take one's theories at face value*; how much ontological commitment one is prepared to make. For instance, if a theory speaks of atoms or molecules, does that mean that these atoms or molecules really exist in a literal way, such as when we speak of "tables", we take tables to exist? An anti-realist stance (instrumentalism for example) would be content to say that atoms and molecules are simply useful fictions that make theories work, but remain agnostic as to wether these entities really exist.

In this thesis, I propose a minimalist ontology:

> There exists a world, and it is independent of human beings and, specialiter, human minds. Everything else is up for negotiation.

Maybe it helps to discern two words: reality, and existence. I will use the word "reality" when I refer to the simple proposition that something exists and that this is not dependent on the human

132 Quoted from Chung-yuan (1970).

133 Some say one can't prove that reality exists. But what could it possibly mean to prove its existence? And, for goodness sake, what could it mean to disprove it?

mind. In this sense, reality, while not being strictly defined, certainly has a discernible meaning, countering objections that "reality" is a word too vague to use in serious discourse.

Existence, on the other hand, will be used in a more technical sense. We must, from case to case, decide what exists; existence of entities is a scientific question. For instance, you can assert with some certainty that the text you are currently reading and you yourself exist. But what about the past? The future? What about imaginary entities? What about abstract concepts? But more on that later; first to reality per se.

As mentioned above "reality" is not, as some would have it, a vague concept. "Reality" is actually not a classical concept, so maybe it is confusing because of this. Concepts are abstractions – abstractions which are either contingent or convenient, as much of our knowledge of the everyday world; or abstractions which exhibit an underlying lawfulness of their own, such as the beauty and hidden symmetry of mathematics.

But how could one abstract from reality? Is not the very instant where you form a concept of reality already another aspect masked[134]? We can conceptualize reality a *little bit*: when we speak of reality, we do not mean a category learned by our brains such as "duck" or "tiger". It is rather that we mean the *yardstick* with which we measure our theories. Reality is the invariance that always springs up when we stratify the world in a way which still merits the name "theory". We all know what we mean by reality because we are part of this reality; in the same way that we all know what it means to be conscious.

Trying to explain these things – reality, consciousness (in their givenness, not in some functional sense) – is to hit a bottom. Descriptions, theories, knowledge – those are how we make our knowledge explicit, so that we can communicate; communication succeeds because of the shared reality; but not all forms of being are communicable; the *inside view* – the first person point of view in case of agents with a personality structure – is ineffable in the sense that it is *raw* being. Every communication about it is necessarily an abstraction and will fail to capture some aspect of this raw being.

The philosophical position of skepticism – that the world is unknowable – rests on the false matter-mind dichotomy. We can know this world because the disconnection so commonly assumed is a misconception. Being is already a form of (unpropositional) knowing, namely, what a certain

134 This actually also concerns abstractions from regions of reality.

physical configuration looks *from the inside*[135]. So we are part of this reality. There never has been a division between man and nature. Nature is not a stage into which man[136] has been placed.

Now, since Kant (recently), but actually since Plato (not so recently), it has become unfashionable to attribute a fundamental role to immediate perception, which is derided as mere appearance and not being the "real" reality, the "thing in itself", which is an unknowable noumena. Actually, there is neither noumenon nor appearance. Reality is, when perceived, simply that. This is not some form of naive realism. Naive realism is the position that things "out there" are perceived as they are; the present position is also not a kind of sense-datum realism where privileged knowledge of internal states is assumed. The position advocated here says that the sum of all physical events leading to the current physical configuration lead to a certain qualitative experienced state – that is how the state feels like. The important thing is the total abolition of any kind of subject-object dichotomy and also of the mental-physical dichotomy. The mental is simply seen as the "inside" view of a certain physical configuration; some of these inside views being more interesting than others. The consequences of this will be explored in section 3.2.

Back to the conceptualization of "reality": the word "reality" works in a similar way as the indexicals "here", "now, and "I"; more poetically, we could speak of *suchness*. It is the simple difference between reference and representation. A reference is simply a pointer to something else with no intrinsic information. For instance, a signpost saying "Vienna" and pointing into the correct direction would be a reference. A representation (such as map of Vienna, or mental images, concepts, ideas) contain intrinsic information. "Reality" is not a representational concept, but a referential one[137].

The pointer nature of the word "reality" and the false grammatical possibilities the word "be" offers in the Indoeuropean languages[138] have lead to a multitude of philosophical confusions on the topic. That reality simply is and can't be captured fully in words need not worry us; language is only part of reality, not the whole of it. Reality as a whole does not care about our language. If we get confused because we misuse our language, because it leads our thinking into obscure paths, that is

135 One is reminded of Russell's knowledge by acquaintance. But his stratification of matters was different than the one proposed here, and the ensuing literature has more to do with philosophy of language than ontology.

136The chauvinistic "man" is chosen here on purpose.

137Incidentally, mathematics is difficult to learn because its symbol are referential – they refer to concepts which have to be learned independently of the symbols; of course, good notation can incorporate some intrinsic information, but only relative to other symbolic knowledge.

138I am not qualified to speak about other languages.

not the problem of reality or the world; it is only our problem because it hinders our epistemic success and subsequently will lead to problems when we try to reach our goals.

Confusion arises when models are created at different levels and then compared with one another in their own respective languages. Reality does not contradict itself. Reality is one, is a unity; it has no levels (Heil 2003). Only models can contradict themselves.

At the bottom of many confusions lies the *Picture Theory,* as Heil says:

> Although my focus is on fundamental questions in ontology, I have a good deal to say about the relation language, or thought, or representation bears to the world. My contention is that metaphysics as it has been conceived at least since Kant has been influenced by an implicit adherence to a Picture Theory of representation. (Heil 2003, p. 5)

What is the picture theory?

> As I conceive of it, the Picture Theory is not a single, unified doctrine, but a family of loosely related doctrines. The core idea is that the character of reality can be 'read off ' our linguistic representations of reality—or our suitably regimented linguistic representations of reality. A corollary of the Picture Theory is the idea that to every meaningful predicate there corresponds a property. If, like me, you think that properties (if they exist) must be mind independent, if, that is, you are ontologically serious about properties, you will find unappealing the idea that we can discover the properties by scrutinizing features of our language. This is so, I shall argue, even for those predicates concerning which we are avowed 'realists'. (Heil 2003, p. 6)

The picture theory then quickly gives rise to a hierarchical view of reality:

> Once set on this course, we quickly generate hierarchies of properties. We discover that most of the predicates we routinely use to describe the world fail to line up with distinct basic-level physical properties or collections of these. We conclude that the predicates in question must designate higher-level properties. Now we have arrived at a hierarchical conception of the world, one founded on the inspiration that there are levels of reality. Higher levels depend on, but are not reducible to, lower levels. My contention is that the

> idea that there are levels of reality is an artefact spawned by blind allegiance to the Picture Theory. (Heil 2003, p. 7)

We can slice reality up in many ways. I can model an object at the Newtonian level (representing the object as a rigid body), at the level of molecules, or at the level of quarks and leptons, or an infinity of other levels – these are only *descriptional* issues. Needless to say ontological levelism is rejected by Heil. Of course, apart from computational difficulties that arise when we go to ever more fine-grained levels, the calculations, if our theories are correct, should be in general agreement (quantum indeterminacy will be dealt with below); they need not be in exact agreement, because abstractions at higher levels buy their power of abstraction exactly by *letting go* of the tracking of fine-grained causal structure, so that, with time, deviations from finer-grained models will arise. If levels are merely descriptional and are not imbued with ontological commitment, then we may ask about what we should be realists? The answer is that our theories capture *real relations* between material things, and it is about those that we should be realists. Being a realist about an electron does not mean imagining a tiny ball buzzing around a bigger ball (the proton), but being a realist about what the theory says the electron will *do*. This is not to be confused with instrumentalism, but to spell out the exact details will have to wait a few paragraphs.

Names – given to concepts, predicates – let us pick us out particular *things* and *relations*. But concepts are learned, because they are useful in relation to us humans. That is why we have the concept of the tiger but not the concept of a neutrino (except if we learn about physics); but the neutrino is, relationally seen, no less real than the tiger. Names and concepts only reflect the attention of the namer and not the "ultimate reality" of the objects named. The ultimate reality lies in the sum of all relations a region of reality has with all other regions. Every "name" of reality will fail to capture it because reality is the plenitude of all relations. Language fails, but this does not mean that there is not an objective reality or that it is unknowable. It just means that when we turn our eye on one aspect, we will neglect another. We are free to turn our attention to another aspect in the next instant.

When we dereference a reference, we look at the value which the pointer points to. For instance, dereferencing the "Vienna" signpost in the example above will lead us to Vienna. Now, in language, we can reference to other language constructs. For instance we can say that x is a pointer to "hello", and every time we invoke 'x' we actually insert "hello". When we dereference to reality, there are no words, there is no data, no system of abstractions – it is basic, uncommunicable, wordless being –

but it's not a bug, it's a feature! This dereferencing from language to not-language is what seems to cause a lot of confusion in the literature (in fact, this certainly contributed to the origin of skepticism, relativism and many other -isms).

Here we can connect with the objective/subjective distinction given by Mario Bunge. We will call propositions *objective*, when we expect that a dereferencing operation will be possible for both of us and lead to the same result. We will call a proposition *subjective*, if not every agent can dereference in the same way. Again we see that objective/subjective has nothing to do with "real" versus "imaginary" or "true" vs. "false". It is simply a categorization of statements in regard to their dereferencing conditions.

But does the above mean that we are completely free in our conceptualization of the world? Would that not entail relativism? The case is this: while we are of course free to conceptualize the world any way we want, there a better and worse ways to do this. A good criterion is given by Wimsatt who suggests the scientific criterion of robustness[139]:

> Things are robust if they are accessible (detectable, measureable, derivable, defineable, produceable, or the like) in a variety of independent ways. Wimsatt (1994)

Heil, differently:

> Our cat concept and our electron concept, in common with many of our concepts, are products of ongoing scientific exertion. Neither concept was invented by dreamers. Both concepts reflect serious engagement with the world. Both concepts circumscribe genuine mind-independent similarities and important worldly divisions. (Heil 2005a)

As always, Yudkowsky is especially lucid. He employs the word *Thingspace* which is a kind of configuration space[140]:

> The way to carve reality at its joints, is to draw *simple* boundaries around concentrations of unusually high probability density in Thingspace.
>
> Otherwise you would just gerrymander Thingspace. You would create really odd noncontiguous boundaries that collected the observed examples, examples that couldn't be described in any shorter message

139 See also Ross' rainforest realism (Ross 2000); or Nozick's invariances (Nozick 2001).
140 See Yudkowsky (2008e) for a more detailed description.

than your observations themselves, and say: "This is what I've seen before, and what I expect to see more of in the future."

In the real world, nothing above the level of molecules repeats itself exactly. Socrates is shaped a lot like all those other humans who were vulnerable to hemlock, but he isn't shaped exactly like them. So your guess that Socrates is a "human" relies on drawing simple boundaries around the human cluster in Thingspace.

...

Presenting us with the word "wiggin", defined as "a black-haired green-eyed person", without some reason for raising this particular concept to the level of our deliberate attention, is rather like a detective saying: "Well, I haven't the slightest shred of support one way or the other for who could've murdered those orphans... not even an intuition, mind you... but have we considered John Q. Wiffleheim of 1234 Norkle Rd as a suspect?" Yudkowsky (2008f)

and

A natural cluster, a group of things highly similar to each other, may have no set of necessary and sufficient properties - no set of characteristics that all group members have, and no non-members have.

But even if a category is irrecoverably blurry and bumpy, there's no need to panic. I would not object if someone said that birds are "feathered flying things". But penguins don't fly! - well, fine. The usual rule has an exception; it's not the end of the world. Definitions can't be expected to exactly match the empirical structure of thingspace in any event, because the map is smaller and much less complicated than the territory. The point of the definition "feathered flying things" is to lead the listener to the bird cluster, not to give a total description of every existing bird down to the molecular level.

When you draw a boundary around a group of extensional points empirically clustered in thingspace, you may find at least one exception to every simple intensional rule you can invent.

> But if a definition works well enough in practice to point out the intended empirical cluster, objecting to it may justly be called "nitpicking".(Yudkowsky 2008f)

Note that the talk of levels in the quoted passage above is only endorsed in an epistemological sense (I am not sure how far Yudkowsky's commitment would go). This conception of reality as presented above is not easily assimilated. Words are only words. Insights can't be pocketed as a Dollar or an Euro; they have to be constructed in one's own mind – new neural connections joined, others pruned. Such is the way of mind.

To convey to you what I mean I would like to introduce some Zen Koans. I urge you to think about them yourself, meditate upon them if you will. The basic reality of the world, when not cast in language – is nicely demonstrated by these two Koans from the Mumonkan:

> Case 7. Joshu's Washing the Bowl
>
> A monk told Joshu, "I have just entered this monastery. I beg you to teach me." Joshu asked, "Have you eaten your rice porridge?" The monk replied, "I have." "Then," said Joshu, "Go and wash your bowl." At that moment the monk was enlightened.
>
> Case 40. Kicking the drinking water jar.
>
> During his stay under Master Hyakujo, Isan was a cooking monk. As Master Hyakujo wished to send a monk to found the new monastery called the Great Mount I, Master Hyakujo told the chief monk and all other monks that he would choose the one who would demonstrate himself as the best among them. Then Master Hyakujo brought out a drinking water jar, put it down and said, "You cannot call it a water jar. Then, what will you call it?" The chief monk said, "One cannot call it a wooden stick." Then, when Master Hyakujo turned to Isan, Isan kicked the jar and walked away. Master Hyakujo laughed and said, "The chief monk lost it to Isan." He made Isan the founder of the Great I-San Monastery.[141]

In the scientific literature of the West, there is a wonderful book which could well be called a manual for a *Daoist ontology.* It is Smith's "On the Origin of Objects" (Smith 1996). Nothing short of a full reading can capture the essence of this book; but I will try to draw from few core passages

141Translations taken from http://www.angelfire.com/electronic/awakening101/mumonkan.html

which will be absolutely essential to understanding my further argumentation. I have put more elaborate excerpts into Appendix B: On the Origin of Objects.

Smith calls my "reality" an "ineluctable deictic flux" (IDF) – I will adopt this terminology and speak of IDF-reality henceforth. IDF-reality is the unified real thing. Analysis, categorization – fracture – is always the operation of epistemic agents who mistake their knowledge for the way the world really is.

First of all, Smith endorses the metaphysical principle of irreduction:

> ...I embraced a methodological criterion that, following Latour, I call a principle of irreduction. Essentially a standard of metatheoretic accountability, it mandates that no theoretical assumption – empirical premise, ontological framework, analytic device, instigative equipment, laboratory tool, mathematical technique, or other methodological paraphernalia – be given a priori pride of place. Every piece of metatheoretic apparatus should be "left open" in order to be subjected to critical assessment, raised up for skeptical analysis, and potentially revamped or set aside. Unless one is willing to adopt this strict a standard of suspicion, ontological biases and unwarranted metaphysical assumptions will slip through and derail subsequent analysis. (Smith 1996, p. 77f)

Note how similar in tone this is to Heil's ontological seriousness[142] mentioned in section 1.3.

Smith goes on to give the irreduction principle a metaphorical *commercial* slant which I will use henceforth because it makes for vivid imagery:

> that for each theoretical assumption (premise, framework, etc.), one be prepared to say:
>
> 1. Where one bought it;
>
> 2. How much one paid; and
>
> 3. How one got it from there to here (Smith 1996, p. 78)

Point one is intended to be seen as a reflexion on the *theoretical backdrop* in which a concept arose; point two on what one *loses* in making a certain distinction (one abstracts aways from finer

142 And to the tenets of critical rationalism (Niemann 2008).

distinctions in reality) and point three is a reflexion on how a *transfer* from one domain into another *affects* a distinction.

Smith also takes a clear stance against forms of relativism, which he calls "politics of pluralism in the large":

> If your culture is different from mine, I can say that we are incommensurable, but I can say nothing more, nor can I learn anything from you, nor can I interpret my scheme for you, or anything else. In spite of its protestations, pluralism-in-the-large does not require a theoretical commitment to make diversity a central metaphysical focus. (Smith 1996, p. 112).

Smith differentiates "particular" – a located patch of metaphysical flux:

> By 'particular,' that is, I mean something like 'occurrent': something that is located or that happens, something that is embodied, something for which there is a Steinian "there there." (Smith 1996, p. 117)

from individuals which are already the domain of abstraction:

> By individuality, on the other hand, I mean whatever it is about an entity that supports the notion of individuation criteria – something that makes 'object' a count noun, something that makes objects discrete. Somehow or other, an individual object is taken to be something of coherent unity, separated out from a background, in the familiar "figure-ground fashion. (Smith 1996, p. 119)

to then posit the *Criterion of Ultimate Concreteness*:

> No naturalistically palatable theory of intentionality – of mind, computation, semantics, ontology, objectivity – can presume the identity or existence of any individual object whatsoever.
>
> The name "Criterion of Ultimate Concreteness" is appropriate because, as will soon be evident, one of the things that individuals have, that physical phenomena lack, is concreteness's opposite: abstraction. (Smith 1996, p. 184)

Now we can introduce the important concept of registration:

> Given … the commitment to honor the Criterion of Ultimate Concreteness, that translates into the following more specific goal: to understand how a conception of objects can arise on a substrate of infinitely extensive fields of particularity.
>
> Except of course that this is an untenable way to phrase it. To say "a conception of objects" makes it sound as if the achievement is the subject's, by assuming a split between conception and what is conceived of. It also fails by making it sound as if the achievement is cognitive. Nor is anything gained by striking a more traditionally realist stance, and asking "how objects can arise on a substrate of infinitely extensive fields." That puts the achievement too squarely on the object. Both ways of putting it violate the mandate of avoiding an a priori subject-world split. In place of these dichotomous formulations, therefore, I will speak, unitarily, of registering the world.
>
> ...
>
> By 'register' I mean something like parse, make sense of as, find there to be, structure, take as being a certain way – even carve the world into, to use a familiar if outmoded phrase. (Smith 1996, p. 191)

The anti-essentialism underlying this metaphysical view is mirrored by epistemological pluralism. One should use different ontologies for different problems. An ontology is a way of ordering the world. From computer science we know that certain forms of knowledge organization (data structures) are more conducive to solve specific problems than others. An ontology is nothing more than a data structure for ordering the real world. So what we need in fact is a *meta-ontology*: a set of rules for constructing *useful* ontologies. A world-view building algorithm, so to speak.

With these concepts in mind, we are safe to proceed, and have a look at levels of reality anew.

3.1.2 Reductionism, Emergence and Complex Systems

> ... ontological pluralism sustained by metaphysical monism.
>
> There is only one world - that is what was important about realism. But its unity transcends all ability to speak. (Smith 1996, p. 375)

The reductive commitment I make is, again, to *one world*; that there is no such thing as a stratified reality, split into different layers according to better or worse descriptional skills.

Of course, it is as always helpful to discern at least the epistemological and the ontological side. Epistemology is concerned with the question of knowledge, and is thus deeply related to understanding. We can only understand systems at the right level of abstraction – not too low, not too high; but this is not an intrinsic feature of the world – the world need not be understood in and of itself – that is a human desire, accomplished in human minds with human cognitive limitations. Especially our limitation in holding many conceptual items in memory at the same time forces us to abstraction and to finding regularities in aggregates.

The ontological side of the equation is at the same time much more simple and much more complex: simpler in that it is unified, thus obviating any need to account for *downward causation*, *epiphenomenalism* or other metaphysical atrocities. More complex, because the actual world that arises by the playfulness of the metaphysical waves whipping in the storms of creation can be infinitely divided up into different description levels, and still fail to be captured completely.

Now, what about emergence, a term which is resurfacing in current debate and reaching inflationary levels? It is, in fact, a non explicatory concept – Yudkowsky would call it a semantic stop sign (Yudkowsky 2007f).

The descriptional concept of **epistemological** emergence is not objectionable – that is just asserting the above mentioned fact that for understanding different domains we have to employ different description languages due to our cognitive limitations. To describe individuals or societies we will always need psychology or sociology in the future: the essential thing is to recognize that these sciences do not introduce new levels; and also that the "physical" is not actually the "real" level; it is just another method of description. The seeming priority physics enjoys is because it is more fine-grained, and thus can explain deviations from regularities discovered at a more coarse-grained look at reality.

The way abstraction works independent of ontological concerns can easily be visualized with a computer system: if we want to program a word processor, we don't use assembly or machine language, but C++ or Java. For programming web pages, we use PHP, Python, we present in HTML etc... Abstractions can be seen as levels of organizing causal forces to get large-scale work done. Causal organization has been historically anticipated by other humans (compiler programmer, keyboard designer, OS and driver programmers, the coding of tables for keystrokes etc). But in the

end – in the actual, real-world computer, the same electrons[143] get shoved around, no matter which language is used to direct their movement. The programming languages have been developed to abstract certain commands; and they work because of the intellectual power invested into the compiler translating one language into another, all the way down to machine code. This even transfers to intentional behavior: if we are playing a computer game against an AI, we do not try to calculate the machine code; we watch how the AI behaves, and compete against it in the way we model the AI.

So, the hierarchy of abstractions we encounter in scientific investigation, but also in daily life, are only epistemological in nature. There are no ontological levels. From this also follows a devastating blow against functionalism; but more on that below. A scholarly account of the issues can be found in Floridi (2008a), who also favors dismissal of ontological levelism, but advocates the use of *epistemological* levelism.

There is also a kind of ontological emergence, but different from how it is usually conceptualized: the traditional way to picture ontological emergence is to suppose that when certain conditions are met, a new level "arises". In the conception proposed here, ontological emergence is simply the triviality that aggregates will behave differently than "isolated" matter. I have put "isolated" in parentheses because there is of course no truly isolated matter. A better way to put it would be to contrast highly interacting matter and matter which does not interact, or only weakly[144]. When I refer to this I would like to speak about the *graininess* of reality. We can zoom into the details, and look at reality at a very fine-grained level, or zoom out, and look at it in a coarse-grained way. Again, the *zooming* imagery avoids thinking of hierarchies; when one zooms into and out of a picture, the picture itself is the same, only viewed at different degrees of detail. And when we zoom out, we see that micro-patterns often form astonishing macro-patterns.

Again, saying that something is emergent does not add toward understanding of the phenomena involved: it simply says: "Something has changed." Apart from that, ontological emergence is actually ubiquitous (thus underscoring the charge of triviality): water could be said to emerge from the interaction of H_2O molecules; weather emerges from large-scale planetary matter interactions; humans emerge from aggregations of carbon, water, and a multitude of other elements; civilization emerges from an aggregation of humans with common goals; and, to jump back to the ultra-small,

143 Electrons equally do not enjoy a priori pride of place. But we have to name some aspect when speaking about particulars, because to speak about them we have to *individuate*.

144Not intended here is the physical technical term of "weak interaction".

in quantum field theory, the very particles which seem to account for rock-bottom grounding can be said to only "emerge" from field excitations[145].

I think we should throw the concept of "emergence" into the philosophical trash can and live without it; if inevitable, we should only use weak emergence (Bedau 1997), which is quite acceptable as it does not introduce additional ontological levels. We should recognize that reality behaves in many complex ways, a feature that is studied by the aptly named new paradigm of *complex systems science* (Heylighen 2001; Heylighen, Cilliers & Gershenson 2007), and here especially *complex adaptive systems* (CAS). Complexity science is about constructing models which show how coarse grained regularities arise through fine-grained interactions.

What is a CAS? A definition is given by Holland (1995, p. 10f):

He proposes four properties and three mechanisms that are common to all CAS.

The four properties are:

- Aggregation (of similar elements)
- Nonlinearity (that is, behavior is not simply the weighted sum of inputs)
- Flows (multiplying and recycling effects known from economics)
- Diversity (product of progressive adaptations)

An the three mechanisms are:

- Tagging (used to manipulate symmetries; visualize a rallying flag)
- Internal Models (agents predict and anticipate and behave accordingly)
- Building Blocks (decomposition of features to enable powerful internal models)

The tool of mathematics, and now computational modeling, so successful in the physical sciences, is becoming increasingly relevant to the social science via CAS and agent-based models and underlying theories (Epstein & Axtell 1996; Turchin 2003; Grimm & Railsback 2005; Epstein 2006; Newman, Barabasi & Watts 2006; Turchin 2008). A seminal work was Schelling (1978) who studied human segregation. This area is vital to understanding human complex systems, because the world we humans live in is largely dictated by "emergent" phenomena: cycles of economy, peace and war, demographical cycles, norms and the mores regulating our daily lives. We humans are more dominated by the emergent macro structure of our society than most other animals (with

145See Kuhlmann (2006) for an introduction.

exception of ants, bees and other hive ecologies): our culture, our ideas, and their rapid dissemination through modern technologies, impact our lives in ways unimaginable to previous generations[146].

3.1.3 An Eclectic Structural Realism

> No Water, No Moon.
>
> When the nun Chiyono studied Zen under Bukko of Engaku she was unable to attain the fruits of meditation for a long time. At last one moonlit night she was carrying water in an old pail bound with bamboo. The bamboo broke and the bottom fell out of the pail, and at that moment Chiyono was set free!
>
> From Shaseki-shu (Collection of Stone and Sand), 29[147]

> Structural realism, as a position in the metaphysics of science that is a form of scientific realism, is committed to causal structures. The metaphysics of causal structures is supported by physics, and it can provide for a complete and coherent view of the world that includes all domains of empirical science. (Esfeld 2009a)

Now is the time to first present the core metaphysical claims of this thesis: the commitment to an eclectic form of *causal structural realism*. The world is, at the finest level of detail, causal/dispositional *and* qualitative in nature. While relation is all that we can know "from the outside" – that is, physics tells us how properties *relate*, we also have a direct inside view – what it feels like to be a relational structure; that is what we call, in sufficiently complex arrangement of material entities, mind. The view is monist – mind and matter are seen to be the same thing, only viewed from different perspectives, and, because each dispositional property is also qualitative the view leads to panpsychism. I prefer to call it panqualicism, because this word is more apt. The monist and panqualicist aspect will be explored in section 3.2. In this section, it is time to concentrate on structures, dispositions and relations. But first a small detour into the debate on scientific realism.

Of what nature is our knowledge? The attitudes taken to our scientific theories in the literature are diverse, too diverse to consider them all. Prominent are Van Fraassen's anti-realist *constructive*

146 An introductory work addressing these issues is Mainzer (1997).
147 http://www.ashidakim.com/zenkoans/29nowaternomoon.html

empiricism (Fraassen 1980) which seems to me to collapse rather rapidly due to the untenable observable/unobservable distinction; but as always, arguments go both ways[148]. Unobservable[149] entities often get center stage, but I think this is a red herring. Why should unobservability be a property conferring special epistemic status to entities? Appealing to the abilities of a "normal human being" begs the question of what is normal and may lead into dubious ethical argumentation. Human enhancement – through technological means, be they mechanical, biomolecular, nanotechonological or otherwise in nature – will further erode any attempt to define a "normal" epistemic agent. Agents have different epistemic abilities, and different abilities will lead to different definitions of what an agent will consider as observable or unobservable. Thus, it can't be a serious criterion for differentiation in a serious epistemology.

On the other end of the spectrum of the debate are more traditional scientific realisms (Psillos 1999) and candidates which everyone can live with like Fine's natural ontological attitude (Fine 1984).

But the most serious current contender to the whole debate is, I believe, structural realism. What does structural realism claim? Knowledge, theories, models – in the end they boil down to *relations between the entities* to which they refer, and relations constitute structures. Words which do not refer to relations or concepts are empty. Scientific theories are mature when they contain mathematics describing features in the world correctly; that is, the theory is useful for prediction. Every physical analysis considers the world in terms of relations; mathematical formulas refer to physical entities and say how the referred entities vary (or stay equal) in relation to one another or with time (where time can be seen as the aggregate over the relations with all other things in the universe).

Let us have a look, for the moment, at a red Ferrari[150]. We move towards the car and touch it – the owner is nowhere to be seen. The car seems quite solid. But that the car has tangibility for us is a relational fact: both the car and we humans are "electromagnetic" entities, made up of protons/neutrons and electrons; therefore participating in the electromagnetic interaction. The neutrino, on the other hand, even though we can't touch it because it (roughly) doesn't share the same interaction space with us, is no less real. When we start to inspect tangible "things" more

148 The problem with philosophy is indeed that everything is *arguable.* To decide what is a good and what is a bad argument already seems to presuppose some stance taken on the world; we are confronted by the hermeneutic spiral.

149 When I write about observable/unobservable, this is not intended to exclude the other senses. Vision is only a paradigmatic case.

150 Insert your favorite cool car here.

closely, they disappear. The atom is largely empty, the protons and neutrons are actually composed of quarks held together by gluons and so on. In quantum field theory, particles, the very basic matter-stuff of the universe, are modeled as field excitations; or irreducible ray representations of groups. Everywhere we look we are confronted by a relational structure; indeed, we can only interact with things that stand in relation to us. Interaction depends on *relational* properties.

Structural realism draws the consequence: in regard to external things, only relation can be *known*, nothing else. Structural realism can account for the increase of knowledge even after drastic theory changes (Worrall 1989), which motivated its development in the first place:

> Perhaps, if we are to believe that the mathematical structure of theories is what is important, then [...] we need a different semantics of theories: one that addresses the representative role of mathematics directly. The advantage of adopting such a view is that we would then be content with the continuity of mathematical structure that is found even between theories that differ radically if taken realistically, and so would not be confounded by theory change. (Ladyman 1998).

Structural realism can either be conceived in an epistemic[151] variant or an ontic variant. Epistemic structural realism concentrates on what can be *known* of the world – its structure – but that there is something beyond structure which is unknowable. The possibility of some unknowable noumenal reality (Platonic Ideas, Kant's Thing in Itself) forever hidden from prying eyes remains. Ontic structural realism asserts that the world actually *is* structure.

I advocate a middle way: that structure, while being indeed all that what we can know about external things, captures only a "part" of reality: the other part being the *inside view* of structures (the qualitative side), and thus quite knowable, from the first person perspective.

What does ontic structural realism (OSR) say about the world?

> Ontic structural realists argue that what we have learned from contemporary physics is that the nature of space, time and matter are not compatible with standard metaphysical views about the ontological relationship between individuals, intrinsic properties and relations. On the broadest construal OSR is any form of structural realism based on an ontological or metaphysical thesis that inflates the ontological

151 Purely empirical structural realism has some problems (Ainsworth 2009), which need not concern us because it is not the position adopted here.

> priority of structure and relations. The attempt to make this precise splinters OSR into different forms …(Ladyman 2007)

Conceiving of the universe as essentially structural in nature addresses the conundrum raised by many a thinker, most notably by Wigner (1960), pondering the "The Unreasonable Effectiveness of Mathematics in the Natural Sciences".

Structural realism tackles this question head on by remodeling our conception of the world, moving away from "naive" object conceptions of the world and incorporating a physically more complex picture, namely, structures of quantum entanglement and spacetime metrics:

> Its [OSR] main motivation is to develop a tenable version of scientific realism in form of an ontology that meets the challenges of modern physics, giving an account of entanglement in quantum physics and of spacetime in the theory of general relativity. The claim is that there are structures of entanglement instead of objects with an intrinsic identity in the domain of quantum physics […]and metrical structures, which include the gravitational energy, instead of spacetime points with an intrinsic identity in the domain of the theory of general relativity[…].Esfeld (2009a)

Everybody who denies the aptness of mathematics to describe reality is welcome to present more plausible alternatives for the effectiveness of mathematics. In this sense, even if structural realism were wrong, it would be a good catalyst for further philosophical inquiry. Wigner's question remains most pressing for the traditional Platonist, who is at a loss to describe how physical brains of mathematicians should come to know about acausal, atemporal abstract objects; and for the working empirical physicist, who is using a tool which "magically" and inexplicably – simply – works.

But there is a problem with ontic structural realism. When we *refer* in physical theories, we see that the *referents* obey these relations. We seem to need *relata* in addition to the relations. There is an eliminative ontic structural realism that says that the *relata* are only nodes in structures; the nodes either being "emptinesses" or there being structures "all the way down"[152]. This is a bit mystical. There are, of course, responses[153]:

> a more moderate version of ontic structural realism has recently been developed in reply to that objection, proposing that physical

152 For an intriguing conception, see Dipert (1997), who models the world as a graph.
153 See also Esfeld & Lam (2008).

> structures are networks of concrete, qualitative physical relations among objects that are nothing but what stands in these relations, that is, do not possess an intrinsic identity over and above the relations in which they stand (Esfeld 2009a).

I opt for a slightly different path – grounding reality in "powerful qualities" (more on that below); as to how these approaches differ, there is already strife in the literature (Heil 2006; Lam 2006). Differences are slight, and in danger of becoming squabbles over words; so I will not pursue these matters here.

A third interesting version of structural realism which is close to my position is Luciano Floridi's *Informational Structural Realism (ISR)* (Floridi 2008b), which, it should be stressed, is not to be equated in any way with the conception of the universe as a Turing machine, a view which he rejects:

> Digital vs. analogue is a Boolean dichotomy typical of our computational paradigm, but digital and analogue are only "modes of presentation" of Being (to paraphrase Kant), that is, ways in which reality is experienced or conceptualised by an epistemic agent at a given level of abstraction. A preferable alternative is provided by an informational approach to structural realism, according to which knowledge of the world is knowledge of its structures. The most reasonable ontological commitment turns out to be in favour of an interpretation of reality as the totality of structures dynamically interacting with each other. (Floridi 2009)

The good thing about ISR is the natural accommodation of Smith's take on ontology: reality is always more than the computational description at one level or another, although patterns are to be found throughout. Now in addition to this relational description of reality as "structures" imbued with more or less ontological significance[154], we need also account for the dynamics: and here is where dispositions come in.

What we need to add to make the universe work – accounting for the dynamics – is *causation* and *dispositions,* causality being dependent on dispositions. We need a connection from structures to causal effectiveness[155].

154 The traditional ontological commitment is foreign to the way ontology is done here. We commit to *ways of doing* ontology, not a specific ontology. There will be an exception: fine-grained dispositions – more on that below.

155 That is also the difference between physical structures and mathematical structures, the latter which are only insofar causally efficacious as they are described on a different level (brain processes constituting the current thinking of a mathematical structure etc).

> Ontic structural realism is the view that structures are what is real in the first place in the domain of fundamental physics. The structures are usually conceived as including a primitive modality. However, it has not been spelled out as yet what exactly that modality amounts to. [...] the fundamental physical structures possess a causal essence, being powers. Esfeld (2009a)

We adopt causal structural realism, which, alas, comes in many versions already (Shoemaker 1980; Hawthorne 2001; Mumford 2004; Bird 2007; Chakravartty 2007). I will mainly follow Heil (2003); Esfeld (2009b,a). What is a disposition[156]? Dispositional *properties* are usually juxtaposed to *categorical properties* (where I will call categorical properties *qualitative* ones, following Heil). A classical dispositional property mentioned in the literature would be *fragility* – the disposition to break upon rough handling. A categorical property discussed in the literature would be triangularity – which I do not endorse, but more on that below.

One definition of dispositions would go like this:

> **Entailment.** *F* expresses a disposition iff there are an associated manifestation and conditions of manifestation such that, necessarily, an object is *F* only if the object would produce the manifestation if it were in the conditions of manifestation. (Fara 2006)

I am unhappy with the word entailment, because it clearly shows the heritage of the definition out of analytic philosophy of language, where dispositions are analyzed in terms corresponding subjunctive conditionals[157]. Well, so be it. Mellor then goes on to show that this definition shows that *all properties* are dispositional (Mellor 1974). We need not go into the details of the argument, which is quite direct. A simple argument will suffice: if categorical properties do not have manifestations, how should we come to know of them? But this does not mean that a qualitative/categorical side of things is denied; indeed, the position I take is that dispositions and qualitative properties are actually the *selfsame*. I follow Heil's identity theory:

> If P is an intrinsic property of a concrete object, P is simultaneously dispositional and qualitative; P's dispositionality and qualitativity are not aspects or properties of P; P's

156 Another name for a disposition is a *power*; I will use the two names interchangeably.

157 We should note that counterfactuals enter the situation here: this is metaphysically intriguing, as counterfactuals also crop up in quantum mechanics. The subject warrants scrutiny, although I can't pursue this matter here. For an introduction, see Dorato (2006). Of course, as Heil says:

> Conditionals provide a defeasible, rough-and-ready way to pick out dispositions, not a reductive analysis. (Heil 2003, p. 196)

> dispositionality, P_d, is P's qualitativity, P_q, and each of these is P: $P_d = P_q = P$.(Heil 2003, p. 111)

So, an entity has properties; these properties are powerful *and* qualitative in nature[158]. A power is the ability to be causally effective, to exert influence (we will look at the qualitative side later on). One remark on the relationship between properties and substances. One could say the following:

> Substances are bearers of properties; properties are ways substances are. (Heil 2005a)

Substances in Heil's conception (he also calls them objects) are not to be conceived of as simple indivisible entities ("billiard balls"), but rather as basic explanatory entities. Substance and properties are intertwined in a sort of way that makes it irresponsible to speak of substance without properties or properties without substance.

One more thing – the properties are conceived of as *modes*, in line with particularity:

> Suppose the world comprised objects distributed about in space–time. These objects possess properties in virtue of which they behave, or would behave, in particular ways. Objects' properties are not universals or instances of universals, they are what, in Locke's day, were called *modes*; what others [...] call tropes. Modes endow their possessors with particular qualities. Modes are qualitative. But modes are, as well, powers. Think of modes as powerful qualities. Objects in the envisaged world are similar by virtue of their possession of similar modes. Modes are similar – or not – *tout court*. Similar objects will behave similarly in similar circumstances because a condition on their being similar is their possessing similar modes; and mode similarity is simultaneously qualitative and dispositional.
>
> Complex objects in the world we are imagining are made up of simpler objects. Characteristics of complex objects are unproblematically fixed by characteristics of their constituent parts and relations these parts bear to one another. From this distance we cannot tell whether our imagined world is granular – consisting of distinct objects arranged in space – or unified – consisting of one or more fields with distributed 'thickenings' corresponding to more familiar objects. [Footnote 4: If this is right, we are in no position to

158 The distinction "categorical" and "dispositional" continues to apply to predicates (linguistic entities).

> ascertain a priori whether familiar objects man beings, trees, rocks, electrons) are, at bottom, substances or modes.] (Heil 2005a)

The rejection of universals is not as problematic as it may initially seem. Universals depend on strict identity; modes exhibit irreducible similarities. One can dump Platonic universals without getting into metaphysical trouble.

In causal structural realism, we will take the structures and relations as actually being constituted by both dispositional and qualitative properties – properties being the "primitives" of our ontology. For that, we need reduce causality to dispositions, and that task is of some complexity and would probably require a thesis of its own. A few guiding ideas must suffice here. First of all, the Humean conception of things is turned around – laws are not generalizations of regularity found in contingent patterns in the world, but are actually derived – necessarily – from a dispositionalist account:

> Dispositionalism, then, radically overhauls the relationship between dispositions and laws. Instead of all dispositions depending upon contingent laws, at least some important dispositions are irreducible features of the natural properties. Moreover, the sorts of laws which are typically used to ground dispositions are — according to the dispositionalist — necessary laws: corollaries of something essential in the natural properties. (Handfield 2009, p. 17)

Causation is also neither the mere observance of regularity or driven by contingent law, but the real thing: causal relations are *manifestations* of dispositions. For the dispositionalist, an analysis of causation will:

> terminate in the essentially power-conferring natures of the basic properties. (Handfield 2009, p. 19)

Ladyman says that the idea of causal structuralism is that:

> causal relations that properties bear to other properties exhaust their natures" (Ladyman 2007).

That is not quite correct – at least not in the present conception. It is rather both the dispositional *and* the qualitative properties, which come part and parcel, which exhaust their natures. The dispositional is more encompassing than the causal: the causal requires a manifestation. The idea that all properties are dispositional and not categorical (that is, qualitative without being effective)

means that there can't be a difference in the world in properties *without* there also being a difference in *detectable* ways. This is important later on in the philosophy of mind. To be blunt:

All real properties have causal influence.

Heil summarizes his view on dispositions in this manner:

> (1) Dispositions are actual, not merely possible features of objects.[159]
>
> (2) Dispositions are intrinsic properties of objects possessing them.[160]
>
> (3) The nature of dispositions is not wholly revealed via a reductive conditional analysis.
>
> (4) Dispositionality is not a contingent feature of the world.[161]
>
> (5) Every intrinsic property of a concrete object is dispositional . .
>
> (6) ...but not purely dispositional.
>
> (7) Dispositions are not 'higher-level' properties.[162]
>
> (8) The manifestation of a disposition is a manifestation of reciprocal disposition partners.
>
> (9) One and the same disposition can manifest itself differently with different reciprocal disposition partners.
>
> (Heil 2005b)

I have said above that no ontological stratification should enjoy a priori pride of place. But there is the fact that physics is more robust than other sciences, and this needs accounting. It can be accounted for in the dispositionalist picture of things. Remember that we have negated a levelist conception of reality: that is, when we look at ever smaller systems in physics, we do not "go down

159 Manifestations, on the other hand, need not be actual.
160 Heil distances himself from a relational account; the exact differences between Heil's intrinsic dispositions and divers relational accounts would need to be analyzed, but that would be taking things too far here.
161 This is the big anti-Humean step. The most important step to solve the mind-body problem.
162 This is the part where functionalism dies.

levels", but zoom in on the microcausal structure of reality. We can then ascribe a dispositionalist/qualitative structure to the universe at the finest scale. This does not make it more real or less real than other "levels" of description. It is simply the scale where the dispositional account bottoms out.

The universe is of a piece[163]. The development of ever-more fine-grained theories should be considered analogously to slicing – parts of reality are "sliced" off, looked at in ever more detail, with increasingly difficult to achieve causal isolation[164]. "Deeper" theories do not correspond to the uncovering of hidden layers. With these metaphysical commitments, for creatures living inside reality, the best working philosophy to adopt is materialism or hylorealism. This is *not,* I have hoped to make clear, to be confused with a picture of minuscule billiard balls bopping around in the spacetime box.

Materialism – at least the Bungean version which I assume as starting point – asserts that everything that exists exists in virtue of its *causal interaction via energy*. I like to think of energy as a currency of transaction or interaction; the more you have, the more you can interact, although if you distinguish structural causes and triggering causes, you may accomplish quite a lot with little triggering energy. Underlying this picture is of course one of fundamental powers; the world is conceived of as a *power net*.

> A world containing properties with built in powers would be one in which objects are embedded in what Martin calls a 'power net' (Martin 1993a)[(Martin 1993)]. An object's behavior, then, would be the result of a confluence of influences grounded in the object's properties and the properties of other objects that influence it and are in turn influenced by it. (Heil 2003, p. 95)

Another core aspect of materialism is the tenet that particularity is what counts in the real world and not abstraction, which is a "mere" tool: Smith's IDF-reality. Which leads us to the next section.

3.1.4 A Mathematical Universe?

> As a mathematical discipline travels far from its empirical source, or still more, if it is a second and third generation only indirectly inspired by ideas coming from "reality" it is beset with

163 Mutually isolated regions notwithstanding; the causal structure can be pictured like a partial ordering, not a total ordering; but the "universal graph" is not separated into subgraphs.

164 Maybe the world is infinitely divisible, but the more you zoom in, the more boring things get. An indication for this could be recent work in quantum gravity Loll (2007).

> very grave dangers. It becomes more and more purely aestheticizing, more and more purely l'art pour l'art. This need not be bad, if the field is surrounded by correlated subjects, which still have closer empirical connections, or if the discipline is under the influence of men with an exceptionally well-developed taste. But there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities. In other words, at a great distance from its empirical source, or after much "abstract" inbreeding, a mathematical subject is in danger of degeneration. At the inception the style is usually classical; when it shows signs of becoming baroque, then the danger signal is up. (Neumann 1947)

If we consider our universe as structural, relational – can we simply say that it is mathematical, as Tegmark (2007) suggest? I believe not, for a number of reasons. An obvious one is that mathematics is full of wild abstraction, and gives us no handle at all *which* mathematics describe the physical universe around us. So, the term *ontic structural realism* used above is already much less charged and thus preferable. Above, I endorse causal structural realism: the relations and structures are themselves constituted by a more primitive modality, one that is at the same time qualitative in nature[165]. But the position is still sufficiently close to ontic structural realism to delineate structuralism from a purely mathematical conception of the universe.

Some remarks on the nature of mathematics are in order before we can proceed: no sophisticated philosophy of mathematics, but rather simple considerations of how mathematics relates to language and thought generally[166]. Snow, in his celebrated book "The Two Cultures" sketched the rift between the sciences and the humanities, which can be expressed in a nutshell with the following sentence:

> A good many times I have been present at gatherings of people who, by the standards of the traditional culture, are thought highly educated and who have with considerable gusto been expressing their incredulity at the illiteracy of scientists. Once or twice I have been provoked and have asked the company how many of them could describe the Second Law of Thermodynamics. The response was cold: it was also

165 But then again – maybe the modal view is only the inside, timed, view of timeless structures, which would then be more fundamental. Who knows.

166 The thesis also wants to address an non-mathematically literate audience.

> negative. Yet I was asking something which is the scientific equivalent of: Have you read a work of Shakespeare's? (Snow 1959, p. 14-15)[167]

This rift has sprung up with the sophistication of mathematics necessary to comprehend modern science. Some people like to stay adrift in fuzzy concepts, whereas mathematics is the paragon of exactness. In a sense mathematics is the most precise means of communication – if you can make something explicit in mathematics, then you can communicate it – provided the receiver can make sense of your axioms. As mathematics can be reduced to set theory, and set theory is grounded in basic embodied metaphors accessible to all humans (Lakoff & Nunez 2000) we have wonderful precision. The problem with everyday language is that it can trigger arbitrary associations in different humans (I guess this also happens to a reduced extent in mathematical communication, as when people unconsciously hold different intuitions about the infinite).

Many people drawn to the humanities do not recognize the importance of mathematics because, in their eyes, it is inapplicable abstract nonsense, not realizing that modern society depends largely on *applied mathematics*. The negligent attitude towards mathematics or even the natural sciences as a whole is probably the result of bad eduction. When a child is shown a table and told that this is to be named a "table", it knows what a table is. When the same child, a little older, sees "F = ma" on a blackboard, it can be difficult to relate this to the real world, especially with a bad teacher. From these early moments stems the misconception that mathematics is strange and otherworldly.

Mathematics is about ideas – precise ideas – and they are captured in a sometimes clear and sometimes an obscure symbolism. Mathematics is not about the abstract symbols themselves, but about relations, patterns, structure; mathematics is about concepts, which may or may not fit reality. The symbols themselves say nothing; it is in our minds that mathematics comes to life. Every human who can think can also think about mathematics; the better her love for precision, the more successful in discerning mathematical ideas. Mathematics means the renunciation of vagueness and the abandonment of superfluous assumptions. Mathematical precision comes from it's rigor – axioms, definitions, theorems, finite rules of inference; where this criteria of precision are not met, we can still continue to suspect, but not be certain.

The syntactic aspect of mathematics is the aspect of rule following. But it constitutes no arcane language; it is no different than English or German; only the concepts which are referred to are not

167 A problem which has not gone away Nature (2009).

so anthropocentric, more of the nature of the world actually. When higher mathematics is involved, the concepts become elusive, difficult to grasp; it takes an effort to understand. Mathematics is successful symbolically not because of some strange magic, but because the notation enables one to offload thinking into the environment (paper, pencil, blackboard, computer). A well chosen notation is the essence of the *symbolics* of mathematics (but not of the *conceptions*; a superior mind could perform our mathematics in her mind without the help of notation). Good mathematical notation enables humans to perform simple cognitive algorithms for arriving at theorems of formal systems. In this one way mathematics is different from natural language: that for higher mathematics the offloading of thinking into the notation and environment can't be forgone. It is a gradual difference; a difference that comes about because of the complexity of the concepts involved or their removal from anthropocentric matters, not because of a strict delimitation of subject matters.

For instance, the decimal system lends itself well to do arithmetic, because of its positional structure, as opposed to the roman system which was very difficult to calculate with. Compare trying to add XXXIII + IIC; and then try adding 33 + 98. The latter is not simpler because the numbers look more familiar – well, OK, that is part of the reason; but it's not the real reason. No, the decimal system is simpler because one has to *learn less rules* to perform addition. The rules are exhausted by learning the addition table for the numbers 0 to 9 and how to handle the carry. No system of similar simplicity exists for roman numerals. But in both versions we simply have to add two numbers – conceptually the task is the same, even if the notation will spawn very different brain processes.

So, the point is this: mathematics is not about the shuffling around of abstract symbols (although it is in a strictly formalist interpretation) but rather the building of elaborate conceptual structures, which can only be handled by our limited minds if supplemented with an algorithmic notation which captures the structure of the theory – but to say more would be to venture too deeply into the philosophy of mathematics.

An especially problematic case of conceptualization is the infinite. Nowhere do we really have access to the infinite. We never see a *real number* all at once. We can't have a concept of the infinite, because we can't see it anywhere. We can't live it, experience it; we do not have direct access to the Platonic concept of Infinity, which is a fantasy. Thought is matter organized in certain ways, due to lawful correlations with the environment. Concepts, to repeat, are formed in derivation of the world (concept empiricism) but with phylogenetic and embodied grounding (Kant's a priori)

which have evolved in relation to this world. Concepts not derived from the world directly can be derived by processes of metaphor and analogy, but must be seen as that, and not more. Lakoff & Nunez (2000, p. 158f) describe how the conception of the infinite arises in finite cognitive entities. They describe the BMI, the basic metaphor of infinity, which goes as follows:

- there is an initial state and
- an iterative process;
- a resultant state after each iteration and
- a final resultant state.
- Entailment: There is a final resultant state that is unique and follows every nonfinal state.

This conception is taken out of a *source domain*, and applied to a *target domain*, where one can forgo the real world observation that all iterations are *finite* and conceptualize an unending iteration with a final resultant state and arrive at the concept of actual infinity. So, even for "infinity", there is no need to go beyond embodied natural cognitive processes; we do not need a Platonic realm from which to draw our intuitions, but can make do with tracking the natural world.

What are the bounds of mathematical thinking? That depends on the axioms we choose. An interesting questions is why humans who do not have a strict axiomatic picture of mathematics nevertheless perform the same operations, think in the same idea-space? Because we implicitly model the physical reality around us (Lakoff & Nunez 2000). Of course, there does seem to be something in mathematics that has a recalcitrance beyond the physical – otherwise it would not be intelligible why, apart from Euclidean geometry, elliptical or hyperbolic geometry are alternative conceptualizations, not immediately grounded in our geometrical environmental experience. But we have to be careful: these geometries are nevertheless grounded in more basic intuitions, which are themselves embodied.

I would say that, as mentioned in section 2.9.5, the concept of consistency also derives from our evolutionary brain architecture[168], and we construe those mathematical structures which are consistent from our perspective. Everything that has been located in Platonia should be relocated to *cognitive architecture*; an architecture which evolved to track structure in the world, thereby making the previous eternal Platonic truths not arbitrary but more *in touch with the material world*. The project of naturalizing mathematics is well underway:

168If somebody says that something does *not* have an evolutionary origin, I am always very interested from where else it should come from.

> Mathematicians frequently evoke their "intuition" when they are able to quickly and automatically solve a problem, with little introspection into their insight. Cognitive neuroscience research shows that mathematical intuition is a valid concept that can be studied in the laboratory in reduced paradigms, and that relates to the availability of "core knowledge" associated with evolutionarily ancient and specialized cerebral subsystems. As an illustration, I discuss the case of elementary arithmetic. Intuitions of numbers and their elementary transformations by addition and subtraction are present in all human cultures. They relate to a brain system, located in the intraparietal sulcus of both hemispheres, which extracts numerosity of sets and, in educated adults, maps back and forth between numerical symbols and the corresponding quantities. This system is available to animal species and to preverbal human infants. (Dehaene 2009)

The idea of the "triangle" or the "sphere" – are abstractions of neural nets – confronted with a messy world which nevertheless contains approximations to such "ideal" objects. The possibility for abstraction should be accepted as a feature inherent in the material world. Abstraction is not mysterious – it is the *leaving away* of aspects in a representation. Basic concepts, such as motion, hardness etc are directly grounded via sensorimotor experience; upon these we build ever more abstract concepts, leading to hierarchies of abstraction, which ideally, retain some aspect of reality so that they are usable in inference tasks or even better, still refer in a sensible way.

But we also need to take care with the structures we create in mathematics. A mathematical theory becomes physico-mathematical when interpreted in a physical way. If a mathematical theory captures some structures of reality in a way that lends itself to prediction, then we can say that it is about relations persisting in the world. There are of course also purely abstract theories, which have no physical referents anymore. And that is one reason why it is unwise to conflate OSR with a "mathematical" universe; because in the mathematical universe, the difference between physical mathematical structures and "fantastic" mathematical constructions gets lost[169]. To clarify, I repeat the above phrased a bit differently: Thoughts are the inside view of certain complex structures, and

169It is also important to distinguish ontic structural realism from a form of Platonism which picks out only *certain* structures as real. In Platonism, reality is a mere shadow of the realm of perfection – the realm of forms. Ontic structuralism is in this sense very contrary to traditional Platonism: it does not say that existence is a mere shadow, an imperfect copy of some eternal object "out there" in some inaccessible realm, but in fact the r*elations (qualitative/dispositional properties) are all there is.* No shadows, no caves, no torches; only the real things themselves.

insofar thoughts are reflections of other structures they can be called knowledge. Thoughts which do not correspond to our immediate reality may correspond to other parts of reality, or to none at all – then they are fantasy.

Mathematics – and the functional relations described via mathematics – does not tell us how to interpret itself, how to relate it to the real world. For this, we need other relations lawfully relating the mathematics to the real world. One could all do it in a mathematical framework, but only in a Platonic one, and it does not solve the problems *which* relations hold, so we might as well go back to materialism. Materialism can then be seen as the philosophical position that *not all* mathematical statements are instantiated and that there is *no abstract realm* where mathematical sentences exist independent of material brains instantiating them. The recalcitrance of mathematics – the "kicking-back" aspect comes from our commitment to the basic consistency learned from the physical world.

The philosophy of mathematics endorsed here is fictionalism, not Platonism[170]. Fictionalism can account for the fact that some mathematics tracks reality, and some mathematics does not, just as some stories are true and other stories are fantasies. The fictionalist account merges well with ISR and the conception of IDF-Reality; the relation to OSR is less clear, though given Esfeld's elaborations that OSR should subscribe to powerful structures (Esfeld 2009a), the delineation to mathematics is recovered and we can reject Platonism, given that fictionalism offers a coherent account of the role of mathematics in the empirical sciences:

> Nominalistic scientific realism is different from standard scientific realism. The latter entails that our empirical theories are strictly true, and fictionalists cannot make this claim, because that would commit them to the existence of mathematical objects. Nonetheless, nominalistic scientific realism is a genuinely realistic view; for if it is correct - i.e., if there does obtain a set of purely physical facts of the sort needed to make empirical science true - then even if there are no such things as mathematical objects and, hence, our empirical theories are (strictly speaking) not true, the physical world is nevertheless just the way empirical science makes it out to be. So this is, indeed, a kind of scientific realism. What all of this shows is that fictionalism is consistent with the actual role that mathematics plays in empirical science, whether that role is indispensable or not. It simply doesn't matter (in the present

170 Balaguer argues, interestingly enough, that what he calls *full-blooded Platonism* and *fictionalism* are the only two tenable philosophies of mathematics, and that the two are in fact indistinguishable (Balaguer 1998; Balaguer 2009). This then lets us reject Platonism for considerations residing outside the philosophy of mathematics.

> context) whether mathematics is indispensable to empirical science, because even if it is, the picture that empirical science paints of the physical world could still be essentially accurate, even if there are no such things as mathematical objects. (Balaguer 2009, p. 86)

3.1.5 Functionalism Negated

> For if one defines the operation of sawing as being a certain kind of dividing, then this cannot come about unless the saw has teeth of a certain kind; and these cannot be unless it is of iron. (Aristotle PHYS)

The philosophy above is very different from functionalism, which will occupy us in detail in section 4.4.3, but some remarks are in order here. What makes the present philosophy different from functionalism is the rejection of a leveled view of reality, and the problematization of the concept of multiple realizability. For instance, consider the traditional Picture Theory conception of pain as a higher-level property, which is then multiply realizable (and for which must pay the price of epiphenomenalism or dualism or other unpalatable -isms).

Not so in the present view: if different organisms are in pain, that is not because the function/predicate pain is realized in different ways in them, but rather their dispositional/qualitative makeup makes them feel pain; and if they have a similar makeup, they will have similar qualitative states.

> I have suggested that the felt need for higher-level properties in such cases is an artefact of the Picture Theory. A simpler explanation of the phenomena beloved by advocates of multiple realizability is that predicates taken to designate so-called higher-level properties are in fact satisfied by members of families of similar properties. These similar properties are just those properties standardly taken to be realizers of the higher-level properties. The pain predicate applies or would apply to creatures in virtue of those creatures' possession of any of a possibly open-ended family of similar properties. These properties fall under the pain predicate because they are relevantly similar: similar, perhaps in the contribution they make to the dispositional and qualitative character of their possessors' states of mind. (Heil 2003, p. 153)

This aside, the case for multiple realizability is empirically problematic (Shapiro 2000).

In sake of terminological clarity, it is also important to distinguish functions from mechanisms; there is a one-to-many relationship between function and mechanism (Mahner & Bunge 2001). An important distinction is also to be drawn between functions as internal processes, such as mechanisms, and functions as relations, for instance, a stool functioning as a place to sit for a human being (Bunge & Mahner 2004, p. 158f).

3.1.6 Time and Space

A few *qualitative* intuitions on time and space shall be offered here, simply because these two categories are so essential for our experience that they deserve some special mention. The below expositions could be considered personal heuristics of thinking rather than a formal analysis.

"Time exists that not everything happens at once"[171], and space, if I may add, that not everything happens at the same location. Both are relational phenomena, and since Einstein we know that they are intricately related.

An intuition I have come to concerning time is the following: from relativity theory we know that there is *no universal now*, and that the invariant between two "points" in the physical universe is neither distance nor time but *spacetime distance* taken together, where the duration of time or the distance of space alone between two events are different relative to observers in different inertial frames. A good way of thinking about special relativity is to consider *everything* as moving at the constant speed c through the block universe, where there is a trade off between the movement through time and the movement through space. The faster you are going through space, the slower you will go through time, so that in the extreme case of photons moving at c through space, no time passes (fort them) as they traverse the universe.

Now let us for the moment consider the position of an observer, ourselves, for instance. Space can then be considered as the dimension over which I have control as a thinking subject. I can move left, right, up, down, front, back – three dimensions[172]. I can consciously change my relation to other objects, which continue on their trajectories through spacetime. But note that we are embedded in *local* spatial relations and only here do we find the freedom to return to locations we have visited previously. If we consider the movement of the Earth, the solar system, the galaxy, the supercluster etc, we can see that we can't control our location as exactly as we would initially think. The larger the scale of structures we consider, the more we enter the domain of time.

171 (Bourke)
172 For an interesting analysis why we may live in a (3,1)-dimensional universe see (Tegmark 1997).

Time can be considered as nothing but the motion of all other objects in the universe in relation to oneself – even some "objects" of which "I" am made of, say, the cells in my body which divide, contributing to the process of aging, the flow of the bloodstream etc.

So, in a nutshell, the conception, from a first person perspective, is as follows:

- Space encodes the concept of self moving in relation to everything else. We have partial control over our spatial whereabouts because that is just what our agency means: to be a power in the world.
- Time is the encoding of everything *outside* the self moving in relation to oneself. We can't freely explore the temporal direction because we can't control the motion of the rest of the universe[173]. This also affects systems very close to us: aging, for instance, is our body parts going through their motions (cell division etc) without our control (except for healthy living, slowing down the metabolism rate etc).

Time results from motions which I can't control, and space results from motions which I can control. So, in this sense, time and space are indeed very much alike – they just represent different points of view (self versus otherness).

Of course in the picture painted above we still need different "states" of the universe, so that the *illusion* of the *flow* of time can appear[174]. I am a bit unhappy with the term "state", because it conveys a too simplistic picture, "state" usually referring to a precisely defined situation. I think that the states of the universe are more unruly than that – what Smith (1996) calls *zest* and *spunk*, a basic playfulness of universal nature.

Anyway, the best exposition of how a timeless universe works, that is, where time appears only in the *inside view* from a purely relational universe[175] is Julian Barbour's "The End of Time" (Barbour 1999). I would just like to sympathetically nod to Barbour's program here, as I do not have the time to explore the issues.

What trouble are we getting into when assuming a timeless universe? Will pain, for instance, be eternal? No: experience, and the preconditions for experience, time and space, arise on the *inside view* – only the hypothetical, in principle unobservable outside view is timeless; it is never experienced; experience demands a succession of states and therefore time. So, pain can only be

173 Time travel is the fantasy of reversing or accelerating all motions outside the self.
174 For a review of the issue of time in physics, see Callender (Forthcoming).
175 Space, needless to say, is also conceptualized relationally here. For a detailed treatment of all the different views on these matters, see the excellent book of Dainton (2001).

experienced in a *timely* fashion. An experienced moment is never eternal, on the contrary, it is always fleeting.

From the inside, the universe is merciful; and from outside, it is nothing at all (because there is no cognitive entity to interpret the "sum" of all relations). What about bad experiences of others in the past that you must think of as real now? Adopting a timeless view, even if only in principle, may lead to empathic suffering in regard to those moments. Yudkowsky offers a placation for the troubled mind – and while it was offered for the multiverse, it works just as well for the timeless universe:

> ...horrible things happened during the 12th century, which are also beyond your ability to affect. But the 12th century is not your responsibility, because it has, as the quaint phrase goes, "already happened". I would suggest that you consider every world which is not in your future, to be part of the "generalized past".(Yudkowsky 2008g)

An exact analysis of how a timeless universe resonates with the dynamic view proposed here would need more careful investigation. But for practical matters, we need the *dynamic view* grounded in powers – the inside view. The other ruminations are only of metaphysical interest.

3.1.7 Universal Darwinism

Often times it is criticized that evolution is applied in non-biological contexts, that the analogy or metaphor is badly chosen – the differences from biological evolution to other forms of evolution are stressed. But evolution is more than a biological phenomenon: it is a meta-principle which is valid in many domains in this universe; it is a universal algorithm[176], substrate-neutral and mindless. Blackmore (1999) calls this principle *Universal Darwinism:* variation, selection and retention (heredity in biological systems) operate algorithmically[177]. Biological evolution is only a special domain where the evolutionary algorithm operates. Applied in wider contexts, there are theories that seek to explain the fine-tuning of our universe to life to evolution happening at cosmic scales – cosmological natural selection operating on universes being produced in black holes Smolin (1997). Also, the stability of the macroscopic world we experience may be the result of *quantum Darwinism* – stable states being selected by the environment Zurek (2007).

176 The word algorithm is employed in a broader context than that of computer science: the requirement of finite termination is dropped; and there is no extrinsic goal evolution wants to solve: the “goal” of replicating sufficiently often to stay in the world is implicit.

177 A detailed analysis can be found in Dennett (1995).

Evolution is not directed; but does evolution lead to more complexity? Only subprocesses of evolution do: mutation and variation, left to their own devices, will lead to more variability in the units of selection. Selection itself – a meta-effect actually, namely, the failure to replicate sufficiently often in a resource-constrained environment, thus leading to displacement – curtails complexity.

A few words on ethics are in order. As Darwinism always seems to raise specters among some people, they should be banished quickly. When Darwin's theory of evolution is invoked, the focus usually lies on natural selection, the process of competition and the survival of the fittest[178].

But that is only part of the story. Systems flourish when they are robust, that is, keep functioning under a wide variety of possibly opposing conditions. We can zoom out and view evolution as an algorithm which proliferates life as a whole (this view is possible without getting into teleological channels). The more diversity the processes of variation, mutation, recombination etc produce, the more robust the whole community of living beings will be against environmental culling. The proliferation of life profits from a large variety in the gene pool. There is no such thing as an *optimal* biological blueprint. – maybe for a certain situation and environment, which will lead to very stable designs such as the class of sharks – but not in general for all times and situations.

This line of reasoning stifles both racism in the biological sphere and totalitarianism in the memetic sphere at their roots; racism and totalitarianism are suboptimal strategies. The infosphere profits from a wide variety of theories and ideas much as humanity or life generally profits from a deep gene pool.

3.1.8 Quantum Fairy Land

What about the quantum? Quantum mechanics is routinely invoked to either justify the inapplicability of rationality to the world or to justify crackpot theories. As soon as you enter the domain of quantum mechanics you find yourself in a place of myth and fantasy, of magical creatures, powerful wizards, and, to take some poetic license, beautiful damsels. This is most certainly due to the fact that no interpretation of quantum mechanics is currently satisfying. But the results of the present thesis do not depend on one or other interpretation being correct; they solely depend on what is certainly not correct, and to elucidate this is the goal of the present section[179].

178 Or the likeliest (Whitfield 2007).

179 For the reader interested in easily accessible overviews of quantum mechanics and the problems the theory poses to classical conceptions of the world, there are a number of excellent texts available, among them Albert (1992); Rae (2004); Camejo (2006). For an excellent philosophical analysis see Putnam (2005).

A good starting point for our investigation is a paper aptly named "Quantum Mechanics: Myths and Facts" by Nikolic, who starts out to list some myths:

> ... myths include wave-particle duality, time-energy uncertainty relation, fundamental randomness, the absence of measurement-independent reality, locality of QM, nonlocality of QM, the existence of well-defined relativistic QM, the claims that quantum field theory (QFT) solves the problems of relativistic QM or that QFT is a theory of particles, as well as myths on black-hole entropy. The fact is that the existence of various theoretical and interpretational ambiguities underlying these myths does not yet allow us to accept them as proven facts. (Nikolic 2007)

Virtually everything of interest in quantum mechanics is disputed in some way or another. And indeed, paradigms accepted in one "interpretational community" are not taken seriously in another. The philosopher can only hope to extract some principles of the discourse which look like they will exhibit some stability.

Of special concern to the topic of the present work are the concepts of *fundamental randomness* and *measurement-independent reality*[180].

So, what about fundamental randomness? Nikolic writes:

> ... classical dynamics is completely deterministic. On the other hand, the usual form of QM does not say anything about actual deterministic causes that lie behind the probabilistic quantum phenomena. This fact is often used to claim that QM implies that nature is fundamentally random. Of course, if the usual form of QM is really the ultimate truth, then it is true that nature is fundamentally random. But who says that the usual form of QM really is the ultimate truth? (A serious scientist will never claim that for any current theory.) (Nikolic 2007)

It often seems to me that every opportunity at destroying a deterministic conception of the universe is seized upon by certain people, who in the process dump their scientific attitude. No theory may enjoy a priori pride of place. And just the same as a scientist will hold totally deterministic theories tentatively, she will hold theories implying randomness tentatively. From

180 The existence – or non-existence – of particles has basically been addressed in the way we plan to to ontology: particles are a commitment to certain microcausal structures dependent on a certain (robust, and therefore scientific) stratification.

whence the wish to have indeterminacy as a fundamental feature of the world? I think it is the wish to smuggle in free will or some conception of agency independent of the laws of nature. But quantum indeterminacy will not suffice to buy this. More on that in section 4.1.

In either case, the relation between quantum mechanics and classical mechanics is not as easily painted in black and white as commonly assumed; it seems as Nikolic has overlooked some myths. Let us have a brief look at the literature. The surprise starts when reading a very respected textbook on Quantum Mechanics:

> ...the quantum correlation is always stronger than the classical one, except in the trivial cases [...]. Are you surprised? If so, this is the result of having been exposed to unfounded quantum superstitions, according to which quantum theory is afflicted by more "uncertainty" than classical mechanics. Exactly the opposite is true: quantum phenomena are more disciplined than classical ones. (Peres 2002, p. 162)

Another of the "indeterministic" features of quantum mechanics is supposed to be *complementarity*:

> Complementarity is the principal impossibility to measure two or more complementary observables with arbitrary precision simultaneously (Svozil 2007)

Svozil goes on to present how complementarity can arise in a finite state automaton – the most simplistic model of a deterministic system.

> Since any finite state automaton can be simulated by a universal computer, complementarity is a feature of sufficiently complex deterministic universes as well. To put it pointedly: if the physical universe is conceived as the product of a universal computation, then complementarity is an inevitable and necessary feature of the perception of intrinsic observers. It cannot be avoided. (Svozil 1996)

But not only that quantum systems are very disciplined and not at all obnoxious to determinism, it seems that Newtonian Mechanics, the standard example for a clockwork universe, is not as well behaved as is commonly assumed[181]:

181 Norton gives mathematical proof for the claim. Note that I do not endorse Norton's anti-causal stance expressed in the cited paper. Although he sympathizes with the account given in Dowe (2007):

> In this regard, the most promising of all present views of causation is the process view of Dowe, Salmon, and others (Dowe 1997). In identifying a causal process as one that transmits a conserved quantity through a continuous spatiotemporal pathway, it seeks to answer most responsibly to the content of

> Even quite simple Newtonian systems can harbor uncaused events and ones for which the theory cannot even supply probabilities. Because of such systems, ordinary Newtonian mechanics cannot license a principle or law of causality. [...] an example of such a system fully in accord with Newtonian mechanics. It is a mass that remains at rest in a physical environment that is completely unchanging for an arbitrary amount of time—a day, a month, an eon.
>
> Then, without any external intervention or any change in the physical environment, the mass spontaneously moves off in an arbitrary direction with the theory supplying no probabilities for the time or direction of the motion. (Norton 2007, p. 22f)

Earman sums it up:

> ... in some respects determinism is a robust doctrine and is quite hard to kill, while in other respects it is fragile and requires various enabling assumptions to give it a fighting chance. [...] determinism is far from a dead issue. Whether or not ordinary non-relativistic quantum mechanics (QM) admits a viable deterministic underpinning is still a matter of debate. Less well known is the fact that in some cases QM turns out to be more deterministic than its classical counterpart. Quantum field theory (QFT) assumes determinism, at least at the classical level, in order to construct the field algebra of quantum observables. Determinism is at the heart of the cosmic censorship hypothesis, the most important unsolved issue in classical general relativity theory (GTR). And issues about the nature and status of determinism lie at the heart of key foundation issues in the search for a theory of quantum gravity. (Earman 2007, p. 1369)

So we see that the common conception – that Newton's worldview was entirely clock like, and quantum mechanics has destroyed this view, is wrong. Where does this leave us concerning the determinism/indeterminism issue? Nowhere. Earman paints an interesting picture, taking a bird's eye view on the whole of physics[182]:

> The fortunes of determinism are too complicated to admit of a summary that is both short and accurate, but roughly speaking the

our mature sciences. Insofar as the theory merely seeks to identify which processes in present science ought to be labeled causal and which are not, it succeeds better than any other account I know.

182 See also Earman (2004).

> story for classical (= non-quantum theories) is this. In Newtonian theories determinism is hard to achieve without the aid of various supplementary assumptions that threaten to become question-begging. For special relativistic theories determinism appears so secure that it is used as a selection criterion for "fundamental fields." GTR, under the appropriate gauge interpretation, is deterministic locally in time; but whether it is deterministic non-locally in time devolves into the unsettled issues of cosmic censorship and chronology protection.
>
> Quantum physics is the strangest and most difficult case. Ordinary QM is in some respects more deterministic than Newtonian mechanics; for example, QM is able to cure some of the failures of Newtonian determinism which occur either because of non-uniqueness of solutions or the breakdown of solutions. But the fortunes of determinism in QM ultimately ride on unresolved interpretational issues. The main driving force behind these issues is the need to explain how QM can account for definite outcomes of experiments or more generally, the apparent definiteness of the classical world — an ironic situation since QM is the most accurate physical theory yet devised. Some of the extant responses to this explanatory challenge would bury determinism while others give it new life. (Earman 2007, p. 1428f)

And so on.

Nobody[183] currently knows if the laws of nature are strictly deterministic or have a stochastic component. Maybe the question is not answerable in principle, because every answer will require assumptions which can be contested. The above is meant to dispel the smugness with which the destruction of the clockwork universe is sometimes pronounced.

But the issue of determinism is orthogonal to a scientific and naturalistic world-view. The methods of science guarantee best modeling practices in *both* cases. And indeterminacy in the laws of the universe does not make the cosmos any less naturalistic.

Of a quite different nature is the claim that there is no *observer-independent* reality, giving observers (humans?) special pride of place, leading to idealism and a rejection of realism. As already discussed in the section on RC, something which is usually ignored is that only the concrete observer herself enjoys this privileged *real* position; other humans are simply part of her physical

183 In the solar system, at least.

reality. That is usually not what is intended: people saying that reality is observer-dependent usually expand the status of observer implicitly to all humans, not only to the speaker. The underlying ontological model seems to be one of humans being possessors of souls or spiritual essences; actors put into a world that is a mere stage.

Of course, that reality is observer-dependent does *not* follow from quantum mechanics:

> ... QM does not prove that there is no reality besides the measured reality. Instead, there are several alternatives to it. In particular, such reality may exist, but then it must be contextual (i.e., must depend on the measurement itself). The simplest (although not necessary) way to introduce such reality is to postulate it only for one or a few preferred quantum observables. (Nikolic 2007)

One possible interpretation which singles out preferred observables is Bohmian mechanics (Duerr 2009). An interpretation which is more removed from common sense and attributes reality only to relational aspects is relational quantum mechanics (Smerlak & Rovelli 2007).

I like the way Heathcote puts the strife between idealist and realist conceptions of quantum mechanics:

> The fault, however, does not lie entirely with Idealism – for there is an ambiguity in the term 'realism' that has been exploited to create misunderstandings where there need be none. In the first sense, to be a realist about quantum mechanics is simply to think that we should believe in the entities and structures that subserve its explanatory hypotheses. Put simply, belief goes along with explanatory success. […] On the other meaning of 'realism' to be a realist is to believe that Classical states exhaust the set of total states that a system might have – and therefore must be possessed by a system at all times. In short, classical states could not be dispositional. Quantum Mechanics only casts doubt on realism in this latter sense. Note, however, that the denial of this latter view does not automatically take one to Idealism but rather to realism in this first sense! […]
>
> In sum: the quantum state is an objective part of the real world; it may well even evolve in a purely deterministic fashion; the particles to which it belongs are capable of interacting causally and when they do their quantum states can become entangled – this entanglement is also something entirely objective. The quantum state gives rise to the

> qualities that we observe, but, contrary to the tendencies within Idealism, that state itself is far richer than the small window of observation is able to reveal. Our best reason for being realists about this state is that it is our only means of explaining the phenomena that are peculiar to quantum theory.
>
> As I write, there is a light breeze stirring at some trees. Crows call invisibly, far off, and a winter sun streams through a break in some branches. All of these things are real. Quantum mechanics has not robbed them of that reality, rather it has made it plain that they are the manifestation of something with truly unforseen complexity. Reality is not less than we thought – it is very much more. (Heathcote 2003)

An intriguing interpretation of quantum mechanics which accounts both for observer-independent reality, is completely deterministic, and solves a host of other problems associated with quantum mechanics is the many worlds interpretation (MWI) of quantum mechanics[184], originally put forward by Everett (1957).

Indeterminism in the MWI only arises for observers restricted to the "inside view" of the universe. An intuitive deterministic model of a "quantish"[185] MWI-universe, implemented via simple gates, is given by Drescher (2006, p. 123f).

There are, perhaps not surprisingly, different variants of the many worlds interpretation. Common to all of them is the explanation of quantum indeterminacy by the fact that *all possible* outcomes of quantum measurements[186] are realized and experienced in one world or another. The MWI hints at ontological and experiential riches which are so baffling that they may just be true – after all, why assume that reality is overly bounded? The MWI reminds me of the sentence in T.H. White's "The Once and Future King", written above every anthill tunnel entrance:

> Everything not Forbidden is Compulsory.

Ah, but what is allowed, what is forbidden, and why? It is intriguing that the quantum realm itself seems to impose serious restrictions on what is possible; I mentioned the stricter than classical correlations above; another obvious example is the *Born rule* which assigns different amplitude to different measurement outcomes; a better name for the many worlds interpretation would be the

184 See Barrett (1999); Vaidman (2002) for introductions.
185 The "quantish" universe is a simplified version of a full-scale quantum-universe.
186 Measurements require neither experimental setup nor observers: it is better to speak of quantum interactions.

"few worlds interpretation"[187]: not everything conceivable, and certainly not everything imaginable, happens. Rieffel writes:

> A typical quote (Deutsch 1998): "There are even universes in which a given object in our universe has no counterpart – including universes in which I was never born and you wrote this article instead." The variety of imaginative examples suggest that anything we can conceive of, even the highly unlikely, happen, if only in a small number of universes. But much of the surprise of quantum mechanics is that certain things we thought would happen, even things we thought were sure to happen, do not happen at all.
>
> Most startling are events that were predicted to happen with certainty by classical physics, but which in fact happen with probability 0. Thus, not only is it not true that everything we can conceive of is predicted to happen in some universe, but things we can hardly conceive of not happening do not happen, not in any universe. To emphasize this correction, I call it "the fewer worlds than we might think" interpretation of quantum mechanics, or the "fewer worlds" theory for short. (Rieffel 2007)

There seem to be serious nomic restrictions at work which are still beyond our comprehension.

The conjecture of the existence of a multitude of worlds – the existence of all quantum mechanically allowed worlds – seems to explain a bit more than there merely being one world, which is a bit arbitrary. But care has to be exercised. If one introduces many worlds via the quantum, why stop there? It leads us to the question: if all possible worlds exist, which worlds are possible? All physical worlds? Mathematically consistent worlds (Tegmark's Type IV multiverse (Tegmark 2007))? The latter position, while infinitely richer in worlds than traditional MWI, is still small compared to Lewis' *modal realism (Lewis 1986),* which seems to be near restrictionless.

But I feel qualms when speaking about many worlds. Speculations that go too far beyond the empirical are always in danger of becoming pure fantasy. I urge caution with MWI (and even more with the other many worlds variants): they quickly become "explain-all" theories. It is better – scientifically more fecund – to stick with the one world we find around us. Relying on the MWI to have solved the mystery of quantum mechanics may be a form of dogmatic slumber; and anything

187 The "few" worlds vastly surpass our most extreme forms of imagination in number.

going further than the MWI (which at least connects with empirical experiment through strict adherence to the quantum formalism) should at present be regarded as science fiction.

But we have ventured too far – the morasses are getting deep, and the water is already too murky to see anything worthwhile. But we can conclude that there is no reason to be found in quantum mechanics to leave the path of rationality. Maybe *strict* determinism – that every state s leads to a definite followup state f – is false. Maybe the world is stochastically deterministic, that is, propensities are real. Either way, we need to evaluate the evidence rationally. And we need to draw our conclusions rationally. Possible indeterminism does not return us to the magical and mythical universe of old.

3.2 Panqualicism

3.2.1 Monism

We spoke above of dispositions and relations. The other important point of consideration is the qualitative side of the equation, informed by a basic *monism* or identity theory; what we are interested here is a monist conception of mind and matter[188]. A good visualization of monism is the following: imagine you have magic glasses that highlight unphysical stuff – mind or soul or spirit-essence or whatever – in white incandescent light. If you believe that there is something over and above those aspects of the world that are material, then when you put on your magic glasses you should start seeing white light around people and maybe also around other entities such as animals, depending on your creed. In monism, either *everything* shines, or *nothing*: you can have your pick, dependent on if you wish to see matter as mind or mind as matter; from monism simply follows that the two are identical. Actually, this identity follows trivially from our ontological ruminations above; but the mind-matter duality is so pervasive that it deserves special mention in this separate section.

From what may we derive *monism*? A starting point is the simple and obvious fact that physical changes in your brain change your thoughts (lesions, chemical imbalances, transcranial magnetic stimulation (TMS), what have you). This is already strong evidence that the physical and the mental are not distinct; parallelisms and dualisms seem already strained from the beginning.

In section 3.1.3 we spoke about structural realism and its causal version. Here we will hook up with the philosophy of mind; we said above that properties are causal dispositions. And that indeed

188 For views like the one propounded in this chapter, see also Rockwell (2005) and Martin (2008).

exhausts their nature – from the outside view; but there is also an inside view – what it is *like to be* such a property. The inside view is the property of *being*, of *qualia*. Let us call the inside view, generally, a qualicist state. Does this not contradict our experience of consciousness being a special state? No, because qualicist states will have very different forms of "inside views". There is an analogy from physics: at high energies we don't find solids and fluids but only gases or even plasma; so the same matter has different properties when different physical circumstances obtain. Mind would be a primary material property – the inside view – but only becoming "visible" as actual consciousness in certain matter configurations.

But first a closer look at the "canonical" form of monism which I merge into the present picture of the world:

> Type-F monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions
>
> ...
>
> This view holds the promise of integrating phenomenal and physical properties very tightly in the natural world. Here, nature consists of entities with intrinsic (proto)phenomenal qualities standing in causal relations within a spacetime manifold. Physics as we know it emerges from the relations between these entities, whereas consciousness as we know it emerges from their intrinsic nature. As a bonus, this view is perfectly compatible with the causal closure of the microphysical, and indeed with existing physical laws. The view can retain the structure of physical theory as it already exists; it simply supplements this structure with an intrinsic nature.
>
> ...
>
> In its protophenomenal form, the view can be seen as a sort of neutral monism: there are underlying neutral properties X (the protophenomenal properties), such that the X properties are simultaneously responsible for constituting the physical domain (by their relations) and the phenomenal domain (by their collective intrinsic nature).
>
> …

> One could also characterize this form of the view as a sort of panpsychism, with phenomenal properties ubiquitous at the fundamental level. One could give the view in its most general form the name panprotopsychism, with either protophenomenal or phenomenal properties underlying all of physical reality.
>
> …
>
> Overall, type-F monism promises a deeply integrated and elegant view of nature. No-one has yet developed any sort of detailed theory in this class, and it is not yet clear whether such a theory can be developed. But at the same time, there appear to be no strong reasons to reject the view. As such, type-F monism is likely to provide fertile grounds for further investigation, and it may ultimately provide the best integration of the physical and the phenomenal within the natural world. (Chalmers 2003)

This kind of monism has found positive reception in the literature Lockwood (1989); Stoljar (2001); Strawson (2006); Strawson & Others (2006); it is a position which near forces itself on the naturalist (Strawson 2006). I would like to call it *panqualicism* instead of panpsychism. Panpsychism smacks too much of the anthropomorphic; of animism and forest spirits. Panqualicism leaves the connotations where they belong: that qualitative properties are normal "inside view" properties of the physical world. Type-F monism is not new territory, philosophically speaking; its modern form can be traced to the Russellian Theory of Mind (RTM) (Russell 1927); or even to Schopenhauer: the inside view is the "Will" of Schopenhauer (1859) in action. The RTM can quickly be characterized like this:

> RTM takes as its point of departure a certain view concerning the nature of the concepts of physical theory, a view which I will also call Russellian'. According to this view, physical theory characterizes the entities in its domain (including properties) strictly *extrinsically*, in terms of the causal, functional and other relations in which they stand to each other (and to experience). [Footnote 4] So, e.g., an electron might be characterized as an entity that sometimes behaves as a particle and sometimes as a wave, that has a certain mass and carries a certain charge whereby it attracts protons and repels other electrons and generates an electro-magnetic field upon moving, that plays a given role in the binding of atoms into molecules, ...and so on. All we have here is an account of causal

> and functional relations in which electrons stand to other theoretical entities such as protons, electro-magnetic fields, mass, charge, atoms and molecules — which themselves are understood extrinsically. Nowhere in all this, according to the Russellian view, does physical theory inform us of the *intrinsic nature* of the entities that are so interrelated. [Footnote 5] So one could know everything that physical theory can tell us about a subject domain without knowing the intrinsic nature of anything in that domain. However (so it is argued)
>
> all these entities can't just be ciphers; they must *have* intrinsic natures. Furthermore, it is in virtue of having these intrinsic natures that the entities have the causal powers and dispositional properties they do. The question is if we can ever know what the intrinsic natures are.
>
> It is held that in at least some cases we can. The having of conscious experiences gives us direct cognitive access to, or 'acquaintance' with, the phenomenal properties of those very experiences, and these properties are intrinsic. The way is then open to say that they also constitute the intrinsic nature of neurophysiological states. Neuro-physiological theory, like physical theory generally, characterizes its entities only extrinsically. But in this case, we are in the (perhaps unique) position of knowing by acquaintance the intrinsic nature of the entities of which the theory treats. [Footnote 6] (Holman 2008) [189]

Chalmers calls this a *property dualism* because the view acknowledges structural-dispositional properties *and* intrinsic protophenomenal properties; I disagree; I subscribe to Heil's identity theory explicated above; categorical properties are not the *bases* for dispositional ones, they are one an the same, just viewed differently[190]. There is no intrinsic nature of the mental, when we take intrinsic to mean "a property not dependent on relations"; we can't take anything in the world out of the relations it stands in. The conceptual shift in going from categorical/qualitative to dispositional/powerful is perspectival. One can view a statue from many different perspectives; one should not imbue this with ontological significance[191].

189 Holman goes on to differentiate different variants of RTM and spells out problems which need be addressed in this framework.

190 We have to make this move to avoid epiphenomenalism, which is untenable (Muller 2008).; and as naturalists we don't want to give up causal closure of the physical (which is motivated by empirical considerations, not wishful thinking, see below).

191 Apart from the fact that one can draw conclusions about the dimensionality of the world one lives in

Similarly here, just because the inside view of structural-dispositional properties is mental in nature – that is, there is something it is "like" to be structure – does not make this position a dualism, not even a property dualism. The position is of course very *related* to property dualism as discussed in the literature, but combined with relationalism and the ontological commitment to a unified world, that is, locating all splits and divisions in epistemology[192], we can call it a pure monism with good conscience. This is not only terminological strife. The point is important because it stresses the relational nature of everything, even the inside view: otherwise – if we would admit pure intrinsic properties – we would have a *categorical* property which *could* be missing in some other possible world; leading to all kinds of metaphysical nonsense, such as zombies[193] and other creatures of fantasy having sullied the halls of academe.

The view of monism as explicated above is in surprising harmony with Eastern philosophy. We find the following in Watts' beautiful "The Way of Zen"[194]:

> According to the Yogacara the world of form is *cittamatra*- "mind only"- or *vijnaptimatra*- "representation only." This view seems to have a very close resemblance to Western philosophies of subjective idealism, in which the external and material world is regarded as a projection of the mind.
>
> However, there seem to be some differences between the two points of view. Here, as always, the Mahayana is not so much a theoretical and speculative construction as an account of an inner experience, and a means of awakening the experience in others. Furthermore, the word *citta* is not precisely equivalent to our "mind." Western thought tends to define mind by opposition to matter, and to consider matter not so much as "measure" as the solid stuff which is measured. Measure itself, abstraction, is for the West more of the nature of mind, since we tend to think of mind and spirit as more abstract than concrete.
>
> But in Buddhist philosophy *citta* does not stand over or against a conception of solid stuff. The world has never been considered in terms of a primary substance shaped into various forms by the action of mind or spirit. Such an image is not in the history of Buddhist

192 Put differently: conceptual irreduction does not imply ontological irreduction.

193 Philosophical zombies are creatures *exactly* like us in every *outward* regard – they behave the same as we do in all situations, also when questioned on matters of consciousness – except that they don't have *any* internal experiences. Zombies are extensively discussed in Chalmers (1996a). For the impossibility of Zombies in the dispositional/qualitative identity view see (Heil 2003, p. 240-249)

194 I would like to be very clear that no Berkeleyian idealism is implied in any remote sense.

> thought, and thus the problem of how impalpable mind can influence solid matter has never arisen. Wherever we should speak of the material or physical or substantial world, Buddhism employs the term *rupa*, which is not so much our "matter" as "form." There is no "material substance" underlying *rupa* unless it be *citta* itself. (Watts 1957, p. 72)

Before this all sounds to fantastic, I would like to point out the following: the move to panqualicism is rather *close* to eliminative materialism[195]. I think it's rather only a choice of emphasis of words. The eliminative materialist focuses on the outside view; the panqualicist focuses on the inside view, how existence is experienced. Both views are firmly naturalistic. The seeming gaping difference – that eliminativists say that certain mental states do not exist – and the alternative position, that everything is mental of a kind – is mitigated by the fact that the property in question, namely mentality, is ascribed or negated *indiscriminately*. The current position is more coherent because it is in tune with experience.

3.2.2 Anomalous Monism

What about Davidson's anomalous monism (Kim 2003, p. 113-136)? The core principles are the following:

> *The Interaction Principle*: Some mental events causally interact with some physical events
>
> *The Cause-Law Principle*: Events related as cause and effect are covered by strict laws
>
> *The Anomalism Principle*: There are no strict laws on the basis of which mental events can predict, explain, or be predicted or explained by other events
>
> *Monism*: Every causally interacting mental event is token-identical to some physical event[196] (Yalowitz 2004)

Anomalous monism, slightly adapted, fits perfectly into the current picture. Token identity of mental events with physical events, without the existence of physical bridge laws can be given a perfectly naturalistic explanation: the reason for the unavailability of bridge laws is the tracking

195 Usually positions at opposite extremes have much in common: they come full circle.
196We do not need Davidson's supervenience principle due to a different metaphysics.

nature of neural states and its dependence on the individual experience and ontogeny of the organism. Mental events thus are lawful, but the sequence of thoughts are associative in nature[197]. The associative connections are formed by causal events in the history of the organism. To know what someone thinks, one would need to recapitulate the ontogenetic history of the specific individual neural net. But more on this in the next section.

3.2.3 Being in a State

> The moment is simply structured that way. Kurt Vonnegut

Let us have a closer look at the perspectival aspect of the mind-matter problem. The structural and dispositional nature of reality can be communicated; that is done via scientific investigation, model-construction and dissemination among humans.

But apart from communication there is actual *being* – for instance, the experience of the color red or blue. Paradigmatic for this distinction between *knowing* and *being* – although it is usually not presented that way – is the thought experiment concerning Mary's room (Jackson 1982). The thought experiment was devised to show that qualia are unphysical in nature. Mary is a scientist investigating the nature phenomenal concept of color from within the confines of a black and white room. Through her diligent scrutiny she gets to know everything physical there is to know about color. Having finished her investigation, she leaves the room – and actually sees color for the first time in her life. Does she learn something new?

Nida-Rümelin (2002) sums the argument up:

> (1) Mary has all the physical information concerning human color vision before her release.
>
> (2) But there is some information about human color vision that she does not have before her release.
>
> Therefore
>
> (3) Not all information is physical information.

197 See the Jets and Sharks model in McClelland & Rumelhart (1988).

We have here before us the distinction between "being" versus "knowing"[198]. Why should knowing, which is always *of* something, be the same as *that very* something? If one takes the distinction of *being* versus *knowing* seriously, and combines it with a panqualicist attitude, the debate between physicalism and the reality of qualia loses much of its force, some minor problems notwithstanding. *Knowing* is merely a subset of being; most *being* is not *knowing* – that is simply the commitment to a non-idealist view of the world. But there can never be knowing without being. Qualia are always incommunicable – that is the curse of indexicality; and the boon of being – what something feels like is the sum of all involved relationships (at a certain level of integration, see below). Every physical state has a feel to it. Of course, because most states are not dynamic, have no memory, no data filtering via perceptors and no annexed neural nets that extract useful information, the feel of most states would be quite empty – maybe as good as not being any "feeling" worth calling a feeling at all. But nevertheless, the position is different from the position that there is *no quality* to physical states apart from in a brain.

We can distinguish raw feels, qualia, primary consciousness, self-awareness. Consciousness seems to come in degrees; and consciousness is not only what we usually perceive it to be:

> Remarkably, consciousness does not seem to require many of the things we associate most deeply with being human: emotions, memory, selfreflection, language, sensing the world, and acting in it. (Koch & Tononi 2008)

To get a gist of how strange conscious states can be, one can have a look at the rich literature on brain pathology (Sacks (1985) is a good introduction), and we will encounter one such case below. Mental states can be very far removed from what we normally perceive them to be, removing somewhat the strangeness of conceiving of all physical states as being qualitative in one way or another.

So, what about a rock? Why not suppose that a stone has qualitative states, albeit very boring ones? The qualitative state of a rock would be in no way comparable to higher-level qualia of mammals or humans. Without memory, percepts, desires, beliefs, cognitive dynamics, without even a sense of individuation. But it is something to be a stone. It is being without awareness. Maybe it is Nirvana:

198 The distinction is not the same as Russell's "knowledge by acquaintance" and "knowledge by description" (Egner & Denonn 2009, p. 191-198) though similar. The main difference is the dissolution of the last vestiges of mind-matter dichotomy implicit in Russell's philosophy.

> Nirvana literally means "extinction," as when a candle's flame is extinguished. In Buddhist thought, nirvana has a very specific meaning and is perhaps the most misunderstood Buddhist concept among people of other religious traditions. Nirvana is not an "absence" or lack. It is instead a state of being. Nirvana is strictly defined as a state without conditioned aspects. Nirvana is without arising, subsisting, changing, or passing away.
>
> In Hindu thought nirvana is a state of liberation from individuality and the suffering of SAMSARA, the cycle of birth and death. But it also assumes the individuality is lost through merger with the divine, or Brahman. (Irons 2008, p. 370)

The brain is ordered. Only an ordered brain can entertain complex thoughts; the degree of consciousness as we value it may depend on the degree of self-reflexivity; awareness comes through reflection of mind-stuff[199], self-awareness through increased reflexion. Cognition is the *inside view* of complex physical relations; the exact structure of which necessary for our kind of cognition to occur is the task of the empirical sciences to unravel.

Sounds strange? Awareness without memory, not integrated into a larger personal narrative would be difficult to notice. This will be illustrated later on with the strange case of Jimmie and his anterograde amnesia (Sacks 1985). Imagine having a dream: while experienced dreams are often very vivid; sometimes we even remember them, but sometimes not, they are not committed to memory. They tend to fade away. In a sense, it is as if they never happened. In this sense we can imagine basic qualitative states that have no access to memory structures such as present in brains. Of course, these states would not even be dreams, because dreams operate on *existing representations in brains*; these basic qualia states would, from our perspective, be so uninteresting that we might not consider calling them qualia, if this would not conceptually reintroduce the rift just closed.

3.2.4 The Intentional

In the present conception of things, the intentional is not "the mark of the mental" – or only insofar as everything is mental. It might just as well be called the mark of the physical. Both Heil and Smith (see the Appendix) give accounts of how intentionality can be incorporated into the natural view of things. Indeed, intentionality is necessarily coupled to ontology:

199 Mind-stuff is of course nothing other than ordinary matter.

> The prospects of a naturalistic grounding for intentionality can be appreciated only if we have some sense of what the natural world has to offer. (Heil 2003, p. 208)

And:

> In any case, we have available a resource ideally suited to account for the kind of projection associated with intentionality: dispositionality. Dispositions are *of* or *for* particular kinds of manifestation with particular kinds of disposition partner. Dispositions preserve the mark of intentionality in being of or for particular kinds of manifestation with particular kinds of non-existent—possible, but non-actual—objects. This is not mysterious or spooky; it is a feature of dispositions possessed by rocks, or blades of grass, or quarks.
>
> My suggestion is that we make use of the 'natural intentionality' afforded by dispositions in making sense of the kinds of intentionality we find in the minds of intelligent agents. (Heil 2003, p. 222)

3.2.5 Evolution of the Mental?

One argument for mind-matter identity is simply the strong empirical evidence from medicine, neurobiology, neuropathology etc; from this evidence follows that one needs to pay a high price for keeping some sort of dualism. But I think there is another strong argument in favor of monism: it seems to me that consciousness can't evolve. It must be a *primitive property* of the universe.

Why this? Evolution always acts on physical configurations, that is, DNA, the cellular environment etc. Morphological changes are selected for because they confer adaptive advantage. The morphological advantage conferred by a brain is its utilization of certain states (such as planning and memory) to enhance survival. But consciousness *qua* consciousness must have already "been there" in the first place. A bit as the laws of physics (especially aerodynamics) need to be in place so that the wing of a bird can evolve. Mind is here from the beginning, mind in itself does not evolve, only *functions* such as perceptions, memory, consciousness. What has evolved is not consciousness, but concrete algorithms and concrete correlations with external world such as planning, backward chaining, logic and self-representation – utilizing the *basic mental properties of*

material nature. Evolution can't bring forth basic constituents of reality such as electro-magnetism, mass[200] etc, and consciousness.

After the panqualicist move – ascribing a state of "what its likeness" to all physical states – brains take on a different role; they are not needed to "generate" mind *per se* as in more dualistic conceptions, they only constitute the *specific* kind of mind as we experience as human beings[201]. Perceptors correlate the brain with the environment. These correlations are integrated via neural mechanisms, that is, filtering and integration capacities such as Bayesian algorithms (Doya et al. 2007); many of these correlations are stored for future reference – memory.

Brains then are *person generators*, associative memories constituting the substrate for a narrative; and *planning machines* (prediction, modeling etc); and of course their complex structure account for higher order consciousness, that is, interesting conscious states; those we so value as human beings. The complex brain structures operate on the representations that have arisen: thinking, planning, reflexion; thought about thought. Without a brain there is no memory. Without memory, there is no person. But more on that below. But that does not mean that without a brain there is no state of being – of primordial likeness.

Another consideration works in favor of why qualia must be there in the first place. The problem is this: how does material evolution find the correct qualia to solve certain problems, that is, pleasure for things that should be rewarded and pain for those that should not? A purely functional account fails to explain: if a state feels pleasurable *because* it fulfills its function, the question arises why it should feel like anything at all, if the function is performed anyway – epiphenomenalism rears its ugly head. Heil is puzzled by this also:

> Even if we knew that neural tissue arranged in a particular way yielded a feeling of pain, the reason this arrangement yields pain rather than some other feeling (or no feeling at all) remains an utter mystery. (Heil 2003, p. 235)

The picture is different if we recognize that all material states already have a basic qualitativeness to them; then, evolution, just as it optimizes body plans and behavioral strategies to solve environmental problems, can start optimizing material configurations so that *configurations that are pleasurable* are causally related to behavior that enhances fitness, and material states that

200 Smolin (1997) and his cosmic evolution was already mentioned; that is evolution on another scale.
201 Note: brains do *not cause* minds. They *are* minds. Special kinds of minds.

feel bad are causally related to behavior that is contrarian to the goals of staying alive and reproducing.

Imagine an organism that was wired "the wrong way": it is wired in such a way that as soon as a massive nerve impulse from one of its limbs due to its having been injured arrives in the brain, the brain reacts with pleasure – such an organism would not survive long; it would seek states of injury, and thus die of any number of causes (bleeding to death, infection, getting eaten); it would most probably fail to have progeny. Only by assuming that material states directly *are* qualia states – and we reject the possibility of zombie worlds or mechanisms disrupting the causal closure of the physical – can evolution *use* qualia to operate on them.

3.2.6 Binding Problem

In the literature the binding problem is often presented as especially problematic for monism, although I don't quite see why other positions should be exempt from solving it:

> There is one sort of principled problem in the vicinity. Our phenomenology has a rich and specific structure: it is unified, bounded, differentiated into many different aspects, but with an underlying homogeneity to many of the aspects, and appears to have a single subject of experience. It is not easy to see how a distribution of a large number of individual microphysical systems, each with their own protophenomenal properties, could somehow add up to this rich and specific structure. Should one not expect something more like a disunified, jagged collection of phenomenal spikes? Chalmers (2003)

We can call this the binding problem; is it a real problem? Some do not think so:

> One of the motivations for models with quantum coherence in the brain was the so-called binding problem. In the words of James [...], ''the only realities are the separate molecules, or at most cells. Their aggregation into a 'brain' is a fiction of popular speech.'' James' concern, shared by many after him, was that consciousness did not seem to be spatially localized to any one small part of the brain, yet subjectively feels like a coherent entity.
>
> However, nonlocal degrees of freedom can be important even in classical physics, for instance, oscillations in a guitar string are local in Fourier space, not in real space, so in this case the

> ‘‘binding problem’’ can be solved by a simple change of variables. As Eddington remarked [...], when observing the ocean we perceive the moving waves as objects in their own right because they display a certain permanence, even though the water itself is only bobbing up and down.
>
> Similarly, thoughts are presumably highly nonlocal excitation patterns in the neural network of our brain, except of a nonlinear and much more complex nature. In short, this author feels that there is no binding problem. Tegmark (2000)

I agree with Tegmark for a simple reason. In the present conception of reality, time and space are not stages on which matter operates; everything is a unified whole and stratifications are imposed by us (see Smith's way of doing ontology above). The neutral monist substance, through its dispositional and qualitative structure, constitutes such things as time, matter and consciousness. The binding problem only arises from an ontology where one has forgotten to pay the price: a naive mechanistic ontology which imbues anthropocentric concepts of mechanism with ontological significance.

3.2.7 Truly at Home

> A monk asked master Joshu: "Does a dog have Buddha-nature or not?" Joshu said: "Mu"[202]

Sitting quietly doing nothing

Spring comes, and the grass

grows by itself.

The wild geese do not intend

to reflect their images

The water has no mind to receive them.

An old tree preaches wisdom.

A wild bird is crying the truth.

202The Chinese character for “mu” means “absence”, “nothing”, “non-being” etc. But the Buddha said that all creatures have Buddha nature. Joshu does not want to engage the monk on a literal level. He wants to urge the monk to “unask the question” and thus transcend to a new level of understanding. (Yamada 2004, p. 29f)

So, in a nutshell we can summarize the results of this section like this: *Mind* is not an ephiphenomenon, because it *is* the very causal dispositional structure of the universe. It is the ultimate nature of existence, not as a categorical base, but as a categorical-powerful identity – Schopenhauer's Will. But mind is something quite different than we humans think it is, because states of being (=mind) are more various than human minds can conceive.

There is reality without observation, because observation and reflection are already *special* states of mind; not all mind must reflect on itself. To let all existence depend on these very human states, as idealism does, strikes me as the peak of human hubris and arrogance. The sun and the moon are not created by the creatures observing them. The observing creatures are instantiations of the same "material" as sun and moon, but in a more complex way due to aggregation of causes and natural evolution. So, despite the prominent role of mind, associating the present position with idealism would locate it in the wrong camp completely.

As a materialist who acknowledges the qualitative aspect of matter, you are at home in the universe. You can even imagine that qualia like love etc feel roughly the same for different persons, as we *share* large parts of our cognitive architecture and, in the present view, due to the identity of qualitative and powerful properties, matter states that are similar will correspond to roughly the same qualia states. This integrated view opens up the possibility for *closing the separateness* felt by some as persisting between human beings in a wholly naturalistic way.

Coming from a Western context, the present metaphysics has left us in a strange world: the elimination of essences, the relational character of knowledge about the world and the monist metaphysical conception of the universe as mind and matter being just two ways of speaking about the same underlying dispositional/qualitative entity. To Buddhist ears, this all should not sound so strange. Two doctrines deserve mention. The first one is sunyata: voidness, emptiness. This is to be understood as the impermanence of all being, that nothing has a enduring identity. The second one is pratityasamutpada: dependent arising; causal powers are the driving forces of everything. Pratityasamutpada is the insight that current thoughts and actions depend on past ones and current ones will originate future ones; the mental underlies the laws of causation as much as does the physical; as it evidently is, in a monist universe.

Mind is everywhere. The basic monism is what allows knowledge and intentionality, reflection, compression, representation, “imported” via dispositional effects well known from physical

relations. Knowledge is possible because there is no epistemic gap to bridge. Everything has Buddha-nature.

4 On Persons

4.1 The Will

4.1.1 The Problem: Free Will

> For as the eyes of bats are to the blaze of day, so is the reason in our soul to the things which are by nature most evident of all. (Aristotle MET)

> Macht ohne Siege. - Die stärkste Erkenntnis (die von der völligen Unfreiheit des menschlichen Willens) ist doch die ärmste an Erfolgen: denn sie hat immer den stärksten Gegner, die menschliche Eitelkeit. (Nietzsche 1879)

The issue of free will must be addressed in this work because misconceptions about the subject are a major hindrance against the adoption of a naturalistic world view. Physicists for instance have developed the “Free Will Theorem”:

> It asserts, roughly, that if indeed we humans have free will, then elementary particles already have their own small share of this valuable commodity. More precisely, if the experimenter can freely choose the directions in which to orient his apparatus in a certain measurement, then the particle’s response (to be pedantic – the universe’s response near the particle) is not determined by the entire previous history of the universe. (Conway & Kochen 2008)

While the mathematical results of the paper are – to my knowledge – beyond reproach, the suggestive naming of "free will theorem" does not of course force itself at all. What is shown is that certain relations must hold between experimenters decisions and particle trajectories; one might as well call it the "No Free Will Theorem" if one chooses a different metaphysical lining. The problem is that such papers are not understood in public discussion – what often remains is the suggestion that physicists have "proved free will", which is of course nonsense.

I claim that free will does not exist. The claim of this statement is not as bold as it may seem at first glance. The proponent of the statement must elaborate at least on three things:

- Why we don't have free will.
- If we don't have free will, why do we have the persistent illusion of it?
- What is a sensible alternative conception to that of free will?

The following section will try to address these points in turn. At first we need to clarify what we mean with *free will*, as for different people this concept means very different things. It is quickly evident that the very concept is incoherent. Especially in the question of free will terminological clarity is of the essence – and it can't be cleared up often enough. How does the intuition behind most people's conception of free will work?[203]

Now why go to such lengths to bring the discourse on free will in line with day-to-day speech, when philosophers have technical jargon at their disposal? I do not believe that the split into an "elite educated philosopher clique" discoursing in an ivory tower and an “uninformed” public is desirable; especially not concerning an important topic for our everyday lives such as free will. Here, if not elsewhere, it is the responsibility of the philosopher to talk clearly and to avoid obfuscation at all costs.

In science, it is inevitable to develop jargon; in highly specialized domains with intricate modeling and a myriad of new entities and associated mechanisms – think of the proteins in molecular biochemistry and their functions – the need for a new vocabulary arises to communicate quickly (see above). But in philosophy, when analyzing commonsense concepts such as free will, we should be careful with jargon; and if we develop sophisticated new concepts, we should give them *new* names. Many people turn to philosophy for guidance of what to believe and how to orient themselves and not, say, to molecular biology, so philosophers should speak and write in a way that is also accessible to the layman; at least on topics of interest to the layman.

Free will is so strongly associated with supernatural conceptions of the self in the general populace, that philosophers obviously can only add to the confusion by constructing different models – naturalistic ones – and still calling the result "free will". I take it that for most people, *free* simply means that humans are in some way exempt from the laws of nature. The idea is that there is *more* to our consciousness than mere chemical processes in action. One can be free *to* do something, and free *of* something. Freedom *of* something is usually taken to mean free of the laws of nature. So, here is the working definition of free will for the rest of the thesis:

203 I mean people who do not pursue thinking about free will professionally like neuroscientists, cognitive scientists, AI researchers, philosophers etc. Now of course many lay people will not even have an account like the one developed below and "free will" is simply the belief that a human being has it without even knowing what it is; it is a script, a cached thought, instantiated in certain cases to justify certain forms of behavior and reactions to oneself and to others.

Definition: *Free will* is the supposed ability of agents, notably humans, to decide in some way independent of the laws of nature; that is, their behavior is not fully determined by the laws of nature[204].

Now, by laws of nature, we might mean either strictly deterministic – causal – laws, or stochastic laws. Strict determinism means that there is a one-to-one mappings of states onto other states; whereas stochastic determinism is lawful behavior of the kind where one state at time t can map into different states at time t+1 with a probability weighting assigned to the possible transitions. In this section, the word "determined" will always include the case of stochastic lawfulness.

What I don't want to address under the rubric of free will here is the ability of some brain states to override other brain states; first of all, separating determined brain processes instantiating "willing" from those instantiating mere "action incentives" can hardly be called *free* in the usual sense of the word. This intuition underlies conceptions of free will coming from the rationalist camp: free will is identified with the reflexive component which is supposed to override more "primitive" desires and drives.

But this conception is better captured by *akrasia:* weakness of the will. Why is there procrastination? Why are people drug addicts, or unable to stop smoking though they *want* to? Why can't overweight people stop eating and do more exercise to lose weight, even though they would so much like to look that sexy commercial man? Seeing that we can have a weak will comes much closer to an explanation of what is at stake. But weakness and strength of will still have nothing to do with freedom of the will. If one has a weak will or a strong will, that is perfectly compatible with a strictly causal universe. What is happening in the brain is that there is a neural battle between different neural regions. Every conception of free will has serious problems accommodating the influence of the unconscious; as the concept of *willing* strongly presupposes some Cartesian rational entity not influenced by "lesser" animal traits; such as emotional currents coming from the limbic system.

I suspect the rationalist intuition which insists on calling this "battle" free will still draws from remnants of Cartesian Dualism[205]: some "real self" suppressing the subordinate animal, material

204 Compatibilism is rejected because it uses words wrongly, violating our Confucian Principle mentioned above.

205 Which is very hard to purge from the system because of the failure of global updating. In the inferential system of a person, even if he/she says "I do not believe that the mind is independent of the physical, no, it supervenes ...." etc, they will often reason in other contexts as if the mind did indeed have a life of its own. This is because mind-body dualism is deeply ingrained in our intuitions and has been fostered, at least in Western culture, since our childhoods. Additionally we don't have access to our neural level, which also lets thoughts seem to pop up magically; out of

self. This kind of reasoning is better captured by the concept of optimal will, which I will propose to replace free will.

Freedom always suggest possibilities. That is the other sense of freedom mentioned above, the freedom *to* do something. And free will means – if the concept is worth anything – that an agent could have willed differently in a *physically identical situation*. That is of course possible if there is randomness at the bottom of the universe; but again, this is not what people usually mean when they speak of free will. Wouldn't our choices being random be even worse than if they were determined? What is important for people arguing this way is a moral issue: that, for example, a criminal *could* have decided differently in the past, that he was free to have taken a moral action instead of an immoral one. But the subject of morality, while interwoven with free will, is a *different* subject matter and will be addressed below.

There is another interesting scenario which highlights the problematic nature of the concept of free will. If you have free will, it should mean that you are free to change your will. For instance, if I have free will, I should be able to decide from one day to another to change my sexual orientation, to change my moral values, to stop being afraid of spiders etc. But clearly, I am not free to do any of these things. A quote attributed to Clarence Darrow comes to mind:

```
    I don't like spinach, and I'm glad I don't, because if I liked it
I'd eat it, and I just hate it.
```

Could we want what we not want? Or not want what we want?[206] Either you want (will) or not, or you are uncertain, and the uncertainty can be resolved by deliberation, that is, calling into consciousness previous experiences or reasons grounded on such experiences; or by seeking new experiences. New experiences or new information can influence our will. And even silent reflection in the cold chamber of a high and forbidding tower is nothing but the continuation of the causal chains set in motion at the beginning of time. But it seems that this kind of free will – that is, the

nowhere, actually. Not even reading a neuroscience paper will suddenly change all those neural connections built up since childhood which embody the concept of mind-body dualism. Getting rid of this intuition often takes years of work, applying the insight in ever new contexts (this corresponds to low-level neural reweighting in different associative contexts). See Papineau (Forthcoming)who argues that nearly everyone is in the intuitive grip of dualism. One possible reason for the iron grip of dualism are phenomenal concepts: phenomenal concepts possess a 'faint copy' of the actual phenomenal experience, say, of the color red. Thinking of them recreates part of the experience when having perceived them. Third person concepts do not possess such a faint copy and thus feel utterly different (Papineau 2003b).

206 Proponents of free will usually do not mean that we are really free to change what we *will*, but that we are free to choose what we *do*. The minimal concession of a stout free-willer would be to speak of "free action" instead of "free will".

ability to really change yourself fundamentally without cause – is not what people want of free will; it is something else. I know not what.

Now on to the problematic – I say, incoherent – nature of free will is supernaturally conceived. There are, metaphysically, four interesting variants of how the world is organized in relationship to the free will issue:

1) completely deterministic, no extraphysical forces
2) completely indeterministic, no extraphysical forces
3) partly deterministic, partly random, no extraphysical forces
4) the indeterministic component in 2 or 3 is influenced by an extraphysical mind; maybe the quantum dice are loaded because of a realm of the mental (where a homunculus controlling the mechanical parts of the human body lives)

With positions one to three, I argue, it is *irresponsible* to talk of free will. Which leaves case four, which will now be analyzed in more detail[207]. The explication below is to be seen as the *argument* someone who believes in free will *should* give, at least in its general outline; I use quantum mechanics for the introduction of indeterminism, but nothing hinges on that. I will call what is argued for, for extra clarity, *supernatural free will* and abbreviate it SFW in the text. The reasoning goes as follows, and, in accordance with the very human approach to reasoning taken in this thesis, psychological steps are included:

The intuition of supernatural free will in four steps

1) Many laws of nature are determined strictly. That means that a physical state s0 passes into state s1 by a transition function t_det which is completely lawful; it is especially not influenced by minds, souls etc. This is the Newtonian picture usually presented as a straw man to begin the discussion and denounce the "outdated" views of scientists denying free will.

2) Quantum mechanics entails basic physical indeterminism[208]. We now have a stochastic transition function t_stoch, that is, it the state s0 passes into states s1_0, s1_1,...s1_n ... with weighted probabilities. We do not have strict determinism as in (1) anymore.

3) Mind is something over and above mere physical law, not mere matter in motion.

207 Conceptions such as Leibniz's preestablished harmony will be ignored completely.
208 We saw above that this is only true in some interpretations. But again, nothing hinges on this.

4) So, the transition function t_ stoch does not define the successor state s1_x strictly – there are a number of – possibly infinitely many – successor states s1_n. Somewhere in the brain, at micro-levels maybe not even detectable, the mind or the soul which is *independent* of the laws of physics influences the quantum indeterminacies, so that at the macroscopic level a decision emerges which is *not the product of the laws of physics alone* (be it t_det or t_stoch), but which has been *nudged* by the little homunculus (mind, soul) into some desired outcome state.

To be especially clear: the mind/homunculus is *bullying* the quantum wave function around (or whatever other poor stochastic law is at hand). That is, the quantum wave function would evolve differently were not a mind present directing it in some sense[209].

The above is what I call SFW, and my denial of free will is dependent on my contention that only a model in the sense above, where natural law is actively influenced by an entity outside natural law, is what *should* be called free will. Now, there are philosophical positions like compatibilism that build different models, with no separate homunculus transcending physical law. But for these other models, one should simply use *different names* to reduce confusion. An excellent paper on "free will" by Levy (2003), impeccable in both its premises and its conclusions, details "free will" in a completely naturalistic sense. The paper does have one flaw: at the end, Levy chooses to call his model "free will".

The above conception – SFW – is not a straw man. Flanagan tells the following story[210]:

> In the spring of 2000, I taught a mixed undergraduate and graduate level course on the philosophy of mind as a visiting professor at Boston University. Among other books, the students read Damasios's *The Feeling of What Happens* […] and Daniel Dennett's 1984 classic, *Elbow Room: The Varieties of Free Will worth Wanting*. They enjoyed both brooks greatly, but Dennett's made them nervous in a way Damasio's didn't. Why? Because, although both argue for a fully naturalistic conception of mind, Dennett explicitly asserts that Cartesian free will is a variety of free will that (a) is not worth wanting and (b) you can't have even if you do want it because it is based on

209 For proponents of the Many Worlds interpretation of Quantum Mechanics free will is not an option at all: they contend that all states s1_n are realized; thus there can be no preferred state that a homunculus nudges the brain into. The model given above only works for interpretations which single out actual from non-actual histories.

210 To understand it, I have to say two things. The "Cartesian Version" of free will mentioned in the quote is a kind of SFW – an unphysical mind causing physical stuff to do things. Secondly, the distinction between the *manifest image* and the *scientific image* was introduced by Sellars (1962), where the manifest image is roughly the one underlying our everyday conception of the world, consisting of an humanistic image and an image of the world. The manifest image is often at odds with the scientific image.

> philosophically and scientifically incredible ideas, and is therefore incoherent, impossible. Over several weeks of seminar I heard the refrain - "But, but, Dennett's conception of free will is not *really* free will!"
>
> It is the specifically Cartesian conception of free will that is part of the manifest image and thus the only contender for an acceptable view of free will. It is *the* variety of free will most everyone wants and thinks they need. (Flanagan 2002, p. 85f)

The passage of Flanagan goes on to support what was said in section 2.5.2 on the "Power of Words":

> When Damasio says free will is real, he definitely does *not* mean Cartesian free will. But because the Cartesian conception of free will is the default view within the manifest image, he is not read as challenging that view. The words "free will is real" are read as comforting, as meaning that Cartesian free will is real, even though when when coming from Damasio's pen, they definitely do not mean that. (Flanagan 2002, p. 86)

Well then, here, I have defined SFW, said explicitly that it is incoherent, and will develop a different account below.

4.1.2 The Strangeness of Supernatural Free Will

> *What is willing*! We laugh at him who steps out of his room at the moment when the sun steps out of its room, and then says: "I will that the sun shall rise"; and at him who cannot stop a wheel, and says: "I will that it shall roll"; and at him who is thrown down in wrestling, and says: "here I lie, but I will lie here!" But, all laughter aside, are we ourselves ever acting any differently whenever we employ the expression "I will"? (Nietzsche Daybreak)(124)

SFW is strange. While strict determinism rules out free will as defined above, there is room in the stochastic deterministic case, but also only in a magical kind of way. We would need to assume that there is something to a person – a decider – over and above the physical instantiations, that can influence transition probabilities[211]. In SWN, there must always be some kind of homunculus or

211 A weaker claim may also be possible, namely that there is a kind of feedback of the physical system onto the transition probabilities – but that is again a very physical view of things, because if this kind of feedback exists it

mental substance *bullying physical law.* The mental substance needs ways to interact with the physical brain[212].

Let us look at what analytic philosophy has to say to *interactionism*, the idea that there is a realm of the mental interacting with the realm of the physical. The premises needed to kill off interactionism are these:

```
RAP: The Real Argument for Physicalism

1. There are causal relations between the mental and the physical.

2. Causation involves the transference of energy.

3. Anything with energy is physical.

-------------------------------------------------

Thus: The mental is physical. Montero (2006)
```

Montero, it should be noted, does not endorse this kind of argumentation; she just shows what premises are needed to exclude dualism. As always, we are now confronted with *empirical* questions. And empirical answers were found. Modern physics currently posits four fundamental forces: electromagnetism, gravity, the strong and the weak nuclear forces. All except gravity have been unified. Philosophers or scientists wanting to find a role for mental causation in this scientific backdrop will be hard-pressed. Physicalism is not a philosophical fad. It rose due to empirical investigations into physiology:

> In the first half of the twentieth century the situation changed, and by the 1950s it had become difficult, even for those who were not moved by the abstract argument from general reducibility, to continue to uphold special vital or mental forces. A great deal became known about biochemical and neurophysiological processes, especially at the level of the cell, and none of it gave any evidence for the existence of special forces not found elsewhere in nature.

must again either be strictly causal or stochastic. McFadden (2002) for instance looks for free will in the EM field generated by the brain. But it wouldn't be, as Flanagan's students would say, *real* free will.

212 Descartes was consequent, believing the body to be purely mechanical, which we can equate with physical here, and the soul being immaterial and *controlling* the mechanical body. This soul-homunculus would then be the seat of free will and Descartes assumed that control – interaction – happened in the pineal gland.

> During the first half of the century the catalytic role and protein constitution of enzymes were recognized, basic biochemical cycles were identified, and the structure of proteins analyzed, culminating in the discovery of DNA. In the same period, neurophysiological research mapped the body's neuronal network and analysed the electrical mechanisms responsible for neuronal activity. Together, these developments made it difficult to go on maintaining that special forces operate inside living bodies. If there were such forces, they could be expected to display some manifestation of their presence. But detailed physiological investigation failed to uncover evidence of anything except familiar physical forces.
>
> In this way, the argument from physiology can be viewed as clinching the case for completeness of physics, against the background provided by the argument from fundamental forces. One virtue of this explanation in terms of two interrelated arguments is that it yields a natural explanation for the slow advance of the completeness of physics through the century from the 1850s to the 1950s. (Papineau 2001)

A detailed historical overview of the *empirical* rise of physicalism is given in Papineau (2002, p. 232f).

Interactionism, given today's evidence, is highly implausible and in serious need of experimental vindication if further attention should be given to it; all arguments for dualism are ad hoc to divert the “most terrible” of all scientific-philosophical results: that human beings are a natural part of this universe.

But let us consider interactionism for a moment; just for fun. The Cartesian conception of free will has a very strange metaphysics; the central question is: if there is a mental substance, how does it *know* where and when to interact with physical matter, when to divert particles in their course? How does the mental substance know that it has a brain to deal with and not a haphazard arrangement of carbon and water molecules? If the mental substance would not only influence brains but everything, it would hardly be able to qualify as *mental* but would be a normal physical law – similar to the present monist account. How does the mental substance know if a protein is still in food, say a hamburger, or in the mouth, in the stomach, in the bloodstream, or finally, when amino acids are transformed into neurotransmitters, integrated into the brain. Where should mental causation begin? At what stage should free will enter? How does "the soul", "the homunculus"

know when to *influence* the matter previously *out of its lawful reach*? Or does the hamburger already have free will, as Conway & Kochen (2008), at least consistently, would probably suggest?

Proponents of free will should be able to give clear criteria not only how the homunculus can interact, but when it chooses (?)[213] to do so; because clearly, purely physical effects can severely influence the ability to exercise "free will": alcohol, anesthetics, even something as banal as a good hit on the head. So it seems that Cartesian free will only works in special circumstances. And what about this mental substance which interacts with physical matter anyway: does it follow spirit-laws? – it must, if it only interacts with brains and not with "inanimate" matter.

Not only is Cartesian dualism/interactionism[214] on a bad footing empirically, it also does not solve any problems; it only adds new ones. Detailed analysis of the free will issue in line with the my work is given in Dennett (1984); Pereboom (2001a). Nietzsche saw this all clearly over a hundred years ago (Leiter 2007). But now to the issue of why the idea of free will is so enduring.

4.1.3 The Enduring Fallacy

There are a number of reasons why people want to believe in free will and why they think they need it. We will now have a look at some of the reasons in turn.

4.1.3.1 Theology

> I am an Agnostic because I am not afraid to think. I am not afraid of any god in the universe who would send me or any other man or woman to hell. If there were such a being, he would not be a god; he would be a devil. Clarence Darrow[215]

Theology has a vested interest in upholding the doctrine of free will. If you have the concept of an all-powerful and all-good God, you have to introduce free will to account for quite a number of things. For instance, in texts of Christian apologists, we read this (these books are published by Oxford University Press, in the case below as of 2005!):

> … God wrongs no one if he allows them to 'lose their soul' in the course of an earthly life as a result of a long series of free choices

213 Does the Cartesian homunculus have free will? Are there homunculi all the way up? Or down?

214 Occasionalism, epiphenomenalism and parallelism are even worse doctrines metaphysically. The rationalist may refute these positions as an exercise.

215 http://en.wikiquote.org/wiki/Clarence_Darrow; Quoted in an eulogy for Darrow by Emanuel Haldeman-Julius (1938). Retrieved 22.08.2009

> to do what they know to be bad, [Footnote 9] until they cease to be a moral agent. [Footnote 10] (Swinburne 2005, p. 215)

Where the first part of footnote 9 is this:

> There are different kinds of free will according to how easy it is for someone to choose the good. There are advantages in God making it easy for us to choose the good, since making such a choice is a good thing. There are also advantages in God making it somewhat more difficult for us to choose the good, since we then have the opportunity to make a more heroic and so a more committed choice of the good. So long as God makes it possible for us to choose the good, God does not wrong us. (Swinburne 2005)

And:

> A central core of a plausible theodicy has always been the free-will defence—much human suffering is due to humans freely choosing to hurt each other (or through negligence freely choosing to allow each other to be hurt), and much suffering provides the opportunity for humans freely to choose to help each other; and it is very good that humans should have those significant choices.(Swinburne 2005, p. 260)

Now, it is clear that, to save the concept of an all-powerful and benevolent God, you need human agents endowed with free will to account for *some* of the evil happenings in the world; God can remain benevolent, even though he does not stop crimes, because he weighs in the greater good of giving his subjects the ability to *decide*. It is an excuse for God (Leibniz's theodicy) not having to stop evil in the world. The theodicy solution fails for two reasons: firstly, natural catastrophes are not accounted for; when sentients die wholesale; this clearly a good God could prevent without interfering with the free will of the little creatures.

But the whole concept of free will to *choose* the good and then *go to heaven,* the sentiment expressed in the text of Swinburne above, is unnatural if you keep God all-powerful, and, hopefully, all-knowing/wise. Let us have a look at Earth from outer space. Seen from a superintelligence view – imagine advanced aliens observing the earth – the behavior of humans and the results of their behavior, that is, their cities, cultures and civilizations, would probably amount – remember, this is from a superintelligence perspective – to little more than something similar to an anthill, albeit of planetary dimensions.

Now, humans would never judge ants on their *morality*; God as conceived by religion would certainly be the most advanced being in our universe, and so every believer would probably have to agree that the gap between humans and ants is *smaller* than the gap between humans and God. But how could such an advanced God be vindictive of humans not following his divine law? It would be comparable to a human torturing or killing ants because they failed to perform well on an optimization problem (carry food to nest, cooperate etc) which the human has set up for, say, his amusement. Such behavior would be euphemistically be described as childish, more correctly as sadistic. So, why should God judge people exercising their free will who are so much more limited than him?

Secondly, the assumption of benevolence is violated all the same: if God were benevolent, why would he put souls in a situation where they can decide against his omnipotent will *in the first place*? Imagine you are a God, and you have some souls left over for deployment. So you create a world, and you know – you are omniscient after all – that some of them[216] will decide against you, and that you have to send them to eternal damnation for that. After all, you want to stay true to your promises as a God.

But now we see why we have already violated the assumption of benevolence! Surely, if there is a heaven, living there is ultimately superior than living out a life on Earth will all its uncertainties and being able to, maybe even unwittingly, decide against heaven. Why not transfer all souls to heaven ab origine? The theodicy solution only works if you assume that God has *no power to change the soul-roulette* (so now, omnipotence is at stake). Let's just leave it there, because going round in circles is boring, except when dancing.

Albeit a seriously broken argument, the gist of the theodicy is still a major motivation for many religious people to continue to believe in free so that they are able to uphold their other beliefs (heaven, God rewarding them for good behavior etc).

4.1.3.2 Free Will as a Value

Free will is not what makes us valuable as human beings; it is not what adds meaning to life – what is often presented as evident is simply cultural backdrop. Life is meaningful qua being lived. Our value stems from being in the universe, not more, not less. Our value is built-in into use via our qualitative experiences (more on that in section 5.5).

216 You don't know which ones, because you give them free will: that is, you randomize some of their behavior in respect to your otherwise omniscient knowledge.

We can learn to value an un-free rational mind-set:

> One should note that our sense of self-worth, our sense that we are valuable and that our lives are worth living, is to a significant extent due to factors that are not produced by our volitions at all, let alone by free will. People place great value, both in others and in themselves, on beauty, intelligence, and native athletic ability, none of which are produced voluntarily. (Pereboom 2001b)

4.1.3.3 Morality and Responsibility

> Irrtum vom freien Willen. - Wir haben heute kein Mitleid mehr mit dem Begriff »freier Wille«: wir wissen nur zu gut, was er ist - das anrüchigste Theologen-Kunststück, das es gibt, zum Zweck, die Menschheit in ihrem Sinne »verantwortlich« zu machen, das heißt sie von sich abhängig zu machen... Ich gebe hier nur die Psychologie alles Verantwortlichmachens. - Überall, wo Verantwortlichkeiten gesucht werden, pflegt es der Instinkt des Strafen- und Richten-Wollens zu sein, der da sucht. Man hat das Werden seiner Unschuld entkleidet, wenn irgendein So-und-so-Sein auf Wille, auf Absichten, auf Akte der Verantwortlichkeit zurückgeführt wird: die Lehre vom Willen ist wesentlich erfunden zum Zweck der Strafe, das heißt des Schuldigfinden-wollens. Die ganze alte Psychologie, die Willens-Psychologie hat ihre Voraussetzung darin, daß deren Urheber, die Priester an der Spitze alter Gemeinwesen, sich ein Recht schaffen wollten, Strafen zu verhängen - oder Gott dazu ein Recht schaffen wollten... Die Menschen wurden »frei« gedacht, um gerichtet, um gestraft werden zu können - um schuldig werden zu können: folglich mußte jede Handlung als gewollt, der Ursprung jeder Handlung im Bewußtsein liegend gedacht werden (- womit die grundsätzlichste Falschmünzerei in psychologicis zum Prinzip der Psychologie selbst gemacht war...). Heute, wo wir in die umgekehrte Bewegung eingetreten sind, wo wir Immoralisten zumal mit aller Kraft den Schuldbegriff und den Strafbegriff aus der Welt wieder herauszunehmen und Psychologie, Geschichte, Natur, die gesellschaftlichen Institutionen und Sanktionen von ihnen zu reinigen suchen, gibt es in unsern Augen keine radikalere Gegnerschaft als die der Theologen, welche fortfahren, mit dem Begriff der »sittlichen Weltordnung« die Unschuld des Werdens durch »Strafe« und »Schuld« zu durchseuchen. Das Christentum ist eine Metaphysik des Henkers... (Nietzsche 1888)

> ….human beings are a species splendid in their array of moral equipment, tragic in their propensity to misuse it, and pathetic in their constitutional ignorance of the misuse. (Wright 1995, p. 13)

> In the mind there is no absolute or free will, but the mind is determined to this or that volition by a cause which is also determined by another cause, and this again by another, and so on ad infinitum. (Spinoza 1677)

> This doctrine contributes to the welfare of our social existence, since it teaches us to hate no one, to despise no one, to mock no one, to be angry with no one, and to envy no one. (Spinoza 1677)

Free will crops up in two situations: negatively, when it is used to assign blame and guilt: "You bad boy, you could have done otherwise!" And positively, when applied to goal-seeking behavior, intentionality, looking for meaning in life etc. The positive aspect will be covered by my concept of *optimal will* in section 4.1.4; the former will be addressed in this section. By splitting up these two components, matter will clear up more quickly.

So, to the negative component: there seems to be a concern to locate moral responsibility outside of strict causation – why? I think that there are evolutionary reasons for this: moral enforcement is coupled with punishment, vindictiveness, feelings of anger. Deriving human actions from causal chains removes the potential for being angry at someone; after all, if one could calculate the motions of atoms and explain why a person acted in such and such a way, there would be little point of being angry at her. An explanation which leads to understanding removes the possibility of rationally holding negative feelings. And exactly this, I contend, is the biggest asset of dropping any appeals to actions having been freely willed. Taking away the "right" to be angry at someone will lead to a more compassionate world. The question "why did you do this" will be meant in earnest; causal explanations will be sought, and then optimal remedies for the situation found.

The negative component of free will is directed toward the *past* (if one wants to express the uncertainty of future actions, again, other words and concepts are more fruitful). The concept of free will, when evaluating past actions, actually entails that the agent in question could have done otherwise, not only in future similar situations, but in that exact *past* situation.

This evokes guilt and blame. While in regard to oneself feelings of remorse and self-loathing may prevail, in regard to other people the result is vindictive behavior: one feels, for instance, the right to punish someone else, even if that would not change her future behavior. Now, in pre-rational times, such emotions may well have been adaptive to enforce morally acceptable actions in society. But the rationalist can do better – again, she need not be unemotional; the rationalist will invest joyous emotions in the optimal pathway she chooses; but she will avoid suboptimal negative emotions.

So, do we have to sacrifice moral accountability[217] when giving up free will, when we reject guilt and blame[218]? Not at all: that would be like saying that gas molecules were only bound to the laws of gravity if they had a free will to do so[219]:

> Instead of treating people as if they were deserving of blame, the hard incompatibilist can draw upon moral admonishment and encouragement, which presuppose only that the offender has done wrong. These methods can effectively communicate a sense of what is right and result in beneficial reform. Similarly, rather than treating oneself as blameworthy, one could admonish oneself for one's wrongdoing and resolve to avoid similar behavior in the future. (Pereboom 2001b)

Responsibility comes with being an agent, not by having free will. But the responsibility I advocate here is of a milder form than full metaphysical responsibility usually invoked to justify moral accountability. Let us have a look at responsibility and where it comes from. I would like to introduce Strawson's basic argument for why we are not *ultimately* responsible. First, Strawson introduces some terminology which I will adopt:

> "R"..."truly and without qualification responsible"
>
> "D" to abbreviate "truly and without qualification deserving of praise or blame or punishment or reward"
>
> "U" to abbreviate "ultimate" when prefixed to a noun and "ultimately" when prefixed to an adjective (Strawson 2001)

217 Incidentally, we are not even free in our moral judgments, which highly depend on situatedness: washing your hands, for instance, makes your subsequent moral judgments less severe (Schnall, Benton & Harvey 2008).

218 As human beings, we will continue to feel guilt and blame; indeed, these are important first evaluations of situations, telling us that something is amiss. The rationalist just suggests to after he has been alerted, he will not wallow in these feelings, but acknowledge their evolutionary function and move on to remedy the situation at hand.

219 A hard incompatibilist asserts that we don't have the kind of free will to account for moral responsibility. I agree with the position of hard incompatibilism.

Strawson then distinguishes RD – normal responsibility – and URD – ultimate responsibility. Strawson denies both, and I think he is right. But here I will simply diverge in terminology and use *responsibility* in a pragmatic, legal, sense, and use *ultimate responsibility* for metaphysical responsibility requiring free will. One form of the basic argument then goes like this (Strawson also presents variants addressing different problems):

> 1.1 When you act, you do what you do — in the situation in which you find yourself [Footnote 7] —because of the way you are.
>
> 1.2 If you do what you do because of the way you are, then in order to be URD for what you do you must be URD for the way you are.
>
> But
>
> 1.3 You cannot be URD for the way you are.
>
> So
>
> 1.4 You cannot be URD for what you do.
>
> Version 1 of the Basic Argument has three premises, 1.1, 1.2, and 1.3. I take premise 1.1 to be obvious and will not defend it. I think that 1.2 and 1.3 are also obvious, but I will give them — or close cousins of them — some explicit defense in due course. (Strawson 2001)

I also view the premises as obvious and will therefore not dwell on them. For a defence, see the cited article above. But the intuition that we can't be ultimately responsible for the way we are is quickly grasped: did you ask that you were born? Who your parents were? How they educated you? Whom you met? Your teachers, yours mates, your friends?

> It is not implausible that good moral character is to a large extent the function of upbringing, and furthermore, the belief that this is so is common in our society. Parents typically regard themselves as failures if their children turn out to be immoral, and many take great care to raise their children to prevent this result. Accordingly, people often come to believe that they have a good moral character largely because they were brought up with parental love and skill. But I suspect that hardly anyone who comes to this realization experiences dismay because of it. (Pereboom 2001b)

This is all backed by empirical evidence: it is well known in psychology that there are *intergenerational effects*. For instance, if we have a look at our simian friends, we find that abusive parenting is transmissible:

> Maternal abuse of offspring in macaque monkeys shares some similarities with child maltreatment in humans, including its transmission across generations. This study used a longitudinal design and a cross-fostering experiment to investigate whether abusive parenting in rhesus macaques is transmitted from mothers to daughters and whether transmission occurs through genetic or experiential factors. Nine of 16 females who were abused by their mothers in their first month of life, regardless of whether they were reared by their biological mothers or by foster mothers, exhibited abusive parenting with their firstborn offspring, whereas none of the females reared by nonabusive mothers did. These results suggest that the intergenerational transmission of infant abuse in rhesus monkeys is the result of early experience and not genetic inheritance. The extent to which the effects of early experience on the intergenerational transmission of abusive parenting are mediated by social learning or experience-induced physiological alterations remains to be established. (Maestripieri 2005)

The way you are now is dependent on people in the past you don't even know! How should we be responsible for things that happened before we were born?

Another problem is that empirical evidence shows that we are neither “good” nor “bad” people (we have differing dispositions of course), but that lack of or display of ethical behavior can depend on seemingly ethically irrelevant situational differences (Doris & Stich 2005).

But, however the world is, we still have to solve the same problems. And there will be agent-actions that upset other agents, and for this we need morality, responsibility and the whole caboodle. But now on a naturalistic footing. A naturalist can anchor responsibility – only in a legal sense, to solve pragmatic problems, not to justify metaphysical vindictiveness – in (limited) agent autonomy, a characterization of which would go along the lines presented by Walter, who calls it “natural autonomy”:

> Freedom of will is an illusion, if by it we mean that under *identical* conditions we would be able to do or decide otherwise, while simultaneously acting only for reasons and considering ourselves the

> true *originators* of our actions. We can do justice to many libertarian intuitions, however, with a neurophilosophical concept of autonomy that includes mild forms of all three components(being able to do otherwise, acting intelligibly for reasons, and being originators of our actions). What remains is a kind of autonomy, that, loyal to the naturalistic approach, I call *natural autonomy*.
>
> Since natural autonomy is not the same as the free will in the strong (libertarian) sense, part of that interpretation *is* lost. Natural autonomy can sustain neither our traditional concept of guilt, for example, nor certain attitudes and hopes about our lives. But we are also not mere marionettes nor puppets without thoughts and ideas that influence events in our lives. The lack of a strong form of free will does not imply that all moral order collapses or that we need abandon every concept of responsibility.
>
> If deterministic chaos should in fact turn out to be a ubiquitous phenomenon within the nervous system, that would explain why we can make different choices in similar situations. It would explain why even in comparable situations we do not always take the same path, how we keep natural alternatives open, and why our thinking is so flexible. It would also explain why the subjective impression of being able to do otherwise seems so irrefutable. Often enough, we do experience a feeling that in comparable situations we would act differently, although we cannot always explain it rationally. Not only can chaotic processes help explain quasi-indeterministic capacities to act otherwise and flexibility; under certain conditions, they also produce stable and predictable behavior. Part of our predictable behavior is presumably within the realm of intermittence — in a realm of order in the midst of chaos. (Walter 2001)

Of course, if we recognize that there is no such thing as free will, we must teach different ethical conceptions; free will seems to be a requirement for ethics appealing to virtuous adherence to moral norms. The following experiment is a bit unfair:

> In Experiment 1, participants read either text that encouraged a belief in determinism (i.e., that portrayed behavior as the consequence of environmental and genetic factors) or neutral text. [sic] Exposure to the deterministic message increased cheating on a task in which participants could passively allow a flawed computer

> program to reveal answers to mathematical problems that they had been instructed to solve themselves. Moreover, increased cheating behavior was mediated by decreased belief in free will. In Experiment 2, participants who read deterministic statements cheated by overpaying themselves for performance on a cognitive task; participants who read statements endorsing free will did not. These findings suggest that the debate over free will has societal, as well as scientific and theoretical, implications. (Vohs & Schooler 2008)

While the last sentence finds me in agreement, I think the experiments do not show us anything important about human morality in relation to "free will"; only about the reactions of humans with certain preconceptions about morality and "free will" and what happens if one selectively picks out and negates one of those preconceptions. The rule underlying the moral thinking of the probands in the experiment seems to be of this kind:

"You should act ethically because some authority commands it and you have been given free will to follow this command."

At least in Western Culture, it is quite common to raise children by connecting morality and free will. People are told – by their parents and by reinforcement through society – that they are responsible for their actions *because* they have free will. Leaving this basic intuition intact and then telling participants of the experiment that they do not have free will, will of course lead to the results described above. A serious experiment would need to examine the preconceptions of the participants and how these relate to the new information presented in the texts.

If you tell human beings about determinism, you should also speak about the signaling function of morality, the responsibility of agent action, the stark reality of qualia states, and the essential non-disconnection, that is, the unified world of which humans are only a part. Only the whole package can be sold without getting into trouble. Accepting that humans do not have free will calls for a *large-scale adaptation of one's internal rule sets*[220].

A quick and dirty rule set which will suffice for most purposes after having been exposed to the fact that free will does not exist:

- I am currently considering an action X
- I am also considering the fact that I am actually not free.

220 Why bother? Because it will lead to more compassion.

- but then, certainly, I am in a highly reflexive state; and this means that I can bring to bear all the ethical rules I have at my disposal.
- I will try to do the most ethical thing I can think of, and I can do this because I am determined (double meaning!) to do this.

If you start thinking about determinism, you already are in an elevated, reflexive state of mind. No excuses remain for not choosing the ethical path. And hurting other people (or yourself: drinking, smoking etc) is never optimal; not with the naturalistic and science-friendly metaphysics proposed in this thesis. These rules have to be practiced, embodied; they have to be taken out of symbol space and neurally instantiated at all levels of brain organization[221]. Diligent practice is the way of the virtuous.

A more sophisticated rule set, which presupposes having read the thesis to the end for full understanding, would go like this:

1. Ethical Premise 1: Agent autonomy is important
2. Ethical Premise 2: There are more and less desirable agent states
3. Responsibility: I am an agent with natural autonomy and thus take responsibility for my actions. Taking responsibility for my action is nothing other than taking into account the real consequences of my actions and factoring these consequences into the state of the universe.
4. Causal Chain: In deliberating about what actions to take in situation X, I am aware that my concrete causal deliberation will lead to certain actions of my agent-self in situation X, and that indeed I can't delegate the computation of possible actions and actual action taken to others, unless I sacrifice responsibility, which I do not want to do because of ethical premise 1.
5. I should choose actions that do not impinge on agent autonomy and which increase more desirable agent states, in accordance with ethical premises 1 and 2.

Responsible agents see oneself as part of the universe who, through their actions and taking into account the real effects of their actions, can continue to behave ethically:

221 That is what Buddhists do in *mettā* meditation; where loving kindness is cultivated toward oneself, family, strangers, enemies and finally to all sentient beings.

> Despite initial appearances, the assumption that we lack the kind of free will required for moral responsibility does not threaten the emotions and reactive attitudes that are crucial to our social life and moral development. Indeed […] the rejection of responsibility-entailing free will holds out the promise of a life that is morally deeper (because less self-centred) and more fulfilling (because less prone to destructive anger).(Trakakis 2007)

Vohs and Schooler (the ones who conducted the cited study above) should have let their empirical results stand for themselves; unfortunately, they choose to foray into philosophy:

> Although the concept of free will remains scientifically in question, our results point to a significant value in believing that free will exists. (Vohs & Schooler 2008)

For a rebuttal, see above. They go on:

> If exposure to deterministic messages increases the likelihood of unethical actions, then identifying approaches for insulating the public against this danger becomes imperative. (Vohs & Schooler 2008)

Well, what if the fiction of free will is a positive one? Should we hide the truth from mortal man? No. This is never an option. A just society can't be built on a lie. The way forward is not to hide scientific evidence from the public, but rather to adjust the conceptions we have of ourselves and the world to the new evidence and live *better* lives with this new evidence. We can leave behind a childish ethics depending on a soul obeying the commands of God and going to Hell if it fails to do so. We can take responsibility for our actions by becoming part of the causal universe, and acknowledging that our causal thoughts shape the future of the real world, the only world that matters. This world is not a stage to prove one's worth for Heaven.

With knowledge comes responsibility; the more knowledge one has, the greater the responsibility; and for those of us who have accepted the *un-freedom* of the will, the responsibility is great; this knowledge requires a dedication to the *well-being of all sentients* to be used wisely; any lesser attitude must lead to catastrophe. That is why I urge that this question – of free will – can only be explored in the right metaphysical embedding. But when this is done, the beneficial aspects of facing nature as it is can't be denied:

> ...there may well be benefits flowing from a thoroughly naturalistic conception of the self and its choices. The retributive impulse, cut

> off from its metaphysical justification in free will, might soften, leading to a less punitive culture. More attention might be paid to improving the social conditions shaping individuals, and no longer will policy makers so blithely blame the victim (remember the multitudes in the U.S. who 'chose' to be homeless during the Reagan era?). On the personal level, dethroning the supervisory 'I' might help us become less self-conscious, more playful, and less likely to wallow in excessive self-blame, pride, envy, or resentment (see Breer (1989)).
>
> Might we become less ambitious, once we see that we don't ultimately choose ourselves or our projects, and that our successes (and failures) result from thousands of combining circumstances? Perhaps, but this might be all to the good, given that the unfettered accumulation of wealth seems likely to compromise the long-term sustainability of resources, or at least concentrate them in a very few hands. And after all, we need not worry that putting the self in its natural, causal context will extinguish desire, any more than we need worry that it will undermine our rights and liberties. Our selves, physically embodied, are virtually constituted by desire, and real freedom lies in having the opportunity to pursue our motives as we discover them arising in us. Seeing that the self neither has, nor needs, ultimate responsibility for itself may well lead to the more responsible use of such freedom. (Clark 1999)

4.1.3.4 The Legal System

> When another person makes you suffer, it is because he suffers deeply within himself, and his suffering is spilling over. He does not need punishment; he needs help. Thich Nhat Hanh, Vietnamese Buddhist Monk[222]

I would like to make a small detour into the legal system, because one objection against denying free will is that the legal system would break down, which is, of course, not true. We have laws to *channel* the behavior of human beings; moral accountability is the assertion that we want these laws to be followed. Morality has a signaling function. Social signals are sent to influence other agents, to direct them to moral behavior; indeed, social signaling, to work, *presupposes* a causal world.

222 Like all good quotes, it's all over the web without a serious source ascription.

Let us look at criminal offense. In an ideal society, we could look closely at every single case, its causal history of origination; understand why the offender *absolutely had to* commit the crime in his place. From this would follow an analysis of how we could school, educate and train this person that he will not commit crimes in the future (special deterrence). Of course, in some persons we will find that no matter of training will change his disposition for criminal offense; these people will have to be detained so as to protect other agents; the justification for detention would follow the same reasoning applied in cases of quarantine:

> A theory of crime prevention whose legitimacy is independent of hard incompatibilism draws an analogy between treatment of criminals and policy toward carriers of dangerous diseases. Ferdinand Schoeman (1979) argues that if we have the right to quarantine people who are carriers of severe communicable diseases to protect society, then we also have the right to isolate the criminally dangerous to protect society. Schoeman's claim is true independently of any legitimate attribution of moral responsibility. (Pereboom 2001b)

Denying free will and advocating determinism does *not* imply excusing crimes or having to live with criminals simply because they can't do otherwise. Now, even in the cases where perfect rehabilitation is possible, there will probably have to be some form of punishment, because this very punishment is *deterministic input* for other agents in *their* deterministic behavioral algorithms (general deterrence). If someone breaks the law, even if un-freely, a response must be made to keep the causal network of action and reaction in place. Giving up accountability would *change* the system humans are living in completely.

Punishment need be enacted to make criminal behavior costly in regard to lawful behavior, thus enticing the agents to optimally choose lawful behavior: that is, defection under detection should always be more costly than cooperation[223], to speak in game-theoretic terms. But there would be no more place for a concept such as "guilt", the secular refinement of "sin". The core concept would be *causal connection* and *responsibility*, in the sense that the autonomous agent, even if not metaphysically guilty or sinful, has certain problems in regard to peaceful coexistence with other members of the society which need be addressed. The details of a such a system would require a separate work in ethics, but there are no problems in principle.

223 We need to take one-time free riders into account, who condition their minds in such a way as to commit only one crime, but then are open for therapeutic intervention so as not to defect anymore.

We have truly understood when we do not judge others on the criteria of their being good or evil, but when we categorize in cause and effect. Then we can rationally decide how to reach good outcomes by influencing other agents. The whole point of not fearing the scientific evidence against free will is that it actually will lead to increased humanization: removal of vindictiveness and more emphasis on healing and therapy.

A strange consequence of the current tension between a scientific causal worldview and the acceptance of "free will" is the following: when confronted with delinquents, those who are obviously different in their evaluative outlook due to pathological brain structures are sent to psychiatric institutions and receive treatment – rightly so, of course – but those who have "normal" brain structures, that is, where less intervention would lead probably lead to complete rehabilitation are sent to prison.

Vohs and Schooler end their paper thusly:

> Or, perhaps, denying free will simply provides the ultimate excuse to behave as one likes. (Vohs & Schooler 2008)

In this thesis, a diametrically opposed sentiment is expressed:

Or, perhaps, denying the *un-freedom* of the will simply provides the ultimate excuse for locking people away instead of giving them benefit of therapy, the former being cheaper, less elaborate, and less challenging on our ethical and intellectual discriminatory faculty.

4.1.3.5 Rationality

> An idealist believes the short run doesn't count. A cynic believes the long run doesn't matter. A realist believes that what is done or left undone in the short run determines the long run. Sidney J. Harris, *Reader's Digest*, May 1979

Some say that without free will we can't be rational, or even worse, we must succumb to fatalism. Both fears are ill-founded, which will be outlined in this section.

Being rational is, as we saw in the first part of the thesis, a way of tracking the world, not access to a realm of the mental disparate from the physical. That is, neural structures in human brains harmonize with developments of world states. Determinism does not threaten the ability to be rational, maximally the ability to be free to adopt rationalism (see fatalism below); if an organism deterministically correctly tracks its environment and acts in accordance with internal models to

fulfill its goals, then it is rational. Even when we go on to analyze our behavior at the micro-causal level (“graininess”) this does not call into question our rationality; “rationality” is simply the ascription of certain properties to large-scale aggregations of matter. Because all our concepts and all our language are at the epistemic level (a process of “registering”), ontological reductionism – the commitment to one world being potent at a micro-causal level – can never threaten our epistemic concepts, so also not rationality.

We can view it this way: thoughts, that feel so causal from the inside, are actually what it feels like if certain brain organizations go through their *physical* causal patterns. Thoughts are neural connections which have become sufficiently dense (they, for example, fire in synchrony sufficiently often) that they become a "causal bundle" – a learned structure, ideally representing something about the environment, which is sufficiently robust to emerge from the subconscious (the latter is the sum of neural connections which do not form into aggregate wholes, that is, they direct the flow of thought without becoming conscious thought). We can of course also call the thoughts themselves causal – if we bear token identity in mind, and see that nothing supernatural is involved; due to much danger of confusion here, I prefer to locate the causal in the *physical description*. Thoughts should be viewed as causality aggregations. In this way we also vindicate the description of the rationalist: being rational means conforming your associative neural net in your brain to the laws of nature, that is: if in the "external" world A is causally followed by B, then the associative net presents B when given A.

Indeed, rationality can even *explain* the persistence of the belief that we have free will. If we are moved by reflex or instinct or unconscious motivations the primacy of the purely causal nature of our behavior is apparent. And this unconscious is necessary:

> From the accumulating evidence, the authors conclude that [...] various nonconscious mental systems perform the lion's share of the self-regulatory burden, beneficently keeping the individual grounded in his or her current environment.(Bargh & Chartrand 1999)

Conflict between physical and mental causation only seems to arise when conscious thoughts come into play, a “higher-order” function of the brain. A part of being rational is constructing plans, weighing their probability of success and the desirability of their outcomes and then *deciding* which one to implement. This latter decision looks like free will; but this decision was of course caused by all prior experiences and genetic factors etc. The illusion of free will arises because we do not have

access to all our brain states – unconscious thoughts etc – so the decision is "simply there" – it seems to have arrived by way of magic, uncaused, but is in fact only the result of the *sum neural computation performed by your brain*. The brain operates on a model of the body from which it draws its sensorimotor grounding. The illusion of separateness of the mind – and thus of free will – comes from the brain not having a *complete model of itself* (qua brain). The brain itself is not modeled; and that is why the mind-body duality and illusion of free will is so persuasive. A nice computational analysis and why the free will illusion[224] is so persistent can be found in Yudkowsky (2008h).

The concept of free will is also useful *heuristic* when reasoning about other people, thus also closely tied to rationality and goes to show why "free will" is such a persistent belief. The brain is a causality aggregator[225]. From the moment where the first cells of the brain are differentiated, the accumulation of causal history begins; of course, this happens in other body parts too, but the brain is the most interesting organ for this analysis because its influence on decisions is more salient, than say, the influence of the cells in your right toe. From the moment the human is conceived, causality begins to aggregate in his cellular make-up. From the moment the human has senses and these senses interact with the central nervous system, causality is also aggregated from more far away sources as the immediate cellular environment (smell: distribution of chemicals in environment; hearing: mechanical air perturbations; sight: photons from across the universe; etc. etc). All these impressions leave their neural imprint, however minute. Again, that is a lot of causality. Of course we can look back to the beginning of life itself; and before that, to the origin of the universe. That may not be very explanatory; but we must keep it in mind when we look at causal histories.

We never fully know what the other person has experienced, what his DNA encodes for etc. That is why humans have invented "free will" from a psychological perspective: we insert "free will" for our uncertainty about the behavior of the other agent. Of course, we can predict the behavior of other people more often than not; that lies in the many commonalities we share, our phylogenetic and environmental history for instance. That it does not always work, that we are sometimes surprised by what others do, or even what we ourselves do – because we can't introspect at the causal level – is due to the perturbations assailing us from all corners of the universe; the chaotic nature of our brain processes; and the vast store of unconscious impressions directing our whims and wishes. "Free will" is a heuristic to *sum over the possible experiences* of the other. "Free

224 See Wegner (2002) for an in-depth account.

225 The aggregation of causality has of course begun before conception via the selection of genes which gave rise to the concrete brain in the first place.

will" is used to designate unpredictability: you will never know what others will do, and maybe not even what you yourself will do ("Free will" is ignorance about oneself). But again, we see that “free will” stands for something else in these cases than SFW, and thus we should use these other words and concepts because “free will” always opens the frame of SFW.

4.1.3.6 Fatalism

> Das Publikum muss denken, dass der Schauspieler auf der Bühne anders hätte entscheiden können. Peter Simonis in Radio NÖ 23/24.01.07

> Moreover, the [...] view that everything that ever has or ever will happen should be regarded as equally real has significant attractions of its own, and ones that are more firmly grounded, philosophically speaking. In fact, the denial of the openness of the future can, paradoxically, prove very liberating. Specifically, those who manage really to take to heart the idea that all events are eternally real will no longer be tormented by thoughts of ‘what might have been’; no longer will they be constantly saying to themselves ‘If *only* I had done such-and such’. For they will acknowledge that at no time are future events anything other than actualities lying in store for us. Any lingering inclination they may have to view their past lives as being littered with missed opportunities and avoidable mistakes will be extinguished by the thought that neither the seizing of the ‘opportunities’, nor the avoidance of the mistakes, ever existed as genuine potentialities. It is, as they will now see it, merely our inability, in general, actually to foresee the future that blinds us to the fact that it is as much a part of reality as are the present and the past. (Lockwood 2005, p. 69f)

A simple example will clarify why "free" is not a necessary word in naturalistic accounts. I define a word ZIP which stands for either *deterministically*, or *somewhat deterministically with random input*, or *purely random* (order emerging in some fashion in the last case). The point of ZIP is to cover all naturalistic cases of state transitions.

Now, we are going to describe a person's decisions with the help of ZIP:

Person A is confronted by environmental state E1. He is in a mind state M1 at which he has arrived by ZIP. Person A now wants to arrive at environmental state E2 because of his mind state M1 encoding certain preferences. His mind at M1 now computes a number of plans of how to arrive

at E2, the computations being effected by A's mind transitioning (via ZIP) into M2, M3, M4 etc, finally arriving at M_N, where he has a number of plans PA, PB, PC available. Due to his knowledge of the world, at which he arrived at by ZIP, he thinks that PB is the best plan, a computation (choice) at which he arrives at again by weighing the plans against each other, by transitioning (via ZIP) through the mind states M_N+1 to M_N+L. Person A now enacts PB and, because he was a good observer and got his correlations with world states right, achieves environmental state E2.

So, we don't need freedom to arrive at choices. But does this lead to fatalism? Fatalism means not trying to achieve your goals anymore. You lean back and say: it's no use.

The best antidote against fatalism is rationality. The rationalist can convince himself easily that even if strict determinism holds, he will still have certain desires (albeit deterministically) and will have to exert effort to achieve the satisfaction of these desires. If you give up *all of* your desires when hearing of fatalism, then you may indeed resign. But I am quite confident that you are not *free* to give up all of your desires.

Let us look again at the example above, in simplified form. It concerns an agent in a world where there is no such thing as free will:

Imagine that the agent starts in state A and wants to arrive at state E. Now let us suppose that to get from A to E it is necessary – strictly causally necessary – to perform the action steps: b-c-d. Now, the agent starts to (deterministically) think about achieving state E, and finds out that he must enact b-c-d. Well then, so *b-c-d* it is. There are no computational shortcuts, that is, the agent will have to perform the actual steps to get to the result.

Fatalism is the erroneous thought that if I am in A, and want to get to E, and it's ordained that I will arrive at E, I needn't bother to transit through b-c-d and instead I can go for a drink. That is not how it works in a lawful universe: here, even if I *know* that I have to pass through b-c-d, I will do it. That is the burden of the autonomous agent. The autonomous agent must actively strive to achieve goals. In a deterministic universe, an agent must become causally active to achieve goals. There is no "kismet" excuse, which in fact presupposes an acausal picture of the universe not in line with naturalism.

Let us draw on a more dramatic example: Your family is caged in a metal box hanging from a rope above a firepit. In front of you is a button, and if you press it, your family will be saved. If you don't press it, the rope will be cut in thirty seconds and the cage will drop into the firepit. Now,

before you is also a TV screen which always shows the world at this place, but twenty seconds in the *future*. In the TV, you see yourself pressing the button and your family is saved. Watching your future self has let you forget the time – twenty seconds have passed; in ten more, the rope will be cut. What will you do? The fatalistic response would be to say it's all no use, and go home (leaving your family to fall into the pit and die). The cheeky and admittedly scientific approach would be not to press the button on purpose to try falsify the video and it's metaphysics of determinism; but I do assume that you love your family, and I think it is quite clear that despite even your foreknowledge of what you will do, and all your scientific avidness, you *will* still press the button. That is actually why the video *could* show your future action: because the video showed yourself having already seen the video and, despite the foreknowledge, pressing the button.

Now, in the real world we usually don't have foreknowledge because matters are too complex and out of our immediate control. So when *even foreknowledge* does not imply a fatalistic stance, but calls us to act on our agent drives, even less so should a situation with *no foreknowledge* imply fatalism. Even if the future is as fixed as the past: there is no alternative to deliberate and perform committed action to achieve one's goals.

Maybe another metaphor will help to avoid falling into the fatalist mindset: think of the cosmos as a symphony. In a symphony, every musician must play the notes of the musical piece before him. Musicians do not lay down their instruments in disgust just because they know what they will be playing in the concert to come. Every single note is necessary to add to the beauty of the symphony. Not one may be omitted. In a symphony, beauty only arises if everybody fulfills his role. The symphony would vanish if everybody could play as he wished. Of course, beauty also arises because of variance from strictly deterministic playing; every musician adds a little personality. This would correspond to the aspect of *zest* and *spunk* in Smith's ontological view of things. But again insofar the universe has these properties, they are properties of the universe of humans only insofar as they are part of the universe. The least contributes to the beauty of the whole. See your actions as part of the cosmic symphony. This picture also removes the threat of the "puppet metaphor" – that there is an independent "you", which, on a previous conception, was free, and under the naturalistic conception is not free. But there never was an independent "you" that is now bereft of it's will and enslaved to the laws of nature. You were the laws of nature in action all along. You are the sum of your experiences, and nature will guarantee that this experience will reflect itself quite deterministically in future actions. Craving free will is actually wanting to being more than you are.

We are not "commanded" by physics to perform our actions; that what we are, what we think, our love and our hate, is simply and integral part of the universe brought forth by physical law as the stone or the tree. Graphically put:

The wrong conception:

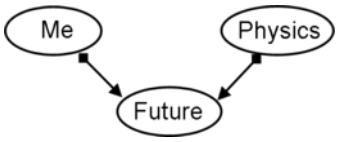


Illustration 3: (Yudkowsky 2008i)

The correct conception:

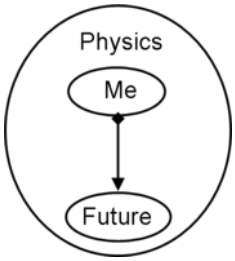


Illustration 4: (Yudkowsky 2008i)

May we still speak about choice in these scenarios? Of course. The will is real. Choice is real (choice as the *implementing of one action out of several possible alternatives*). There is just no sensible way in calling a choice "free". Autonomous, yes – meaning largely independent of current environmental variables – but that's still a far cry from free. Choice, the selection between alternatives can be perfectly deterministic[226]. Let us say you love chocolate ice and you absolutely hate vanilla ice. Now, once a month you eat ice cream, and every time you will choose chocolate before vanilla – you know it yourself; others know it; because it is determined by your taste – but it's a choice nonetheless. That is what choice means: an agent acting in the world according to his preferences.

Saying "Person A chose to do so-and-so" is actually the same (in kind) as saying: "The stone dropped to the floor". It is the lawful universe in action; only in a much more complicated system.

226 The model of Marvin Minsky presented in the next section will elucidate this further.

Imagine not a simple Newtonian clockwork, which may be rather depressing, but a playful brook meandering through a lush valley; sometimes moving swiftly, sometimes languishing in full sun in little pools; or sometimes flying through the air – a joyous waterfall breaking the smoothness of the flow. Be a brook, a river, an ocean; *not a clock*. The scientific conception behind this metaphor is that of deterministic chaotic systems.

When judging other humans we won't get far by looking at micro-dynamics. That is why psychology should (and does) concern itself with belief states, emotions, instincts, reflexes, intuitions, mental models etc and not nonlinear calculations of complex systems.

There is a last difficulty: how about becoming a rationalist? Can we all become rationalists? I think not. If you are currently reading this thesis (and you've made it this far), chances are you are already a rationalist or are on your way to become one – strictly lawfully, of course. But there are certainly people in this world who, given their upbringing and experiences, are not *able* to choose to become rationalists (just as we can't choose our sexual orientation or our athletic ability). That need not be a bad thing in and of itself. There are different way to happiness; maybe they live a good life naturally; but if not, then it means that they will continue to strive for their goals in a suboptimal way, and that is a tragedy. What I hope to do in this work is to contribute that many people opt for a rational way of living: that they understand what is at issue and what is not; that they find the courage to adopt the rationalist way; and that they also adopt a good metaphysics, enabling them to strive for the right goals (that is, living a life filled with value, meaning and love).

Accepting the doctrine of the un-free will or even strict determinism should not change your behavior in any *immediately* noticeable way: it is a minimal reweighting of the meta-knowledge about yourself. Behavioral weights can, on first approximation, stay the same. But in the long run, there will be effects on behavior, hopefully positive ones (see the next section on optimal will). And don't forget: always play as if you *could* have done otherwise.

4.1.4 The Solution: Optimal Will

If we relinquish the concept of *free will*, we need something new to describe our everyday activity – to describe that what *feels* like free will but is not – that delivers a better framing without being awkward. The concept I propose in its stead – is **optimal will**.

In regard to the goal of improving your actions in the future, the concept of optimal will is much more interesting than free will. Optimality is directed toward the future. Past actions are analysed

and checked for sub-optimal outcomes. Goals are re-evaluated, new behavioral strategies thought through and maybe even practiced in safe environments (with friends etc). Reflexivity is schooled. "What did I do? Why did I do it? How would I have *preferred* to perform?" So, in the future, you can act differently under *similar* circumstances. And that is what we all strive for, or at least what we should strive for. The new conception automatically leads to the adoption of *self-improvement techniques*, and is perfectly compatible with a deterministic worldview. Determinism even guarantees improvement[227] because past situations *will* have influence on the future, which free will does not guarantee. Free will, being locating in some metaphysical realm of dubious nature, guarantees neither change nor betterment. It does not supply any cogent mechanism for improving.

Achieving optimal will is a difficult process. And optimality of course, begs for the question of *what* to optimize. Goal-optimizing, yes, but which goals? That means looking for criteria of *eudaimonia*, the good life, and ultimately, wisdom. Goals worth adopting could be enacting love and kindness towards other people[228]. But a closer look at this will have to wait till section 5. Here, let us dwell a bit more on optimality.

For instance, imagine the goal of leading a healthy life. Seeing the body as a machine is the first step to achieving optimality in this regard. The machine model of the body will start letting you think in term of *causal mechanisms* when considering nutrition, regeneration, sleep, exercise etc.

The same holds for the mind. Seeing the mind as a machine will lead to a more rational life[229]. Seeing the mind as a machine will open yourself up to learn every day, to "install" new rules and "delete" bad ones. *Your identity will not be coupled to your current state of perceiving the world, but to the ongoing process of rational improvement.*

Let us look at a model of cognition presented by Marvin Minsky. We will not concern ourselves with the merit of this specific model, it will only serve for illustrative purposes. He models cognition as being made up of a layered hierarchy, from bottom to top:

Self-Conscious Reflection
Self-reflective Thinking
Reflective Thinking
Deliberative Thinking

227 If one is able to – deterministically – adopt the self-improvement goal in the first place.
228 If, for instance, you want more love, compassion and empathy in your life and your interpersonal relationships, you should better start being rational: being rational means you are working towards those goals. Are you fighting with loved ones? Do you do things which you regret? That is suboptimal behavior. Improve.
229 For viewing yourself as a machine see also Turkle (2005, p. 247f).

Learned Reactions
Instinctive Reactions
Minsky (2006, p. 160)

We of course know from our metaphysical excursion that the layered model is not to be interpreted as their being actuals *layers of being*: the brain exists at one (better: *no*) level; the levels above rather represent modes of organization in the brain. The levelist picture captures our intuition that there are processes which are very modular, probably innate and with less degree of sophistication, and other, "higher" processes, which are the result of large scale integrated neural firing and more dependent on learning and experience. But sophistication in reality does not arise by abstraction and building hierarchies – as in human epistemic models – but by non-linear large scale complexes of causal interactions; the most complex are those systems that achieve a homeostasis in the sense that they persist. In the following, talk of levels will be used to designate different modes of organization.

Now, no matter how many levels; or, no matter how large-scale the integration – we are finite beings, so there will be a level/scale where there is *nothing* beyond. That last level is as determined as the rest; but for the cognitive entity who examines "lower-level" thoughts with the "topmost" level it will feel as if the topmost level is the real self and exercises free will, because that level is not available for introspection. Now, the more access the topmost component has on lower level ones – that is, the better the self-representation – the more your behavior is amenable to evaluation and change.

Now that we have put it this way, there is a sense in which we can speak of *freedom*: the more accurate our self-models and environmental models are, the more we know what we want and how we can get what we want, the more free – *free in the sense of the ability to achieve that what we really want* – we become. The more powerful the reflective skills – the building, step by step, of ever more levels of self- and world-representation – the greater the freedom for action. If there is such a thing as freedom – then it is attained by a long an arduous process of personal development; it is not a divine gift everybody has a priori; it is the prize you receive for a life full of virtuous striving. So, ironically, the acceptance of the un-freedom of the will and the modeling of the world in mechanismic detail will lead to more freedom. But, in line with everything said above, I will call this *optimality* and not freedom, to avoid association with SFW.

An example of how adopting mechanismic models (instead of magical free will) can help to achieve goals is the process of *scaffolding* deployed to overcome shortcomings of willpower, which can be eroded:

> Much of the time, what looks like sheer willpower is the result of more-or-less well-orchestrated attempts by individuals to arrange their lives in such a way as to *economize* on willpower, by avoiding situations that call for its exercise. We refer to this as *distributed willpower*, since it involves individuals creating more than one locus of self-control. [Footnote 14]
>
> Self-control strategies can usefully be thought of under four general categories, as part of a progression that involves movement away from the purely psychological toward the environmental. (Heath & Anderson Forthcoming)

There are direct psychological, self-management, environmental and social strategies (Heath & Anderson Forthcoming). Environmental strategies consist of making desirable activities easy to do and undesirable things hard to do via concrete physical arrangements. Social strategies include selective association with certain people, to direct oneself into certain modes of being via group conformity etc.

A concrete example:

> The importance of these environmental strategies can be seen in the phenomenon of "college procrastination" – the fact that college students, particularly during first and second year, experience much higher levels of "problem procrastination" than the general public. What is particularly interesting about this phenomenon is that it has little predictive significance, when it comes to determining work habits in other contexts. [Footnote 26]

From an internalist perspective this is perhaps mysterious, but when seen from the perspective of environmental scaffolding it is entirely unsurprising. College students are given a fairly high degree of autonomy, when it comes to determining a plan of work for themselves, yet they are deprived of all the scaffolding that they have used, in the past, to offload motivational resources. Often they are living away from home for the first time, and so are missing whatever "system" they had developed for the timely completion of tasks. For example, merely studying in the same

location has been shown to decrease procrastination among college students. [Footnote 27] (Heath & Anderson Forthcoming)

So, environmental and social scaffolding can radically change our performance – a fact easily integrated into the view of optimality, but which every proponent of *free will* will have problems accounting for. This of course also raises issues of paternalism – how much scaffolding the public should provide to prevent individuals from making suboptimal choices:

> ... many of the most effective support structures – especially the social ones – are not built by us but built for us, part of an institutional and material heritage. [...] When a traditional institution such as the relatively early "last call" for drinks at British pubs is abolished, it is of course possible for people to institute, perhaps with friends, various strategies for avoiding procrastinating about getting to bed on time. But such arrangements are typically effortful and fragile, relative to taken-for-granted structures. To take just one example, consider the sleep-deprivation that has become a source of complaint in our society, which can plausibly be attributed, at least in part, to a tendency to procrastinate going to bed on time. [Footnote 33] It used to be the case that TV stations would end their broadcasts at around midnight, bars and restaurants would close, subways and buses would stop running – the clear message being sent was: "time to go to bed." These institutional arrangements also made it much easier to go to bed on time, since there was little else to do after a certain hour. Now individuals must exercise more self-control about when to go to bed. (Heath & Anderson Forthcoming)

So, to wrap things up: the insight that we don't have free will – is it for everyone? I think yes, but only if embedded in the right metaphysics. For people who don't see themselves as foreigners in this universe but at home here – being an *integral part of reality ("no-essential-disconnection")* – a reality which is constituted by moments, so that every moment already has some intrinsic value simply be being part of the real world – there is neither the danger of falling into fatalism nor into meaninglessness.

Freedom lies in the universe – in the universal exploration of all living and un-living moments; the laws of nature are *playing out* the freedom of the universe. Our part of universal "freedom" is our ontogenetic history, our specific becoming. If we can love our physical becoming and what it

represents, then we can love that this is what determines our future actions. We can let ourselves fall into the world. By realizing all this, we can, ironically, be freer than before. And we can maximize this freedom by trying to achieve *optimal will.*

4.2 The Self

Having divested us of free will, the next assault on our self-conception will be even more difficult to accept, at least to Western ears: the conception of the self as something of substance over and above informational patterns. Again, this does not reduce our value. Our value derives from the value that is inbuilt into the universe: qualia states, which encode value directly and which can't be reduced to functional descriptions on pain of confusing epistemology with ontology. But more on value in section 5.5.

We have the naive and stubborn feeling of a stable "I" to which our experiences belong. An idea of Drescher brought forward in a slightly different context – qualia – will let us take a first step to understanding what the "I" or the "Self" could actually be, if it is not a substance. Drescher likens qualia to certain symbols in the programming language LISP:

> In the computer language Lisp, a gensym (short for generated symbol)is an object that has no parts or properties, as far as Lisp programs can discern, except for its unique identity — that is, a Lisp program can tell whether or not two variables both have the same gensym as their value [...] A Lisp program cannot examine whatever internal ID tag distinguishes a given gensym from any other; yet the program can tell whether or not something is indeed the same as that gensym. Similarly, we have no introspective access to whatever internal properties make the red gensym recognizably distinct from the green; Drescher (2006, p. 81)

I do not know if it is helpful to understand qualia in this way. But we can use the account of Drescher to understand the representational structure of the "I" as a first approximation. The "I" symbol – that is, that what is activated in one's brain when one thinks "I" – fits the gensym description nicely.

The "I" symbol is a non-descriptive reflexive designator needed to performs certain cognitive functions, but which need not refer to anything particular[230]:

230 See also Sloman & Chrisley (2003); Pollock (2008).

> We investigate the functional role of "I" (and also "here" and "now") in cognition, arguing that the use of such non-descriptive "reflexive" designators is essential for making sophisticated cognition work in a general-purpose cognitive agent. If we were to build a robot capable of similar cognitive tasks as humans, it would have to be equipped with such designators.
>
> Once we understand the functional role of reflexive designators in cognition, we will see that to make cognition work properly, an agent must use a de se designator in specific ways in its reasoning. Rather simple arguments based upon how "I" works in reasoning lead to the conclusion that it cannot designate the body or part of the body. If it designates anything, it must be something non-physical. However, for the purpose of making the reasoning work correctly, it makes no difference whether "I" actually designates anything. If we were to build a robot that more or less duplicated human cognition, we would not have to equip it with anything for "I" to designate, and general physicalist inclinations suggest that there would be nothing for "I" to designate in the robot. In particular, it cannot designate the physical contraption. So the robot would believe "I exist", but it would be wrong. Why should we think we are any different? (Pollock & Ismael 2004)

These accounts are strongly simplified. Additionally, it is not at all clear that a designator such as "I" is really necessary – see below[231]. How an "I" can come about is detailed in Metzinger (2003), which is too technical for this section, and can thus be found in Appendix C: The Self. The self can also be conceptualized as a dynamical system, incorporating more or less of the environment, but this also leads us too far astray[232] (Gelder 1998; Rockwell 2005). It suffices here to know that reductive computational/representational accounts of the "I" are possible; we can conceive of the "I" as a representational structure. In spite of the computational wording of the above, these insights are old; see Appendix D: A Question of King Milinda for a beautiful traditional passage.

We see that there is something to the "I" – but it is not what it seems to be: it is, as the Buddha well knew, an illusion. If Eastern accounts are to be believed, there is merit in seeing through the illusion, although, again, following through behaviorally is more a meditative task than one of

231 I believe that it is necessary but has been *evicted from symbol space* in the case described below.
232 I would especially like to endear the book by Rockwell (2005) to the reader.

theoretical insight. Seeing that there is nothing of import tied to the self, one can become less selfish and less fearful:

> If there *could* be a [...] way for the sense of self to be lost, but with the consciousness remaining unimpaired, then this would compel the philosopher of mind to think about consciousness and its capacities in a new and different light.[...] On this matter, the Eastern accounts of how *Arahants* act in the world are telling. Not only are such persons reportedly able to 'get by'; their ordinary actions – those normally associated with autonomy and survival – are depicted, without exception, as proceeding more effortlessly and efficiently than the comparable actions from persons *with* a sense of self. From an ethical perspective, their conduct is invariably described as exemplary in its virtue, wisdom and compassion – exceeding, even, Aristotle's *phronimos*. Should these reports be correct, they would raise a plethora of questions (Albahari 2006, p. 210)[233]

In a similar vein Parfit, who also arrives at the conclusion that there is nothing of substance to the self (and to personal identity, more on that below):

> It makes me less concerned about my own future, and my death, and more concerned about others. I welcome this widening in my concern. (Parfit 1984, p. 347)

And especially concerning death:

> Instead of saying 'I shall dead', I should say, 'There will be no future experiences that will be related, in certain ways, to these present experiences'. Because it reminds me what this fact involves, this redescription makes this fact less depressing. (Parfit 1984, p. 281)

The *self*, we must remind ourselves, is a conceptualization like any other: if we imbue this concept with ontological significance, as is done when conceiving of the self as a "soul", we have to pay a price – in this case, one pays with *separateness*. The reductionist account of the self is actually giving back the concept and collecting the refund – *unity*.

233 An Arahat or Arahant has achieved liberation by destroying attachment, hatred and delusion.

In case you find the views concerning self and identity expressed above and below disturbing, there are traditions which have ample experience in dealing with this way of viewing things (Smullyan (1977); Goldstein (2002) are possible first ports of call).

4.3 Identity[234]

What about our identity? We are agents: our agency captures our embodiment and our personal histories, reflected in beliefs, desires, goals, emotions etc. We have discarded the notion of a substantial self. What remains are fluctuating patterns of perceptions and reactions to perceptions; an identity always constructed on the fly from memory, from the body, from the environment. This follows already from all of the above: naturalism, causal structural realism, monism etc. That is *ontological reductionism,* contrasted with the supernatural view that we have Cartesian egos or souls. Noonan distinguishes the *Complex View* from the *Simple View*:

> The proponent of the Simple View of personal identity will say that personal identity is an ultimate unanalysable fact, which resists definition in any other terms. By contrast a proponent of the Complex View will maintain that an informative account of what personal identity consists in is possible, since personal identity is nothing over and above those observable and introspectable facts of physical and psychological continuity which provide the only evidence for it. Again a proponent of the Simple View will say that persons are 'separately existing' entities, distinct from their brains, bodies and experiences, whilst a proponent of the Complex View will say that persons are nothing 'over and above' their brains, bodies and experiences. (Noonan 2003, p. 93)

I follow the Complex View. What are we, actually? Material patterns in motion. In a naturalistic universe, everything changes – we, like the whole universe, conform to the laws of physics, and they guarantee that we always undergo gradual transformations. True isolation is impossible in this universe. Real things change, and real persons change. Our beliefs, our goals, our desires, our emotions, yes, even the qualia we experience change[235]. Is there anything that stays constant? I think not – there are only varying degrees of constancy which may create the illusion of permanence.

234 This section can be skipped, as it contains nothing that does not already follow from our metaphysical commitments. I only want to show connections with the literature on identity. And, of course, reading will add to further understanding of what the metaphysical commitments entail.

235 For instance our feeling of wholeness and integrity increases with age.

Even in our short lives, we undergo considerable transformations. The old man at the age of ninety has only rudimentary similarity to the teenager of eighteen; maybe not much more in common than some memories and a scar acquired at an early age. For people living active intellectual lives and *seeking* the process of transformation – of improvement – transition times into new personality structures may be much quicker. Personal identity is a fickle thing. There are various conceptions of personal identity floating around in the literature. Staple points in every discussion of (diachronic) personal identity are examples of fission and fusion, that is, the splitting or merging of persons. As neither fission nor fusion constitute a problem for the naturalist view propounded here, I will not dwell on the subject.

For practical matters – to judge, for instance, who is who over the course of normal time periods, we still need a criterion of personal identity. For a detailed argument concerning matters of identity, see Parfit (1984)[236]. Parfit considers different degrees of physical and psychological connectedness. They constitute the *physical spectrum*, and the *psychological spectrum*. We can traverse the physical spectrum by replacing body parts of a person. This happens every day in yourself. Some body parts are shed (hair, nails, skin cells etc), others are built anew out of the food you eat. The atoms in our body are constantly changing via metabolism. What stays roughly the same is the pattern. We can imagine more extreme cases of physical changes like teleportation[237], complete with malfunctions such as destruction of the original or multiplication of the original; or diverse forms of surgery. Then there is the *psychological spectrum:* here we can imagine varying degrees of connectedness via memories. Parfit argues that neither in the physical spectrum nor in the psychological spectrum – both combined are the *combined spectrum* – is there are *sharp borderline* where are person becomes a *different* person. There is no *deep fact*[238] about personal identity – no fact over and above similarities and continuities constituted by the physical or psychological facts. Given our investigation into the nature of the self, this should not come as a surprise.

Roughly corresponding to physical connectedness is *animalism* in the literature on personal identity. Personal identity is construed as coinciding with biological continuity:

> if X is a person at t1, and Y exists at any other time, then X=Y if and only if Y's biological organism is continuous with X's biological organism (Olson 1997; DeGrazia 2005). Note that Y may or may not be a

236 Another excellent book is Wilkes (1988).
237 As in Star Trek.
238 This is Parfit's terminology.

> person, which allows that X might be one and the same as a fetus or someone in a PVS. (Shoemaker 2005)[239]

But here we see immediately that identity is indeed not what matters: while under the conception of animalism you are identical with your comatose self, you yourself are more interested in psychological continuity:

> instead, what matters has to consist in psychological continuity and/or connectedness (what Parfit calls "Relation R"). As long as that relation holds between me-now and some other person-stage — regardless of whether or not it holds one-one — what happens to me is just as good as ordinary survival. Call this the Identity Doesn't Matter (IDM) view. (Shoemaker 2005)

In the real world, of course, psychological continuity usually only occurs when animal continuity is also the case. Psychological continuity without animal continuity becomes interesting when you believe in the ability of *uploading*, that is, the transfer of your mind into a different substrate. More on that below. The intuition behind psychological continuity is correct – it is indeed what matters.

How devastating the loss of memory – and thus the disruption of psychological continuity – is to personal identity is illustrated in Sacks (1985). Sacks describes a patient, Jimmie, who developed severe retrograde amnesia due to heavy drinking[240]. Jimmie was around the age of fifty when admitted to Sack's clinic. All his memories from his present age back to when he was twenty years old had been "erased" by the disease. An especially chilling episode occurs during the first interview of Sacks with Jimmie. Sacks recounts:

> Jimmie was a fine-looking man, with a curly bush of grey hair, a healthy and handsome forty-nine-year-old. He was cheerful, friendly, and warm. 'Hiya, Doc!' he said. 'Nice morning! Do I take this chair here?' He was a genial soul, very ready to talk and to answer any questions I asked him. […]
>
> 'And you, Jimmie, how old would you be?' Oddly, uncertainly, he hesitated a moment, as if engaged in calculation. 'Why, I guess I'm nineteen, Doc. I'll be twenty next birthday.' Looking at the grey-haired man before me, I had an impulse for which I have never forgiven

239 PVS is a permanent vegetative state.
240 Korsakov's syndrome.

> myself — it was, or would have been, the height of cruelty had there been any possibility of Jimmie's remembering it. 'Here,' I said, and thrust a mirror toward him. 'Look in the mirror and tell me what you see. Is that a nineteen-year-old looking out from the mirror?' He suddenly turned ashen and gripped the sides of the chair. 'Jesus Christ,' he whispered. 'Christ, what's going on? What's happened to me? Is this a nightmare? Am I crazy? Is this a joke?'— and he became frantic, panicked.
>
> 'It's okay, Jimmie,' I said soothingly. 'It's just a mistake. Nothing to worry about. Hey!' I took him to the window. 'Isn't this a lovely spring day. See the kids there playing baseball?' He regained his color and started to smile, and I stole away, taking the hateful mirror with me. Two minutes later I re-entered the room. Jimmie was still standing by the window, gazing with pleasure at the kids playing baseball below. He wheeled around as I opened the door, and his face assumed a cheery expression. 'Hiya, Doc!' he said. 'Nice morning! You want to talk to me — do I take this chair here?' There was no sign of recognition on his frank, open face. (Sacks 1985, p. 23f)

Jimmie lives in an "eternal now", being reset every few minutes. Jimmie, being alive now, in effect died at the age of nineteen. But he did not die *when* he was nineteen, because he lived a relatively normal life – albeit with alcohol abuse which led to his disease – until his *mid forties*, when his retrograde amnesia started to develop. So, here was a human being who had lost half his life because his memories had been erased, or were inaccessible at the least.

What we value in ourselves are our relationships, our things, our *anchoring* in the world – but this is mediated through our memories; if a disease such as the above, or others such as Alzheimer, destroy our memories, we *die* despite still being bodily alive.

So, let us have a closer look at the psychological continuation criterion, adopted here in the variant explicated by Noonan (2003). But first some terminology: the *Only x and y principle:*

> This is the principle that whether a later individual y is identical with an earlier individual x can depend only on facts about x and y and the relationships between them: it cannot depend upon facts about any individuals other than x or y. Otherwise put, what the principle asserts is that whether x is identical with y can only depend upon the

> *intrinsic* relationship between them, it cannot be determined *extrinsically*. (Noonan 2003, p. 127)

We need the *Only x and y principle* to exclude some versions of psychological continuity which restrict identity when fission takes place. As mentioned above, fission, if possible, poses no problem for personal identity in a naturalistic account and thus is ignored in the present thesis. But we need to address it here simply for completeness sake. The psychological continuation criterion then goes like this:

> what is crucial for personal identity is neither identity of body nor brain, but psychological continuity, in the wide sense which includes other continuities as well as continuities of memory. Where I disagree with such psychological continuity theorists as Shoemaker and Parfit is in my adherence to the Only x and y principle, and my consequent rejection of any 'best' or 'no rival candidate' version of a psychological continuity account of personal identity. The crucial difference is that I am committed to saying that any sufficiently strong line of psychological continuity represents the history of some person irrespective of what fissions or fusions have taken place, or will take place. (Noonan 2003, p. 210f)

Now, psychological continuity is of course mediated by the brain (for differing opinions see again below). This leads to the question: *is* the person the brain? Many people think so, and this is certainly true in academia, where personal identity is strongly coupled to intellectual achievement. But what about a body-builder? A super-model? An athlete? Take away their bodies – what they have worked for years to shape with great effort – and you take away much of their identity: so there is also a strong sense in which the body qua body contributes to identity, quite apart from brain processes. It all depends on what matters to your conception of yourself in the first place; identity in this sense is agent-relative; what would constitute identity for one agent's self-conception may not hold for another's. Burwood (2009) goes even farther and considers bodies getting a "brain-transplant". That view is too extreme[241]; but the body certainly has a great influence on the human being, so that a new body for an "old" brain would most probably constitute a new person, for all practical purposes.

Let me wrap everything up:

241 There is no narrative to lament the loss of the brain, only a narrative to lament the loss of a body.

Bodily continuity ends when we die. But even psychological continuity is no guarantee to stay the same person in the long run. Imagine that we can live for thousands of years. To remain cognitively stable, it would be necessary to perform constant neural pruning so that we do not drown in associative memory deluges. In other words, we must forget, lest we suffocate in memories[242]. Let us meet another of these long-lived people, Mr. Smith. We speak with him extensively and get to him quite well. Now, let us say that ten thousand years have passed, and we meet Mr. Smith again. Mr. Smith has experienced things undreamed of by us short-lived homo sapiens. He has read books; written books; enjoyed art and performed art. He has lived through hundreds of relationships. But there is a causal history from the first Smith to the second one. He never died. Every night, he went to sleep and the next day he woke up again. He possesses animal and psychological continuity.

Despite this, I think we can confidently say that the older Mr. Smith will not in any way be the same person as the younger Mr. Smith. Not in any sense as we use the word person. If we could meet the two Smith's simultaneously, we would not recognize any similarity between the two. If there is no soul carrying some *meta-person* responsible for essentialist personal identity – and the naturalist does not believe in souls – this leads to a gradual[243] degradation of personal continuity: at some point (this is not a specific point in time), divergence of personal characteristics is so huge that we will not speak of the same person (this is movement along Parfit's psychological spectrum). It is happening to us right now: we change all the time; some people accept this and carry on their companionship with us; others say: *this one has become a different person, I want no dealings with him*. It depends on what is important to people; but there are certainly changes where everybody would agree that personal continuity is not present anymore.

We can construe personal identity as largely parallel to bodily identity only because we are a short-lived species, and most of us do not change significantly after having stabilized as mature personalities. But identity has nothing to do with physical continuation. As shown above, *physical continuity does not preclude personal discontinuity.*

That does not mean that bodily continuity is not necessary: our memories constitute who we are; but we can't extract our memories from our brains like blood from our arteries – memories are too much causally imprinted into the neural net which constitutes our selves. *In this sense, then, our bodies – our brain being part of our body – constitute who we are.*

242 Read Borges (1942) for a fantastic account.
243 Sorites-style. See below.

4.4 Material Beings

4.4.1 Q-Beings, N-Beings and P-Beings and Ethics

Qualia, as argued above, are an ontic and perspectival property of dispositional structures. But if we make this move to solve the mind-body problem, another problem opens up: we must decide which qualia bearing structures can be targets of ethical concern.

We can discern two criteria interesting for morality: agency/personhood[244] – I will use the terms largely interchangeably in what follows; and the ability to have qualia of a certain sophistication, most notably the ability to feel pain, but also other states such as boredom, longing etc.

Our desire for ethical behavior from others stems from *our feelings* (qualia) – we feel pain, disillusionment, sadness; but also happiness and joy. If it wouldn't be *like* anything for us, there would be no incentive for ethics and morality. In short, one could say: *no qualia, no ethics*. But in a panqualicist universe that is not a very exclusive claim. Of course, a rock is evidently not the concern of ethics (but see Foridi's information ethics (Floridi 1999; Stahl 2008). I will use the term *Q-Being* for an entity which is *sophisticated enough* to be eligible for ethical considerations[245], where Q-Being is short for Qualia-being. What is a Q-Being is an empirical question in the end, and, needless to say, much care must be exercised given the momentous consequences of such evaluations. Some possible criteria for evaluating Q-Beingness are presented below.

First of all, a strong indicator is similarity to human beings. Indeed, we infer that other people are conscious from ourselves being conscious, and other people being similar to ourselves. The same most probably holds for higher mammals. The farther removed organisms are from our make-up, the more difficult it will be for us to reliably judge what they feel and of what they are conscious of. We will have to fall back on behavioral criteria or general assumptions about what makes entities conscious.

244 There can be no morality if there is only one entity (only a God), or no entity at all (a lifeless universe). The question of ethics may very well apply to those scenarios – albeit in a form not interesting to most humans. Ethics has a broader focus than morality; ethics is about inquiry into questions of meaning, value and moral conduct in the broadest possible sense; ethics asks question about a good life per se. Morality has a narrower focus; here, I will use morality to designate the evaluation – in a normative sense – of behavior. Morality comes into play as soon as agents interact with one another. We still need qualia – agency is no substitution for that; for qualia-less agents – such as simple software agents – interaction rules are simple *coordination* or *transaction* rules; rules of protocol. To make norms between agents *moral* norms, I contend that we need *qualia*.

245 This is not a work of ethics. That is why I do not explore the question of when a being is sophisticated enough. I do not know the answer anyway. What is enough for the present thesis is that the naturalist conception of things is ethically more expansive – see below – than traditional (religious) conceptions.

We will sometimes be wrong, but the direction is clear: we will be *more expansive* in our assignment of ethical value to other beings than we are today. Debates on ethics which deviate from the current dominant view have the peculiar nature of deteriorating quickly[246]. To be especially clear, I would like to define a c*lausula expansiva:* whatever the current state of affairs is, the ethics endorsed by the present metaphysical view wants to *expand ethical concern, minimize suffering and generally improve the quality of the living experience of all living sentients*. If you read what I write in a different way, there must be some misunderstanding involved.

An example for the expansiveness is the inclusion of animals: when we feel pain – one of the primary qualia serving as an example here – this is an evolutionary adaptive trait – it is feedback at the organizational level of the central nervous system about remote bodily events. The organism should change it's behavior. It would be very astounding if pain qualia should suddenly spring into existence *fully formed* in the human brain, and animals just behaved *as if* they had it; the conjecture that pain *evolved* in central nervous systems having the right categorical/dispositional properties and was kept because of its adaptiveness is much more plausible. But if animals feel pain and other qualia – why should they be exempt from moral considerations? Why should it not be forbidden to cause unnecessary harm to an animal? The same as we have overcome tribalism and and racism (at least in theory), we should overcome speciesm (Singer 1993; Bernstein 2004). From a Buddhist perspective, this is nothing revolutionary:

> For Buddhism, humans are a part of the community of sentient beings in a conditioned world where suffering is endemic. Humans are not seen as set over non-human nature as 'stewards', but as neighbours to other, less intelligent, sentient beings. The spiritual potential of humans means that they are to be more valued than members of other species, but that very potential is expressed and enhanced by compassionate regard for any being. To kill or harm another being deliberately is to ignore the fragility and aspiration for happiness that one has in common with it. When it comes to indirectly causing harm to sentient beings, Buddhism's emphasis on an ethic of intention means that such actions are not necessarily blameworthy. Yet its

246 Especially in German speaking countries ethical debates get out of hand quickly:

> In a country where memories of the Nazism still haunt the national psyche, some questions have become taboo, tainted by their association with the Third Reich. One such taboo is euthanasia and one professor who dared write about it is Norbert Hoerster [...]. The result was a hail of protest leading to his resignation. (Niemann 1999)

Niemann aptly calls *reason* a victim of Nazi Legacy. Singer includes an Appendix: "On Being Silenced in Germany" in his "Practical Ethics". (Singer 1993, p. 337-359)

> positive emphasis on compassion means that the removal of causes of harm to beings is praiseworthy. (Harvey 2000, p. 185)

There a different theories of how consciousness as we value it arises[247]. The most compatible with the general gist of this work are given by Edelman, Tononi and Koch (Edelman & Tononi 2000; Koch 2004; Tononi 2004; Koch & Tononi 2008; Tononi 2008).

In all these approaches, some kind of integration is performed; Edelman assigns consciousness[248] to reentrant processes in the dynamic core – which is mainly the thalamocortical system sending reentrant signals to itself. Reentry is used here in the following sense:

> Reentry is a dynamic process of ongoing spatiotemporal correlation occurring between functionally segregated neural areas that is mediated by signaling through massively parallel, reciprocal fibers (Edelman & Gally 2001)

Tononi gives his approach – the integrated information theory of consciousness (IIT) – a more information-theoretic[249] coloring:

> According to IIT, consciousness implies the availability of a large repertoire of states belonging to a single integrated system. To be useful, those internal states should also be highly informative about the world. (Koch & Tononi 2008)

and:

> … IIT, is grounded in the mathematics of information and complexity theory and provides a specific measure of the amount of integrated information generated by any system comprising interacting parts. We call that measure Φ and express it in bits. The larger the value of Φ, the larger the entity's conscious repertoire. (Koch & Tononi 2008)

I would like to stress the following: the approach detailed by Tononi et. al. seems to be functional, because it measures the repertoire of states of a system and the degree of integration the system is able to perform. But in the above quoted passage, we also read "those internal states should also be highly informative about the world"; again we face the problem of how meaning

247 These are all "easy" problems. The "hard problem" has been answered by the commitment to monism.
248 Edelman sees consciousness as an integration of many different qualia.
249 For the relevant mathematics, see Tononi (2008).

enters the systems. I think that the criterion of information integration given by Tononi is necessary, but not sufficient for consciousness. We also need what I call *deep integration*. More on that below.

One can hope that methods such as the one developed by Tononi will be refined enough in the future that one can – together with theoretical assumptions – directly evaluate a species for the presence of sophisticated qualia states.

Q-Beings are those entities in the universe which can be and should be the *subject* of ethical concern; in Kant's words: a Q-Being is an entity which is an *end in itself*. Consciousness and qualia are not a prerogative of humans but result from certain matter configurations. Other possible candidates for Q-Beings are aliens, and, given the restrictions[250] mentioned in the next section, Artificial Intelligences (AIs). I have already mentioned the most obvious contender for ethical concern, animals[251].

That is not to say that agents without qualia can never be the target of moral consideration. In a complex world, things are never so clear cut. In human society we will need a derivative ethical status from agency – or potential agency – alone. There are two cases imaginable: reduced qualia, and no qualia.

If agents have *reduced* qualia, it would nonetheless be unacceptable to interfere in their autonomy. For instance, there exists the (fortunately) rarely occurring affliction[252] of being unable to feel pain. You could argue that there would be no moral commitment to abstain from inflicting injury on these humans, because they would not feel pain. But their autonomous agency would be compromised; even if they would not feel pain, injured body parts are detrimental to their agent status. They would impede or at least aggravate the attainment of agent goals. Of course, the inability to feel pain does not make these people unable to feel happiness, sadness and other emotions. So, even if aspects of agency creep in, we can still ground ethical concerns in other qualitative agent states.

250Concerning AIs I recall the materialist doubt, but, just to be perfectly clear, not any "soul-doubt". AI's, if possible, will have to be materially created, not programmed on traditional sequential/linear hardware.

251 See Kemmerer (2006) for an introduction.

252 Congenital insensitivity to pain with anhidrosis (CIPA)

If there is a *permanent total absence* of qualia[253], such as in brain dead patients[254]; they would still be objects of ethical consideration because of their families. Quite differently is the evaluation of coma-patients: here, an ever so slight possibility of regaining qualia makes them ethical subjects of course (see N-beings below). To be especially clear on a sensitive topic again, I would like to insert a quote which I fully endorse:

> ...soll hier unmissverständlich festgestellt werden, dass aus evolutionär-humanistischer Perspektive *jeder Mensch von Geburt an – und dies ungeachtet seiner geistigen Kapazitäten! – das uneingeschränkte Recht auf Leben (incl. der damit einhergehenden Menschenrechte) besitzt*, das nur in extremsten Sonderfällen (Notwehrprinzip, Tyrannenmord) in Frage gestellt werden darf. *Dieses unbedingte Recht zum Leben bedeutet jedoch keineswegs eine "unbedingte Verpflichtung zum Leben".* (Schmidt-Salomon 2006, p. 127)

A being which is the subject of ethical concern is a bearer of *rights*. Agency and personhood also makes one a bearer of *obligations* – only persons can be targets of moral norms. So not everybody has the same obligations and the same rights[255]. At the moment you have qualia, you have a certain set of *minimal* rights. At the moment you have agency, you have an increased set of minimal rights and *obligations*, among those the duty to behave morally. Let us call a Q-Being that is a person a *P-Being*, short for Person-Being. How to differentiate P-Beings from Q-Beings? First of all, in the real world there will always be cases where there is no sharp distinction:

> Contemporary thought and experimental studies indicate that there is no sharp dividing line between persons and non-persons. For example, if newborn babies are not persons because they lack the capacity for self-consciousness and rational thought, then there is no exact point at which they become persons. Likewise, if we contend on empirical

253 There is a case imaginable where there are *no qualia at all, but agent status* is present. Chalmer's philosophical zombies (Chalmers 1996a) are beings of this kind: they are ill equipped to be subjects of ethical considerations – they do not feel love, pain, passion, kindness, empathy and what have you.

254 As we see here, there exist organically human beings, possibly being kept alive with machines, which are not Q-Beings anymore. The set of human beings is neither identical nor a subset (nor a superset, to make things complete) of possible Q-Beings.

255 Persons are also bearers of *additional* rights to those of a Q-Being who is not a person. A core right of all persons should be self-ownership (More 1997). With increasing aptitude, rights may also increase, but only slightly. Usually, such rights will be those which come with great responsibility, such as to wield a technology for which it is necessary to be very diligent. Such rights may be acquired by taking diverse tests and for which thus not all will qualify to pass. We already have this situation today: the right to drive a car is associated with the ability to pass the test for a driver's license. Not every person will be able to pass this test. Obligations, on the other hand, will rise correspondingly with rights: consider the measure of negligence: this measure will be quite different for different sentients (imagine super-intelligences being able to monitor more parameters than we humans are able to).

> grounds that chimpanzees, gorillas and dolphins (or some other species of being apart from humans) are persons then we might admit also that there are some other species about which there will be no clear answer to the question: Are some adult members of this species persons or not? For example, perhaps, orang-utans fit into this grey category of being neither clearly person nor clearly non-person. This indicates that there is no sharp moral dividing line between persons and non-persons. (Thomson 2008)

But we can give a distinction for most practical purposes. An old blunder by Aristotle's will helps us in this because it has drawn the attention of philosophers, leading to work which will come in handy. So, how to decide who is a P-Being and who is not? Aristotle distinguished between persons and *natural slaves*, which could be likened to Q-Beings that are not P-Beings:

> We should, then, state baldly Aristotle's argument on *behalf of slavery* (**BS**):
>
> 1 Slavery is just if and only if there are natural slaves.
>
> 2 There are natural slaves.
>
> 3 So, slavery is just.
>
> Importantly – and this is yet another reason for facing Aristotle's brief for slavery head on rather than indulging in cultural apologetics – it follows directly from (**BS**-1) that the enslaving of those who are not natural slaves is unjust. As he says, 'No-one would say that someone is a slave if he did not deserve to be one' (Pol. 1255a24–25). Consequently, since there are in fact no natural slaves, slavery is, by Aristotle's argument, unjust.
>
> Or are there natural slaves? What can be said on behalf of (**BS**-2)? It is in fact a bit difficult to determine who Aristotle takes the natural slaves to be. We know they lack deliberative faculties (Pol. 1253b9–32, 1254a10, 1255b36–37). Happily, they are not Greeks, but are drawn from the inferior barbarian hordes. Perhaps the suggestion should be that as a matter of fact half the Greeks, the male half, all arrive on the face of the earth with sound deliberative faculties, which would preclude their being enslaved. Why mental faculties should

> be distributed thus unevenly is nowhere explained – or explicable. (Shields 2007, p. 371)

Carol Rovane has given an apt response to Aristotle by her definition of what constitutes a person, and I will adopt her definition for my P-Beings. For Rovane, a person is an entity that possesses full reflective rationality:

> … something cannot qualify as rational unless, in addition to conforming to the requirements of rationality, it also grasps those requirements and apprehends their normative force. In other words, a rational being must see that it ought to be rational. [Footnote 1]
>
> Let us call this kind of rationality full reflective rationality. It is the kind that persons possess. […] Being reflective, persons can inspect their own thoughts and actions and evaluate the extent to which they do and don't conform to the requirements of rationality. When they don't conform, they can respond to such rational failure by engaging in self-criticism and efforts at self-improvement. These self-critical activities show that persons are committed to being rational, and it is by virtue of this commitment that they can qualify as rational even in the face of rational failure. [Footnote 2] (Rovane 2004, p. 321)

So, P-Beings are Q-Beings who possess full reflective rationality. In what way is this a response to Aristotle? Because with this definition people can only deny *other persons* their status as persons in an obviously hypocritical way:

> … rational modes of influence are ubiquitous in interpersonal relations. In fact, they are the distinguishing mark of all distinctively interpersonal relations. Whenever persons treat one another specifically *as persons*, they are engaging one another's points of view and interacting with them from within the space of reasons. We have just seen that this is so even when persons are disrespectful and abusive of one another. Even in such cases they are usually treating one another as persons as opposed to mere things, precisely because they are appealing to and exploiting their common rational nature. […]
>
> There is no doubt that to embrace this definition and equation is to take a somewhat exclusionary view of persons. We would deliberately

> exclude from the kind "person" anything that we can't treat in distinctively interpersonal ways—for example, fetuses, the severely insane, the irretrievably comatose, and the hopelessly senile. But if this seems unduly exclusive, consider that the definition is highly inclusive as well. It entails that if we can treat something *as* a person, then it *is* a person. And herein lies its moral advantage. For, whenever we find we can engage something in distinctively interpersonal ways—for example, in conversation and argument—then we cannot deny that we are confronted with a person. Any attempt at such a denial would be a form of prejudice.
>
> Moreover, it would be a *hypocritical* form of prejudice, because one would be explicitly denying someone's personhood while at the same time implicitly acknowledging it through interpersonal engagement. So, for example, it was necessarily hypocritical of the slave owners of the American South to deny that their slaves were persons, given that they implicitly acknowledged their personhood whenever they talked to them (and, even more revealingly, when they passed laws against their education). In contrast, it would not be hypocritical to deny that fetuses, the irretrievably comatose, the severely insane, and the hopelessly senile are persons (note that it would be nonsensical to pass laws against educating them). In making such a denial one would not be depriving these human beings of ethical significance. They would remain objects of affection, concern, respect, legal rights, and so on. All one would be doing is registering that there is an important ethical kind to which they do not belong, namely the kind that can engage one another in rational ways and, thereby, treat one another specifically as persons. (Rovane 2004, p. 326f)

Not treating "fetuses, the severely insane, the irretrievably comatose, and the hopelessly senile" as persons is not problematic in an ethics based on Q-Beings[256]. Fetuses simply are not reflectively rational as opposed to persons. But fetuses are interesting because of something else: they have, as opposed to, for example, mice[257], the potential to become full P-Beings. And that should account for another distinction.

To all beings that have not yet achieved, but *can* achieve, the status of full autonomous personhood, that is, all Q-beings that can become P-Beings, we should adopt the stance of nurture;

256 See the remarks above for the especially problematic case of comatose individuals.
257 Both are Q-Beings!

we should try to do everything in our ken to lead them to full personhood. Let us call these beings *N-Beings*, for Nurture-Beings. There are of course beings which are Q-Beings but *neither P-Beings nor N-Beings*, such as cats (or mice as in the above example). We should treat these creatures with respect all the same; albeit not with the intention of turning them into P-Beings. The distinction between Q-Being and N-Being is very difficult and less clear cut than that between P-Being and Q-Being; because the former depends on technological possibilities. For instance, what if we develop technologies that can radically enhance cognitive abilities? Do we have the obligation to "uplift"[258] creatures or even whole species, that is, turn Q-Beings that don't even have the potential to become P-Beings into the latter? Imagine, given the possibility of turning your cat into a person, opting for *not doing this*. Is this unethical? But these questions are for another time and day.

The taxonomy of Q-Beings, N-Beings and P-Beings is especially important for *weaker* sentients. The strongest (rational) species on a planet can opt to *define* the maximal criteria it possesses as being *necessary* to qualify as an ethical agent. This is the status of current mainstream ethics, which focuses on humans and ignores living beings which do not have a language to articulate their suffering. Basing ethics on Q-beings automatically leads to more compassion with all living beings; and it can be justified rationally, as opposed to the singling out of species. In the long run, humans can hope that this kind of ethical taxonomy is adopted by all rational sentients.

Imagine super-intelligent aliens that require higher conditions for their equivalent of P-Beings than humans are able to meet. Then humans should hope that they do not possess a species-centric ethic, but one similar to the one proposed here. For then at the very least we are Q-Beings, and maybe even N-Beings for their kind of personhood.

The sketch of an ethical system presented here scales well – even if we are not the biggest fish in the pond.

4.4.2 Deep Integration

Given the naturalist outlook of this thesis, and the elaborations above, it is time to tackle an important question – will AIs be conscious? To answer this question is of utmost importance, so as not to inadvertently create entities that are then forced into slavery of humanity. The question is also of interest in its own right, because it elucidates the question *of what actually is conscious*[259]. In the

258 http://en.wikipedia.org/wiki/Biological_uplift

259 Again, it should be stressed that this is now the question of what is necessary for specific kinds of consciousness, such as those giving rise to personal narratives, having memory etc. The primitiveness of qualia is already accounted for.

following, when I speak of consciousness, I will refer to conscious states which have content and intentionality; phenomenal states will simply mean conscious states. All these aspects deserve their individual discussion, but for here, it is OK to conflate the concepts.

Above, I have stated my suspicion that maybe brains "only"[260] contribute to planning, prediction and memory; a brain is an introspective machine, that is, a complex physical structure which is capable of traversing through a multitude of states, where some states mirror inside processes of the brain itself.

What about *meaning*? – meaning is not generated *in* the brain; meaning depends on the external environment. The classic argument for meaning externalism is given by Putnam (1975): Putnam imagines another planet, Twin Earth, which is exactly like Earth except in one respect: water is not made up of the chemical compounds H_2O, but of XYZ; XYZ behaves just like H_2O in all noticeable respects. When an inhabitant of the Earth now refers to water, he clearly means H_2O and not XYZ; from which follows that meaning is dependent on external circumstances. The thought experiment only works correctly in our ontology if we note that H_2O and XYZ are still distinguishable somehow; but maybe not by Earthen and Twin-Earthen laymen. That is enough to make the example work (and this conception coincides with Putnam's account).

Let us take a walk again and have a look at the Ferrari we encountered above. We naively think that the Ferrari is "red" and that the car is the bearer of that property; but a tangential familiarity with science already reveals this thought to be mistaken: "red" is only a relational property of ambient light, the reflective properties of the material under scrutiny (here the the lacquer of the car), which results in specific light frequencies being absorbed and reflected, and the specific neural and sensory makeup of the perceiving organism. This position is different from naive realism (locating "red" out there) and radical constructivism (locating "red" solely in the mind):

> Colour experiences, on the view defended here, are mutual manifestations of reciprocal dispositionalities of incoming light radiation and the visual system. (Heil 2003, p. 202)

260 It's enough for starters.

What follows from the above is *semantic externalism*[261] – the position that meaning is not located in the head (alone). Well, then, how does meaning arise in a physical system? Good contenders for this are teleological theories of mental content:

> According to teleological theories of content, what a representation represents depends on the functions of the systems that use or (it depends on the version) produce the representation. The relevant notion of function is said to be the one that is used in biology and neurobiology in attributing functions to components of organisms (as in "the function of the pineal gland is releasing melatonin" and "the function of brain area MT is processing information about motion"). Proponents of teleological theories of content generally understand this notion to be the notion of what something was selected for, either by ordinary natural selection or by some other natural process of selection. (Neander 2004)

Note that, as mentioned in the quoted passage, the sense of function which is used here is not the same as the sense of *functionalism* in the philosophy of mind, but more what Bunge & Mahner (2004, p. 158f) have in mind with "fungieren"[262] – playing a *role* in a physical system. Teleological theories of content, while coming in different guises, notably input-based and output-based ones, agree on one thing: namely that semantics should be derived from functions:

> the content of beliefs [is analyzed] in terms of actions they prompt. (Papineau 2003, p. 26)

Input-based theories such as indicator semantics look at the conditions which produce the beliefs – raising the problem of misrepresentation[263]. Of course, output-based views also require that

---

261 Heil himself does not subscribe to externalism. His position is internalism where he locates the acquiring of intentionality and meaning in the dispositionality of the system in question. I try to synthesize the two positions.

262 Bunge and Mahner would prefer the term teleonomic function, and not teleological function, though.

263 This problem is the target of Fodor's assymetric dependency theory:

> ...we can define a predicate, "x is locked onto y," to capture this asymmetric causal structure:
> A symbol "S" is locked onto property F just in case:
> 1 there's a (ceteris paribus) law that F causes tokenings of "S";
> 2 tokenings of "S" are robust: i.e. are sometimes caused by a property G other than F;
> 3 when Gs (other than Fs) cause tokenings of "S," then their doing so asymmetrically depends on (1) i.e. on the law that F causes "S"s,
>
> where X's causing Ys "asymmetrically depends" on a law, L, if and only if X's causing Y wouldn't hold but for L's holding, but not vice versa: L could hold without X's causing Y. Thus, smoking's causing cancer, depending upon many laws, asymmetrically depends upon Newton's, since Newton's doesn't

the functional roles have been *selected* for – the "historical-etiological" view, so in a sense, input is imported via the backdoor:

> The historical-etiological account of function, by contrast, explicitly restricts attribution of functions to traits that have in some sense been designed to produce the effect cited as the function. On this account of function, functions are the upshot of prior processes of selection. A trait has a function if it has been designed by some process of selection to produce some effect. (Macdonald & Papineau 2006, p. 10)

> Put as briefly and generally as possible, the etiological account says that functionality arises because some individuals in a group acquire novel traits with capacities that are favourable to their ability to reproduce. Such features are transmitted to their descendants, proliferating within the group in the process. Those features will then have as their function the exercise of the favourable capacity. (Macdonald & Papineau 2006, p. 11)

and:

> In such cases, it is natural to adopt teleological terminology, and say that, in the normal case, the trait exists because of an effect the trait can produce, or in order to fulfill its function. (Macdonald & Papineau 2006, p. 11)

What about actual beliefs and desires, which do not (proximately) depend on a phylogenetic evolutionary history, but on the ontogenetic history of the actual thinking being?

> This kind of ontogenetic selection has been termed 'vicarious' or 'secondary' selection by Donald Campbell (1974). Campbell's thought is that the relevant developmental mechanisms have themselves been selected for by genetically based natural selection to be non-genetic selectors. They operate so as to be less severe selectors than death, permitting learning and other adaptational processes to occur.
>
> Campbell developed an explicit 'blind-variation-and-selective-retention' (BVSR) model of learning. There were three essential aspects to Campbell's BVSR model:

depend upon smoking's causing cancer. Fodor's proposal about content then is:
(M) if "S" is locked onto F, then "S" expresses F
(Martinich & Sosa 2001, p. 458)

> (a) mechanism(s) for introducing variation,
>
> (b) consistent selection processes,
>
> (c) mechanism(s) for preserving and/or propagating the selected variants.
>
> (Macdonald & Papineau 2006, p. 15)

This means that non-genetic selection – through individual learning processes for instance – also plays a role in assigning representational functions.

I think that one should not opt for input-based theories/output-based theories in exclusion of one for the other: that beliefs arise in a certain way is due to their causal history; but their output, that is, that the actions they give rise to lead to the satisfaction of desires, is the necessary feedback to cull misrepresentations. Taking heed of both aspects can shed light on things left in the dark by focusing on one approach only.

From the question of semantics and their external nature we must differentiate the question of qualia: are they, at least, in the head, or not? Remember that we have already committed to panqualicism above – the question is not if other physical states than brain states are qualia-like, the question if these other states are required such that *brain states themselves experience the qualia that they do* – it is a question of the *distribution* of qualia states.

There is a funny thing about consciousness: if it truly resides only in the brain, this means that your real skull – not the one you are thinking about, which is only the internal representation, but the real, physical one – is actually beyond everything you currently represent: look over the fields; look over the sea; look out into space – your real physical self will be beyond those, because what you are "looking" at is actually only representations. This picture is strange enough that it may call into question the *natural intuition* we have that qualitative states are located strictly in the brain. So, maybe it isn't *that* strange that qualia also should not be in the head[264].

So, in the present account – which could already have been guessed at in the metaphysical section – intentionality, phenomenal states and meaning arise through *causal connections with the external world, going in both directions* – from the world to the organism and from the organism back to the world. An organism has access to meaning only in regard to its environment in which it

264 This view is proposed in Dretske (1996).

evolved – much like a computer program, which must be embedded in the correct context to be executed correctly.

But what if one were to duplicate my neural structure exactly, down to every atom? What if these process would happen through mere chance? Would meaning, intentionality, consciousness be present in that system? This is what Swampman (Davidson 1987) is all about, and it has raised both much consternation and discussion.

Davidson enters a swamp and is hit by lightening and dies. Another lightning hits a tree stump and lets molecules spontaneously assemble into a new Davidson, absolutely indistinguishable from the "real" Davidson. Swampman now leaves the swamp and carries on with Davidson's life. Nobody notices anything unusual. But does Swampman have intentional states?[265]

First of all some metaphysical objections: even though Swampman may be correctly constituted at some level of description, he violates the one world conjecture in that his connections have formed spuriously; in a way which will certainly not happen in this world, and probably not in any world (see the “few worlds of quantum mechanics” above). Swampman will probably break apart in the near future due to internal stress, because the microcausal structures are not perfectly aligned. But we are being too realistic for philosophy here. Assume that Swampman is *really* identical. Let's play along; maybe there is something to learn here.

The physicalist response would be to say that of course Swampman would have intentional etc states, because these *supervene* on physical configurations and these are, in the present case per definitionem, the same.

Now, there is the strategy to deny that Swampman is in possession of contentful states; and if one let's consciousness go hand in hand with contentful states also deny that he possesses conscious states. That is advocated by Dretske (1995, p. 141f) who let's a Twin Tercel[266] appear – a perfect replica – produced, as a such things go, by lightning – of a normal Tercel; there is one difference though: the gas-gauge pointer of the Twin-Tercel is not responsive to the amount of gas in its tank. Is it broken? That might be the initial reaction of an onlooker. But, after all, Twin Tercel was *not designed*. Twin Tercel is a random aggregation of molecules, so it would be false to ascribe an intended functionality to its gas-gauge, and therefore also false to contend that it was broken. Maybe, in the same sense, it would also be false to ascribe intentional states to Swampman.

265 Or, put differently, is duplication of specific states of consciousness possible?
266 A Tercel is a car.

Michael Tye constructs a similar example, thereby arguing for *qualia externalism*. Qualia, the same as meaning, seem to be a property which are deeply integrated into the micro-structure of the world. Tye imagines the planet Xenon. On Xenon, pod plants grow which contain an organic stuff that resembles human brains. By coincidence, one pod – XP1 – is infused with electricity – by lightning of course[267] – and thereby becomes the exact microphysical duplicate, for fifteen minutes, of a woman having sex on Earth:

> XP1, then, is not a brain. It was not designed by nature to function as a brain, nor has it become a brain by taking on the appropriate control role with respect to a body. Does XP1, for the period of time during the storm in which it is microphysically identical to a particular human brain, undergo experiences phenomenally identical to experiences of the relevant human on Earth? Indeed, does XP1 undergo any experiences? It seems to me that the intuitive answer to both of these questions is No. There is something it is like for the human being during the specified 15 minutes. She experiences a variety of pleasurable tactile, visual, gustatory, and olfactory sensations. But intuitively, I would say, there is nothing it is like for XP1. [Footnote 13] The Xenon example provides us with a possible case in which a standardly embodied creature with a brain and a microphysical duplicate of that brain differ phenomenally (or so it seems to me intuitively). (Tye 2009, p. 195f)

Tye believes that physical indistinguishable individuals have different phenomenal states due to different histories:

> If microphysical duplicates can have different histories, different beliefs, and different desires, and if they can see and touch different objects, why not also hold that (in some possible cases) they can be acquainted with different phenomenal characters? If meanings ain't in the head, then why insist that qualia are in there? [Footnote 15] For the thoroughly modern materialist, the thesis of phenomenal internalism, like the doctrine of phenomenal concepts, should be ''committed to the flames.'' [Footnote 16] Only then will all vestiges of Cartesianism be eliminated from the materialist worldview. (Tye 2009, p. 199)

267 What would philosophers do without lightening?

So Tye would say, in analogy with his pod plant, that Swampman does not experience phenomenal states. Others hold that Swampman is a refutation – or at least, a weakening in argumentative force – of external semantics (Heil 2003, p. 214f). I contend that the position taken up by Heil – that the content is grounded in the dispositional intrinsic nature of the agents, and therefore internal, is not so much at odds with external semantics as it may seem at first. I would like to repeat, due to its importance, a quote already presented in the section on intentionality:

> In any case, we have available a resource ideally suited to account for the kind of projection associated with intentionality: dispositionality. Dispositions are *of* or *for* particular kinds of manifestation with particular kinds of disposition partner. Dispositions preserve the mark of intentionality in being of or for particular kinds of manifestation with particular kinds of non-existent—possible, but non-actual—objects. This is not mysterious or spooky; it is a feature of dispositions possessed by rocks, or blades of grass, or quarks. (Heil 2003, p. 222)

Certainly, subtle points will differ if one favors a purely externalist account or a dispositional internalist account[268]. But those are of no concern to me here; I will gladly leave *those* feuds to the ivory tower of philosophy – what is important to take out of both approaches is where they both agree – there is a *nexus with the world* in both accounts: externalism looks for some kind of causal grounding, dispositionalism derives its grounding – its *of-ness* and *for-ness* – from metaphysics directly. Heil's version is more elegant, but probably more prone to misunderstanding in the current philosophical climate. This *nexus with the world*, this ineluctable connection, is what I call *deep integration*. It is a final renouncement of every Cartesian conception concerning the mind.

On Heil's account, Swampman is conscious and intentional because he has the right dispositional states, no matter that the causal connections are spurious. Swampman mimics those very connections *exactly* – that is the incredible coincidence. The more the exact mimicry degenerates, the less likely the resulting entity will still be conscious in any discernible sense; maybe we will find gradual degradation here.

268 But I am not sure if these are real world differences. After all, they usually depend on bizarre thought experiments – stupendously impossible events giving rise to complex causal structures which usually require complicated evolutionary histories to arise. I think that external semantics correctly captures the intuition that meaning depends on the world around us; but the dispositional account is metaphysically more correct, namely, that at the end of the day, physically indistinguishable objects have indistinguishable phenomenal states. The merger of the two happens in the real world: where real states have real histories, not fairy-tail ones.

The important point is this: if, in this world, Swampman is created by chance as an exact duplicate, he will indeed be intentional and conscious because the very matter he is constituted of mimics, by near impossible chance, the complex dispositional structure which usually only arises through intricate evolutionary histories. But what grounds the intentional states, lastly, is the dispositional nature of his makeup, which is apt to interact (in the "correct" way) with the rest of the world. That is quite different from pure internalism. What would happen if one would lift the physical brain out of its environment – say, into a different universe? Then no consciousness or intentionality would be present, because it would not be in the right physical environment for which the dispositional interactions were apt to constitute content.

Imagine a key – a key has intrinsic properties, and because of these, it is able to unlock a certain lock; that it is able to unlock that lock of course also depends on the shape of the lock – where it different, the key would not be able to unlock it[269]. Such is the way to imagine deep integration – consciousness and intentionality as the key, the world as the lock; both are necessary to go together. Internalism disregards the latter intuition. Internalism is like a key without a lock. And a key without a lock, as everybody knows, is useless.

This conception of things could solve a problem in another domain. While Swampman here on Earth would be a veritable curiosity, in a cosmological scenario, he is not so seldom – in fact, he should be the rule, and comes under the name of a Boltzmann Brain (BB):

> A century ago Boltzmann considered a "cosmology" where the observed universe should be regarded as a rare fluctuation out of some equilibrium state. The prediction of this point of view, quite generically, is that we live in a universe which maximizes the total entropy of the system consistent with existing observations. Other universes simply occur as much more rare fluctuations. This means as much as possible of the system should be found in equilibrium as often as possible.
>
> From this point of view, it is very surprising that we find the universe around us in such a low entropy state. In fact, the logical conclusion of this line of reasoning is utterly solipsistic. The most likely fluctuation consistent with everything you know is simply your brain (complete with "memories" of the Hubble Deep fields, WMAP data, etc) fluctuating briefly out of chaos and then immediately

269 This example is inspired by (Heil 2003, p. 124).

> equilibrating back into chaos again. This is sometimes called the "Boltzmann's Brain" paradox... (Albrecht & Sorbo 2004)[270]

A simple visualization makes everything clear[271]:

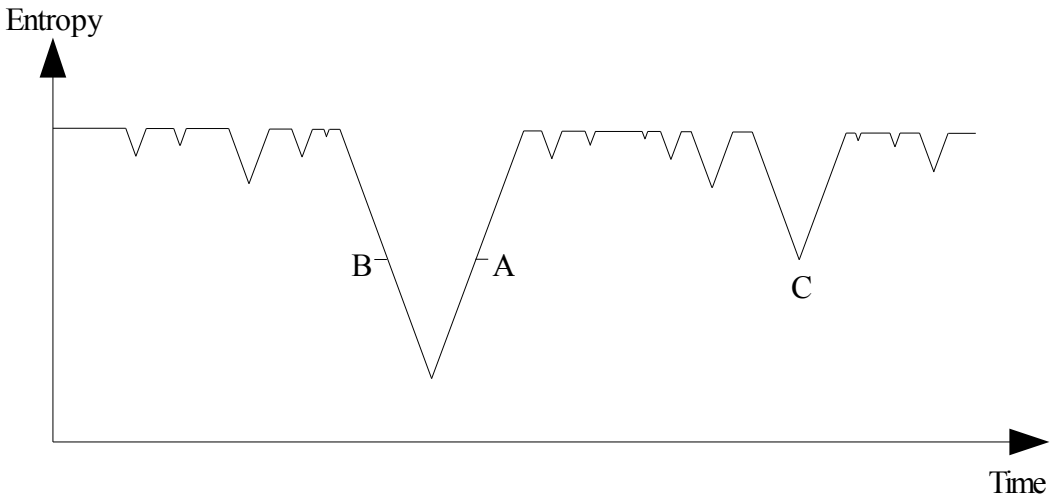


Illustration 5: Boltzmann's Entropy Curve

A normal history point in a universe that is not a farce would be A or B, both of which would look into the past when looking "downward". A farcical BB would be located at point C – complete with a history pointing back to a Big Bang. C, of course, is a random fluctuation much more likely than the one in which A and B live.

This scenario can definitely be avoided by an externalist account of content – BBs would have no contentful states at all. But also Heil's account is sufficient, if one acknowledges that the view as a metaphysical whole is inimical to the conception of the world arising out of a *random mess*. In Heil's view, the world has the structure it has because of the dispositional character of the objects therein. An agent has intentional states because of his dispositional makeup – but this only makes

270 See Boltzmann (1895), the last paragraph.
271 (Carroll 2006)

sense with dispositional interaction partners; you can't take such a dispositional creature out of his world; that intuition is Cartesian.

With the tool of *deep integration* tucked away in our backpack, we are fit for the next leg in our journey.

4.4.3 Computational Functionalism in the Philosophy of Mind

4.4.3.1 Computationalism and Functionalism

> It is often argued that because brains and computers basically are composed of the same atomic and electrical components, there is no reason 'in principle' why computers could not some day be capable of having inner experiences such as sentience and consciousness. Yet all matter is basically composed of particles with electrical charges, but that does not prevent the elements in the periodic table from having quite different properties and causal powers based on their atomic numbers. Moreover, the atomic composition of substances, such as CO_2 and HCl or NaOH and H_2SO4, makes all the difference as to their properties and causal capacities, despite their being basically electrical. The chemical composition of genes is not protein but that of nucleic acids. Should not the fact that our neural networks, unlike computers, are chemical-electrical along with being internally programmed by evolution make a crucial difference as to their capabilities? (Schlagel 1999)

The following section is a bit technical. The reader who has no investment in computational functionalism (CF) may skip it; although it may be interesting for the purpose of understanding how materialism differs from functionalism, two doctrines often held and endorsed simultaneously by scientists[272]. Those who believe that a computation can lead to consciousness *qua (functional) computation* – independent of physical considerations – are the main audience of the following section. What is at issue is the following[273]:

> There is a common intuition among people to the effect that computers and robots lack an essential aspect of what makes us human. [...] The most outspoken advocate of the exclusionist view, however, is John Searle. As we saw in the case of chess, Searle's main argument

272 Indeed, CF comes from the wish to accommodate consciousness in a materialist worldview. Consciousness can be accommodated, but not by CF, because, as we will see, it clashes with materialism.

273 In fact, the position advocated here is similar to Searle's Biological Naturalism (Searle 2007).

> is based on the putative lack of original meaning in computers. Like Nagel, however, he is also concerned with consciousness, which he takes to be a direct product of "the causal powers of the brain" (2002: 67). Thus, although he poses "no objection in principle to constructing an artificial hardware system that would duplicate the powers of the brain to cause consciousness using some chemistry different from neurons," he argues that "computation by itself is insufficient to guarantee any such causal powers" (ibid.). Searle's main argument relies on his famous thought experiment of the Chinese room, (Ekbia 2008, p. 81)[274]

The contrary position is, of course, that consciousness can be "lifted" from the substrate – by ascribing it to computational processes, not to causal features of matter. Let us bring some order into the debate. *Computationalism*[275] can be defined in a harmless way:

> Computationalism is the view that intelligent behavior is causally explained by computations performed by the agent's cognitive system (or brain). (Piccinini 2009)

There is nothing problematic about *explaining behavior* computationally; indeed computational, mechanistic models are very *good* explanations for cognitive processes. But this question is independent (as Piccinini also later notes) as to how cognition and mind relate to the brain. It could well be that computationalism is a good epistemic shortcut for describing cognitive states which themselves are something quite different.

This all raises the question of what exactly we mean with *computation*. We can view a computation as a process that produces outputs from inputs. Together with a liberal view of states – every physical state is a computational state – we will find computational processes everywhere. Saying that everything is a computing system in this sense simply boils down to the assertion that the universe follows (strict) laws. Computationalism can also signify processes that perform information processing; again, we will find very many computational processes of this sort in nature.

But I will concentrate on a more stringent kind of *computationalism* here -- the notion of computation prevalent in computer science and mathematical logic, namely that of Turing computability or e*ffective* mechanisms. Effective mechanisms is an informal notion formalized –

274 The reference (2002) is to (Searle 2002).
275 Computationalism as used here is not opposed to connectionism.

according to the Church-Turing Thesis (Church 1936; Turing 1936) – with the equivalent formalisms of the Universal Turing Machine, the Lambda Calculus or Recursive Functions. *Computationalism* as relevant here is then Putnam's machine state functionalism:

> According to Putnam's *machine state functionalism*, any creature with a mind can be regarded as a Turing machine (an idealized finite state digital computer), whose operation can be fully specified by a set of instructions (a "machine table" or program) each having the form:
>
> If the machine is in state S_i, and receives input I_j, it will go into state S_k and produce output O_l (for a finite number of states, inputs and outputs).
>
> A machine table of this sort describes the operation of a *deterministic* automaton, but most machine state functionalists […] take the proper model for the mind to be that of a *probabilistic* automaton: one in which the program specifies, for each state and set of inputs, the *probability* with which the machine will enter some subsequent state and produce some particular output. (Levin 2004)

The core assumption of *functionalism* is *substrate independence*: what counts is the relation of inputs to outputs and internal states; if they are realized by carbon creatures or silicon computers is of no concern:

> Stones, trees, carburetors and kidneys do not have minds, not because they are not made out of the right materials, but because they do not have the right kind of functional organization; their functional organization does not appear to be sufficiently complex to render them minds. Yet there could be other thinking creatures, perhaps even made of Swiss cheese, with the appropriate functional organization. (Shagrir 2005)

I would like to point out for the sake of clarity that computationalism and functionalism are logically independent:

> To get functionalism from computationalism, we also need an additional assumption, such as that the nature of mental states is (entirely) computational. To get computationalism from functionalism, we also need the independent assumption that all functional states are

> computational. Both of these assumptions are controversial; neither is especially plausible. (Piccinini 2009)

So here, I will take a look exactly at this "implausible" combination – computational functionalism (CF), which AI researchers in the strong AI[276] community nevertheless hold, at least implicitly.

CF then asserts that:

- mental states are computational states, and only computational states
- all functional states are computational in nature

This position is of interest because it holds promise that one can *program* an AI. Denying CF makes this project largely hopeless. To be precise: it does not make AI *per se* hopeless, only the notion of *programming* one.

The materialistic view – that matter matters, and function is not enough -- is often deemed chauvinistic, because it seems to accord special pride of place to *organics* such as humans. That may be so; but the present view is more subtle. Let us proceed slowly. First of all, my motivation, as can be guessed from everything written above, is not to reserve any special place for mind or humans in the order of things. The denial of CF proposed here especially has *no* religious overtones; the rejection of CF follows for *ontological* reasons alone. CF is a result of the picture theory, and must be thrown out together with the picture theory.

What I also do not intend to say is that human minds are more powerful than "mere machines". It should be noted that Turing's conception of a Turing machine places *limits on human cognition* (Dresner 2008); anybody who claims that human minds are superior to computers is welcome to compute a function that a Universal Turing Machine can't compute.

What is at issue here is not a cognitive superiority of humans to machines -- an attitude which is totally foreign to this thesis. -- but the question of if the *fundamental physical structure of the world*, and thus, consciousness which is the "inside view" of certain physical structures, is completely captured by a high-level computational view. Maybe, of course, the universe as a whole is a

276 Strong AI is the doctrine that it is possible to create *thinking* machines, not merely intelligent machines.

computation – pancomputationalism[277] – and then mind is trivially also part of this computation – more on this below.

The CF view of mind is deeply related to conceiving the mind and ourselves as machines. But there are machines and there are machines. We can continue to conceive of ourselves as machines without adopting CF. Following Harnad it is is difficult to see how we could *not* be machines:

> ...if we do follow this much more sensible route to the definition of "machine," we will find that a machine turns out to be simply: any causal physical system, any "mechanism." And in that case, biological organisms are machines too...(Harnad 2003)

This can, I think, be taken for granted by every scientifically minded person; and indeed conceiving of yourself as a machine in such a way can be a very healthy process, because it leads to inquiry into *mechanisms* of behavior and the adoption of strategies to optimize behavior – we saw this above in the section on optimal will. But physical machines need not be Turing Machines; that would be committing the Church-Turing Fallacy, which

> is to believe that the Church–Turing thesis, or some formal or semi-formal result established by Turing or Church, secures the following proposition:
>
> If the mind–brain is a machine, then the Turing-machine computable functions provide sufficient mathematical resources for a full account of human cognition. (Copeland 2004)

This is not true:

> But Turing had no result entailing that "a standard digital computer ...can compute any rule-governed input–output function." What he did have was a result entailing the exact opposite. The theorem that no Turing machine can decide the predicate calculus entails that there

277 Piccinini has something funny to say about pancomputationalism in one of his papers:

> I have encountered two peculiar responses to pancomputationalism: some philosophers find it obviously false, too silly to be worth refuting; others find it obviously true, too trivial to require a defence. Neither camp sees the need for this paper. But neither camp seems aware of the other camp. (Piccinini 2007)

This is also true of other philosophical positions. Something that is regarded as obvious in one research community is regarded as unbelievable in another. I have encountered this phenomenon a number of times in my interdisciplinary ventures. I say this because many people may hold the philosophy of computational functionalism to be obviously wrong, and not understand why I am at pain to show that it can't be held upright. Functionalism is far from dead (Buechner 2008).

> are rule-governed input–output functions that no Turing machine is able to compute – for example, the function whose output is 1 whenever the input is a statement that is provable in the predicate calculus, and is 0 for all other inputs. There are certainly possible patterns of responses to the environment, perfectly systematic patterns, that no Turing machine can display. One is the pattern of responses just described. The halting function is a mathematical characterization of another such pattern. (Copeland 2004)

The CF view is stricter than merely saying that we are machines: CF is closely related to the idea that we could take the computational description of ourselves *at certain levels* and build a new version of ourselves with this description, in, say, another substrate, a substrate which can instantiate all the relations of that the level under scrutiny. These are the ideas underlying whole brain emulation (WBE):

> An important issue to be determined is whether […] a cut-off exists in the case of the human brain and, if it does exist, at what level. While this paper phrases it in terms of simulation/emulation, it is encountered in a range of fields (AI, cognitive neuroscience, philosophy of mind) in other forms: what level of organisation is necessary for intelligent, personal, or conscious behaviour?
>
> A key assumption of WBE is that, at some intermediary level of simulation resolution between the atomic and the macroscopic, there exists at least one cut-off such that meeting criteria 1a and 1b at this level of resolution also enables the higher criteria to be met. (Sandberg & Bostrom 2008, p. 12)

Where 1a and 1b are

> 1a: An inventory of all objects on a particular size scale, their properties and interactions. (Low level neural structure, chemistry, dynamics accurate to resolution level.)
>
> 1b: A complete 3D scan of a brain at high resolution.
>
> (Sandberg & Bostrom 2008, p. 11)

And higher levels include:

> 'Mind emulation': The emulation is truly conscious in the same way as a normal human being. (Sandberg & Bostrom 2008, p. 11)

While Sandberg and Bostrom do acknowledge details of organic processes, they express the view that there is a functional description *at a certain level* which can be emulated; and then, optimally, the emulation will be as conscious as the original brain. If that is not the picture theory at work, I know not what.

I have, like Searle, two main objections against the CF view[278]. In my case, they come from the metaphysical picture painted in this thesis – one coming from the notion of *meaning* and the concept of *deep integration,* the other from notion of the identity of dispositional (causal) and qualitative powers. Let us have a look at each in turn.

4.4.3.2 Meaning

One of the most well known arguments for the problematic nature of meaning in computational systems is the Chinese room argument put forth by Searle (1980). A human sits inside a room, and manipulates Chinese symbols according to instructions – the human does not understand Chinese, but manipulates the symbols in a purely "mechanical" way. As computers do nothing but manipulate symbols mechanically, they, as well, do not understand. Searle is opposing strong AI: the idea that formal symbol manipulation can lead to actually thinking machines. Searle's intention was to show that meaning lies at the heart of understanding, not mechanical symbol manipulation.

In short, programs can't, *qua running as program*, constitute understanding, intentionality etc:

> We might summarize the […] argument as a *reductio ad absurdum* against Strong AI as follows. Let L be a natural language, and let us say that a "program for L" is a program for conversing fluently in L. A computing system is any system, human or otherwise, that can run a program.

278 There is a third objection. The moment we accept that digitizing persons and copying them arbitrarily is possible, we make a farce of science and the universe. Bostrom argues that in the case future civilizations will actually run simulations of the past (and thereby also simulating your current experience), we are probabilistically forced to believe that we already live in a simulation now (Bostrom 2003).

Marchal goes even further; in an argument of eight steps he shows that in case we assume that persons can be "run" on a computer, we must conclude that physics follows from the sum of all (Platonic) computations (Marchal 2004). The argument has some problems in detail, but the core point is this: if there is a possibility to make duplicates of yourself, and you assume that these duplicates are actually run, then what you expect to happen in your next moment of experience depends on what happens to the *majority* of your duplicates – indeed, the physical world loses it's power if *enough virtual worlds exist* which simulate your continuations.

> (1) If Strong AI is true, then there is a program for Chinese such that if any computing system runs that program, that system thereby comes to understand Chinese.
>
> (2) I could run a program for Chinese without thereby coming to understand Chinese.
>
> (3) Therefore Strong AI is false.
>
> The second premise is supported by the Chinese Room thought experiment. The conclusion of this [...] argument is that running a program cannot create understanding. (Cole 2004)

The computer does not understand anything, because it is only manipulating symbols without any intrinsic sense. And indeed, if we look at the bootstrapping process for a computer, there is nothing but machine language, relating to itself; gates switching and switching and switching...it has come about by human designers, having their own goals in mind; goals of calculation, of exact mechanical behaviour etc

Meaning is grounded – via evolution – in reality. This grounding is so basic that it is easy to miss. That does not mean that it is impossible to create thinking, feeling, conscious machines; it is just that restrictions are more severe than initially supposed.

> The fact that the primary function of the central nervous system in higher organisms seems to have evolved to represent a world to these creatures explains the difference in meaning of 'designation' and 'interpretation' for humans compared to computers. Although proponents of AI attempt to mask this difference by referring to 'the symbol manipulating machine' and 'neural net representations,' the physical symbols lack designation and cognitive meaning, while by 'representation' in neural nets is actually meant the encoding of patterns of physical stimuli, not an actual representation in the sense in which our brains represent a world.
>
> *In fact, consciousness would seem to consist mainly in having representations, which is why consciousness is not attributed to computers, however marvelous their computational abilities.* (Schlagel 1999)

We are selected by evolution to *represent* the world – that is one of the points of deep integration. Computers are not selected in this way. A prediction of *deep integration* would be that AI's must be *trained in the physical environment* to develop sufficient meaning; they must learn meaning via perceptors and actuators, feeding back into a representational substrate; meaning can not be programmed via symbols. But that is not enough[279]. Apart from that, in accord with materialism, the *representational medium can't be ignored* – even if we accept Tononi's functional account of information integration – not every material, it seems, can achieve a *sufficiently dense* kind of information integration. Which kinds of matter are sufficient to generate interesting kinds of consciousness are empirical questions. What we can exclude is that something that was not conscious, such as a computer, *becomes* conscious by running the right kinds of programs on it. But now we turn to dispositions.

4.4.3.3 Dispositions Again

A computational description is primarily that: a description at a *certain level of abstraction.* Here is where the tension arises: the physical world is all about particularity, specificity, and detail; and mathematics is all about abstraction, equivalence classes, and structural similarities. Mathematical equivalence does not smoothly translate into physical equivalence. We could, for instance, construct a computer out of silicon, or even out of some more outlandish ingredient – Chinese people, for instance (Block 1980). If we were to draw a state space diagram with possible nomological transitions of the systems in question (the silicon computer on the one hand, the Chinese people computer on the other), these would turn out to be *completely different*; in fact, different beyond recognition. Yet *functionally* – when we draw maps of the computations they can perform – they would be indistinguishable. Now, one can follow through with functionalism and ascribe mental states to the Chinese people computer; but that is clearly not physicalism or materialism anymore, because then we do not look at nomological (physical) properties. Mental states are attributed to systems solely due to functional relations they instantiate at some *level* or other.

But matter matters. Let us consider a harmless fellow – the gecko. The gecko's abilities depend on the very physical world he lives in:

> Geckos have evolved one of the most versatile and effective adhesives known. The mechanism of dry adhesion in the millions of

279 That is, sensorimotor grounding will not be enough, as Kiverstein (2007) argues. Sensorimotor grounding will most probably be necessary, but not sufficient.

> setae on the toes of geckos has been the focus of scientific study for over a century. We provide the first direct experimental evidence for dry adhesion of gecko setae by van der Waals forces, and reject the use of mechanisms relying on high surface polarity, including capillary adhesion. (Autumn et al. 2002)

If physics were not as it were, the gecko would not have his cool adhesive toes. There is no functional abstraction – adhesiveness – that is satisfying for the gecko. Certainly, we can define a *predicate* "adhesiveness" and categorize *different forms* of physically occurring adhesiveness under this one form. But the abstraction we make would come at a cost: we could not be sure that the kind of adhesiveness we mean is the kind of adhesiveness we need in a physical situation. The gecko would be very unhappy with office-glue type adhesiveness, for instance.

Why should mind be independent of material properties, when adhesiveness is not? That seems very Cartesian/dualist to me. Relying on functional relations entails a commitment to substrate independence; that there is a reasonable abstract level of description below which we are free to substitute different matter elements to fulfill the role prescribed by the functional description. Computer gates can be realized by semiconductors, relays, vacuum tubes, water pipes and troughs; but their behaviour will differ (in reliability, speed of computation, ease of reading off of results, size of the system, effectiveness under different environmental conditions such as winter, summer etc). But clearly, even here we can't abstract away from *all* material properties: for instance, we couldn't work with air exclusively. There is the tension again between material processes and abstract descriptions thereof. Materialism takes nomological properties of matter seriously. Causal patterns are considered relevant, down to the *finest granularity* of detail. But abstraction is the deathsman of detail.

Kary and Mahner give illustrative examples of how function can't be separated from material properties:

> Now consider a slightly more demanding process, namely what one might call 'wheeling'. Recall that the difference between what one might in general call a 'roller', namely anything with a round cross-section, and a wheel, is the presence of a hub and axle. We find that, while solids, liquids and even gases might roll, and that while rollers do occur frequently in nature, only solids can be used to build wheels, and no wheels occur naturally. Furthermore, not every combination of solids makes for a good wheel: for one to be used in a

> simple pushcart, for example, the materials must be chosen so that among other things, the friction between the outer rim and the ground is always appropriately greater than the friction between the hub and the axle. Thus, not only are there new restrictions placed on the possible components, but also upon the internal structure of the system (the friction at the hub), the external structure (the friction between the rim and the ground), and even upon the relationship between the internal and external structures.
>
> The point of this example is to show that the more complex and specialized the function, the more it becomes tied to the special properties of specific materials and systems. This is only a general rule of thumb; sometimes even very simple properties are tied very specifically to special systems. Consider for example the atomic property of being able to join together with like atoms to form long chains and branched molecular systems: there is but one atom that has this property, namely the carbon atom. There is a simple molecule that has a similar property, namely SiO, silicone; but the dissociation energy of such bonds between SiO molecules is significantly greater than that between carbon atoms, so only carbon is suitable as a biomolecular building block for the temperatures encountered here on planet Earth.
>
> It is then simply a reflection of the facts of nature that while some process classes are very large, others are extremely small. (Kary & Mahner 2002)

Or consider Wieland[280] the Smith, a character out of Germanic mythology[281]. He forged the formidable sword Mimung. He forged it three times anew: every time he fragmented the sword into little metal shavings, mixed them with wheat and let geese eat them. He then remade the sword out of the geese droppings; the final sword was incredibly strong and sharp. In modern terms, what Wieland did was to perform the process of *nitrogen hardening* of steel (the nitrates being contained in the bird excrement). We see that the addition of only a few key chemical elements changes the property of the whole significantly. A more modern example would be the doping of semiconductors to change their electrical properties. In this case, often only as a few as one atom per hundred million atoms is needed to change the properties of the material in question.

280 Anglosaxon: Veland.
281 Thidrekssaga, 57-79

Let us move closer to biology and the brain. First of all, neurons come in all forms and shapes. That should surly make a difference. But let us concentrate on something else: neurotransmitters, which also come in great variety and types (Kalat 2007, p. 59f). Most neurotransmitters are amino acids, proteins. When we look at proteins, we discover primary structure – the amino acid sequence; secondary structure – alpha helix or beta sheet; and tertiary structure – the fold of the protein due to mainly noncovalent interactions. Even a quaternary structure – multiple-subunit proteins – can be discerned. All these structures make a difference in one setting or another. Eliminate the structure and you eliminate the function.

What about the relation of carbon and silicon? In philosophy there is a well known class of arguments that go by the name of Sorites-Paradoxes[282]: they highlight the vagueness of language. A classical Sorites-style argument is the "inability" to make a *heap* of sand from sand *grains*. One sand grain is not a heap of sand; two neither; and the addition of one more sand grain can certainly not make a heap of this allotment; and so on. One can also do this argument the other way round by beginning with a heap of sand and consecutively removing grains; at the end, one will have a heap without any grains in it! This is of course only a conundrum for philosophers afflicted by the picture theory or those who suffer from even more serious conditions[283].

Now let's start, sorites-style, substituting silicon neurons for the real thing in the brain. Surely, one silicon neuron will not make a person much different; and certainly not unconscious. And, so the computationalist reasoning goes, if we only keep functional equivalence we can go on substituting biological neurons for silicon: at the end we should have a conscious entity – the same person actually – but made out of silicon instead of carbon[284]. The problem is of course this: at what *level* do you want to *ascribe functional equivalence*? At the material level, silicon and carbon (to pick out two main chemical ingredients) are certainly *not* equivalent, otherwise they would not have been called differently by chemists and would not merit different entries in the periodic table. The reader may convince himself of the different properties of carbon and silicon in every elementary chemistry textbook[285].

282 "soros" is Greek for "heap". The sorites-puzzle is attributed to having been first presented by Eubulides of Miletus (Hyde 2005).

283 Computer scientists should have no problem debunking sorites style arguments: consider highly available computer systems which operate with redundancy – you can let some parts break, and the system will still work; with luck, when the load isn't high, nobody will even notice anything. But go too far with this game, and the system breaks down. The system will definitely not be highly-available when all supporting computers have broken down.

284 Read carbon as *pars pro toto* for all matter that makes up a neuron.

285 For instance Zumdahl (2007).

Different material properties at the micro-physical granularity have big consequences in the macroscopic world: the properties of silicon and carbon, while being mutually more similar than any one of them compared with an element of a different group, give rise to very *different* structures when aggregated. They differ in their dispositional properties, and that means, under the present metaphysics, also in their *qualitative* properties. There is no "higher-level" functional substitution that leaves the qualitative properties intact.

So, a gradual substitution with silicon neurons will eventually lead to different qualitative properties; a single silicon neuron will most probably not make a difference; substituting the whole brain definitely will. Maybe there are silicon substitutions that will be able to sustain conscious states – if they support enough information integration as in Tononi's theory. But if that is the case, that will be because silicon can support consciousness – silicon consciousness – *sui generis*, and *not* because of *functional equivalence*.

It is also interesting to consider altered states of consciousness. Drugs – alcohol as the most common form – alter states of consciousness. More invasive is anesthesia; in all these cases, the alteration of the current *material* makeup – even if only temporary, that is, the mere alteration of the causal "runtime" behaviour – can alter or even eliminate conscious states. The functionalist will simply say that *function* has been disrupted. But how? By the introduction of different *materials* into the conscious system. Again, material properties do the actual work.

The functionalist or computationalist will be hard-pressed to suggest material substitutions which have *no* effect. In short, I do not see how changing the material substance – which always suggest changing the dispositional and qualitative nature of the entity in question – can leave conscious states unaltered. But if this is not possible, then functionalism (or computationalism) as a metaphysical doctrine is false (albeit epistemologically helpful). It means, in effect, that there is no "substitution" level in human beings where neurons can be replaced by something else and lead to the *same* qualitative states.

> Because a computer lacks neurons and the chemical components of neural transmitters, such as dopamine and serotonin, one does not expect computers to develop schizophrenia or Alzheimer's disease, although they can 'crash' for other reasons. But sentience and consciousness are dependent upon the same chemical components, so why should one reject the former and expect the latter? More over, because it is believed that dopamine is the chemical that also plays a major role in creating the sensation of pleasure, it is unlikely that such

> sensations will occur in mediums that lack that chemical, as well as the underlying organic structure in which it functions. Considering how the slightest chemical imbalance or deficiency in neurotransmitters in an organism can affect its experience, is it likely that consciousness or sentience might emerge in machines entirely devoid of chemicals? (Schlagel 1999)

Little changes in neurotransmitter levels amount to big changes in experience – so why should big changes (circuits instead of organic compounds) not lead to big changes in experience, so big as to maybe even lead to *no* experiences?

Our consciousness is not independent of micro-states realizing the same higher-level functional states– that is a Picture Theory fantasy. Functionalism is scientifically respectable because it gives high level descriptions of regularities of interest to human beings. Functional descriptions are epistemic shortcuts of the unified, complex world. The functionalist account extracts higher order regularities but should not be imbued with ontological significance. In this restricted sense functionality certainly plays a role[286], but the functional role can't be abstracted away, rather it depends both on the causal history which led to exactly that kind of function in the organism and on the current microcausal structure (which of course originated historically).

We *are* the finest level of reality in operation. W*e are the quantum*[287]. There is no atom, molecule, cell, organ or organism independent of the quantum. The smallest "parts" we are made of are just the smallest dispositional causal slices we can make of the world: that is, when we look with great precision at the fundamental constituents of nature, we are not actually discovering lower levels which can support independent higher levels, but rather we are discovering the *fundamental causal entities*. We may, with technical prowess, isolate causal entities -- slice them off – and then discover how they behave in isolation (as when we isolate atoms for experiments) but we are never looking at a "lower level". We are the "lowest" level in the first place.

To continue Harnad's machine quote above:

> … and the answer to our question "Can a machine be conscious" is a trivial "Yes, of course." We are conscious machines. Hence machines can obviously be conscious. The rest is just about what kinds of machines can and cannot be conscious, and how -- and that becomes a

286 That qualia are tightly related to function is explicated in Cole (2000).
287 If that is the finest level.

> standard empirical research program in "cognitive science"..(Harnad 2003)

Indeed. What kinds of machines are conscious? Remember that Harnad defines machines as causal physical systems. That is something different than a Turing machine. Sloman recognizes this point:

> … there are (at least) two very different concepts of computation: one of which is concerned entirely with properties of certain classes of formal structures that are the subject matter of theoretical computer science (a branch of mathematics), while the other is concerned with a class of information-processing machines that can interact causally with other physical systems and within which complex causal interactions can occur. Only the second is important for AI (and philosophy of mind). (Sloman 2004)

Could AIs be conscious in the sense of strong AI[288]? First of all, there is no theory whatever which predicts that such a thing were *not* possible. Consciousness is very much part of the fabric of the world. And, we observe strong "I's" (intelligences) every day: your fellow humans, yourself etc. We are conscious, and we live in this physical world, made of the same physical stuff as everything else. Our consciousness is a property of the matter we are made of, not something magical tacked on as an afterthought. Physical matter configurations can become conscious; so much is clear.

Now, in line with panqualicism, different kinds of matter – not only carbon compounds – probably *will* have mental states – that is, *if these different kinds of matter can form sufficiently complex structures.* Here carbon compounds *do* enjoy considerable advantage over other materials. But, if mind arises, it will be a *different kind of mind* than a human mind; mind can't be abstracted and lifted onto another substrate. Different substrate, *different* mind. And another point is important: if these other entities are conscious, they will not be conscious in virtue of computational properties but in virtue of material properties.

Maybe the necessary and sufficient criteria for consciousness are the following: what we need is a material substrate which supports highly complex parallel interactions, satisfying Tononi's IIT

288 Separate from the question of consciousness is the question of intelligence; weak AI is trivially possible, indeed, we already have it. Weak AI is about *simulating* intelligent behavior. The behavior need be connected neither to understanding, insight, nor consciousness. Simulation, it should be noted, is not the real thing. To bring well-known examples: a simulated fire is not hot, a simulated stomach does not suggest, and a simulated thunderstorm is not windy, cold, and wet. Now, a simulated stomach may not digest, but maybe an artificial stomach having the right causal powers will digest; again, we are driven back to the question of what the right causal powers are – *not* what the right computations are.

criterion, standing in the right causal relations to the world so as to give rise to representational structures. Currently, these conditions are satisfied by certain biological organism. Maybe, with good engineering – including molecular biology and nanotechnology – we can build artificial entities that are also conscious. But the solution will be in *matter*, not in *computations*.

To connect all this with our discussion on identity above: while psychological continuity is very important for identity, it seems we *can't change the material substrate just so*. So psychological continuity is largely mediated by *animal continuity*, in line with animalism; the present position does justice to both intuitions on identity.

Those who are not satisfied by the objection from meaning and the objection of right dispositionality I refer to Appendix E: Maudlin's Olympia.

4.4.4 A Computational Universe?

Maybe there is a method to reconcile some aspects of computationalism with our attaching the utmost importance to material differences: namely that computationalism is not a theory of cognition but is more apt to be a tool for modeling causal relations, especially micro-causal relations. That is, computationalism may be well alive as a *physical* doctrine, not one of cognition, when the whole universe or at least "local patches of flux" are conceived of as computations (Fredkin 1992; Fredkin 2003; Svozil 2005; Lloyd 2006); also Schiff (2008)

The notion of the universe as a computer has other problems. I do not have the time to dwell, I just want to clarify that these are *separate* issues. Ruling out a computational theory of mind does not a priori rule out a computational theory of micro-causal relations. Of course, mind would then also be a micro-causal (computational) relation, but *not one separable from the material substrate. S*uch a theory of computational mind would then be fundamentally different in kind than current computational theories of mind which have functionalist (abstract) connotations and which want to explain mind in computational terms, as opposed to just "carrying it along". The explanation of mind in the present picture is given by the qualitative/dispositional account.

An example: carbon has a different nomological state space than silicon. Spoken computationally, one could say that carbon *supports different computations* than silicon. Maybe only the complex computations of carbon (what we normally refer to as its material properties) can give rise to sufficiently complex structures that have a mind worth speaking of. Clearly, nowhere do

we speak here of cognition anymore. Computation conceived in this way does not pick out anything special about minds, and thus would not be explanatory in cognitive science:

> If pancomputationalism is true, so that everything is a computing system, then minds are computing systems too. But at the same time, computation ceases to be a specific kind of process among others. If the fact that minds are computing systems follows trivially from the fact that everything is, it is unclear how computation could explain how minds exhibit their peculiarly mental characteristics. (Piccinini 2007)

5 Conclusion; or How to Proceed

5.1 Introduction

The Sage falls asleep not

because he ought to

Nor even because he wants to

But because he is sleepy.

(Smullyan 1977, p. 3)

In this last chapter, we will see what follows for us human beings in the world as painted above. Ethics have already been touched above, but will concern us again, from a slightly different angle. But first it is time to touch on a few spiritual matters. Why does spirituality feature in a philosophical text aimed at locating humans in the natural world? Because that is a perfectly natural thing to do. Many humans have spiritual desire. Why should the naturalist be bereft of the right to exercise this desire? Spirituality is about finding *your relation to the universe*; your place in the world. In this sense, it is also clear why ethics is tied closely to spirituality in many traditions: because other agents are also part of the universe to which you relate[289].

But now to the subject at hand. The physicist Steven Weinberg writes:

> The more the universe seems comprehensible, the more it also seems pointless.
>
> But if there is no solace in the fruits of our research, there is at least some consolation in the research itself. Men and women are not content to comfort themselves with tales of gods and giants, or to confine their thoughts to the daily affairs of life; they also build telescopes and satellites and accelerators, and sit at their desks for endless hours working out the meaning of the data they gather. The effort to understand the universe is one of the very few things that lifts human life a little above the level of farce, and gives it some of the grace of tragedy. (Weinberg 1977, p. 149)

That is the kind of writing which gives the scientific worldview a bad image. Why should such a reduced view of value follow from the scientific worldview? It may *seem* that by placing humans

289 I will draw (only lightly) from existing spiritual traditions. I am always somewhat ambivalent when presenting connections with traditional texts; they contain deep insights, but also a lot of historical nonsense – often the metaphysics went awry at some point, leading to false conclusions. How, then, to tell the difference between insight and nonsense? By applying the standards of rationality, of course.

thoroughly in the material world there can be no place left for value – but nothing could be further from the truth. A whole universe – literally – of value arises. Discarding the supernatural or the immaterial spirit/mind does not mean that we also have to discard the values that we hold; they just receive a different justification.

We have come to the conclusion that mental states *are* certain matter states; that is, the qualitative dimension of our experience is *built into the very fabric of the universe*. Meaning, in fact, is everywhere: meaning lies in our experiences, and these experiences are natural, physical states; they do not need a *further* purpose derived from outside the universe; from a God; or from a teleology; or an eschatology. When I love someone, why should this be devalued by being a physical state? Love can stand for itself – it is no less real or important just because it is a normal aspect of the physical universe. On the contrary, it is *more* real. Many people hear of such things as ethics and value only if they are embedded in religious discourse. If you show the religious discourse to be largely untenable, the impression may arise that meaning is lost too. But that is only because the two are cognitively intertwined in respective human beings, not because of logical or metaphysical necessity.

We should see that everything *real* around is is actually everything we had from the beginning, and that is enough. The spring flower is enough. The loving couple is enough. The cat playing in the garden is enough. And yes, also the game of life and death in the jungle – is enough. The universe is a playful place; an insight captured in the vedantic doctrine of Lila:

> Brahman is full of all perfections. And to say that Brahman has some purpose in creating the world will mean that it wants to attain through the process of creation something which it has not. And that is impossible. Hence, there can be no purpose of Brahman in creating the world. The world is a mere spontaneous creation of Brahman. It is a Lila, or sport, of Brahman. It is created out of Bliss, by Bliss and for Bliss. Lila indicates a spontaneous sportive activity of Brahman as distinguished from a self-conscious volitional effort. The concept of Lila signifies freedom as distinguished from necessity." (Misra 1998)

We can imagine that Lila is the splitting Godess; enacting reality as a playful sport. Eternity is a long time; you have to be quite imaginative to pass it; why not play a little? – and take joy in this play:

> the perfect devotee does not suffer; for he can both visualize and experience life and the universe as the revelation of that Supreme Divine Force (shakti) with which he is in love, the all-comprehensive Divine Being in its cosmic aspect of playful, aimless display (lila) - which precipitates pain as well as joy, but in its bliss transcends them both. (Zimmer & Campbell 1969)

This is not the time to indulge in Indian religious scholarship; I will have to gloss over the concepts above and just take what comes easily. On Brahman, we read that one possible translation is:

> the ultimate Reality; the ground of the universe; the Absolute; the Divine...It has nothing similar to it and nothing different from it, and it has no empirical distinctions from the acosmic viewpoint. (Grimes 1996)

For present purposes I will equate Brahman with the Dao, as I am more familiar with that terminology. The Dao is playful; it enacts everything, and we are part of this enactment. We can take joy in this.

5.2 Why the Ultimate is not God

> The most preposterous notion that Homo sapiens has ever dreamed up is that the Lord God of Creation, Shaper and Ruler of all the Universes, wants the saccharine adoration of His creatures, can be swayed by their prayers, and becomes petulant if He does not receive this flattery. (Heinlein 1973)

> Tung-kuo Tzu asked Chuang-Tzu (Chuangtse), "Where is that which you call Tao?" Chuang-Tzu said, "Everywhere". Tung-kuo Tzu said "You must be more specific". Chuang-Tzu said, "It is in this ant". "In what lower?" "In this grass" "In anything still lower?" "It is in tiles". "Is it in anything lower still?" Chuang-Tzu said, "It is in ordure and urine". Tung-kuo Tzu had nothing more to say.(Creel 1970, p. 31)

If I use the word "Dao"[290]. to refer to ultimate reality – all of it; the ineffable suchness of being transcending any *one* description, some people are bound to ask: is the Dao not simply God? Everybody has different conceptions of God. But we can dispel some conceptions, and what is left will *not mean* God for most people, at least in the Western tradition.

290 A wonderful introduction to the Dao and Daoism for Western minds is Smullyan (1977).

First of all, the Dao, as is even expressly said in the Dao Dejing, is something different than God:

> The Tao is like a well: used but never used up.
>
> It is like the eternal void: filled with infinite possibilities.
>
> It is hidden but always present. I don't know who gave birth to it.
>
> It is older than God. (LaoziMitchell 1995)(4)

The Dao also has no beliefs, desires and intentions. It knows not good and not evil. It is the mother of all things but does not lord it over them. It does not command and it does not punish:

> The Tao doesn't take sides; it gives birth to both good and evil. (LaoziMitchell 1995)(5)
>
> The Tao is called the Great Mother: empty yet inexhaustible, it gives birth to infinite worlds. (LaoziMitchell 1995)(6)
>
> The Tao is infinite, eternal. Why is it eternal? It was never born; thus it can never die. Why is it infinite? It has no desires for itself; thus it is present for all beings. (LaoziMitchell 1995)(7)
>
> The Way is like a great flooding river. How can it be directed to the left or right? The myriad things rely on it for their life but do not distinguish it. It brings to completion but cannot be said to exist. It clothes and feeds all things without lording over them. (LaoziMitchell 1995)(34)

The Dao, above all, is not a person; and it does not reserve a special place for human beings. Humans *ride* on the meaning of the universe, they do not play a privileged role in it.

In recent years, a number of prominent scientists have written on the topic of "the sacred"; most prominently Kauffman (2008). Sacred can mean a number of things, among them "forbidden" and "deserving absolute respect". The concept of Dao has no connection with either of the two meanings. The Dao is not sacred. It simply is. It does not want to be revered. It does not want anything. When thinking about the Dao, laugh

The Dao is also very different from God in another respect:

> In the Judeo-Christian religions, one hears much of "fear of God" and "love of God" - also "obedience to God". In early Chinese Taoism, one speaks not so much in terms of "love of Tao" – and certainly not "fear of Tao"! - but rather of "being in harmony with the Tao".
>
> Fear of Tao is completely ludicrous! Tao loves and nourishes all things, but does not lord it over them! Tao is something totally friendly and benevolent – friendly to *all* beings, not just those who believe in it or "accept it as their Saviour!". Thus Tao is the sort of thing which is impossible to believe in without loving. But the loving of Tao is not stressed for the simple reason that it is obvious. To command love of Tao would be as silly as commanding one to love his closest friend. (Smullyan 1977, p. 50)

5.3 Redemption: The Materialistic Soteriology of Change

Religion does not only talk about ethics, meaning and God, but also about an afterlife. Wouldn't an afterlife be cool? Humans seem to think so, otherwise they wouldn't strive for it. There is the hope to overcome death by processes such as cryonics (Ettinger 1964) or "uploading" of the mind, that is, transferring a person from an organic substrate into a different physical substrate, one which is amenable to digital manipulation.

Cryonics may work (Merkle 1992). That depends on the development of nanotechnology and the ability to manipulate, in detail, complex biochemical processes. There is no principled reason why it should not work. As such, it is a way to ensure psychological continuity in accord with animalism. Medicine is about life extension: sicknesses that would lead to death are cured, thereby postponing death. Cryonics is the continuation of this idea: if a person is sick in such a way that he is dead according to *current* medical standards, the state of the body is stabilized via the cryo-process to keep the causal structure making up the person intact until a future age where the person can be healed. Death, if cryonics works, only occurs when the causal structure constituting a person is *irretrievably lost*, radiated out into the environment by physical processes.

However, with cryonics, one is still bound to an organic material substrate, and it is questionable that even excellent rejuvenation medicine would lead to *eternal* life. But it is a good bet for life *extension*. Some people hope for more. They hope to transit to an age where they can change to a substrate – go digital – where life is less fragile. But our investigation into computationalism above has shown that it is highly unlikely that this should work. Uploading a person will fail not because

the human mind is not a machine, but because the material processes in a brain – and the qualitative dispositions – are so very different from those in a different material medium. Achieving immortality in that way seems to be barred by metaphysical reasons.

This is actually a happy thought: imagine that it were possible to upload your personality onto a digital medium. Then it would be possible to make as many copies of your personality as possible. Sounds like heaven? Personal Backups? Eternal life in virtual scenarios? Well, it seems you can't create heaven without creating hell; because in such a scenario your *utmost priority* would be to *safeguard* your digital copies; if someone else would get hold of them, he could start duplicating you, make slaves out of you; or torture you in virtual environments.

But, as I said, metaphysics seems to rule out eternal life in a computer simulation after a personality-preserving upload. The same metaphysics also rules out negative scenarios – where your copies are stolen and abused. Such is the soteriology of a materialistic universe: that while there is no heaven, there also is no hell.

It seems that the price of community – that we can *share* experiences with each other – is death. To communicate, we have to have a *world* to refer to; that is, we need to be situated in a physical world and not be disembodied, eternal souls occupied in internal dialogue. A world which is static can't develop enough complexity to support cognitive entities. The world needs to be dynamic and full of change to give evolution a handle to build complex structures out of simpler constituents, complex structures which then relate to the world. But the cognitive entities thus generated, being of this world, have to underlie the very laws of the world and the *conditions of change*. This implies that beings of this world can't be immortal – because everything about these beings will change with time, so much so that finally we will have to admit that something has died and something new was born – see the next paragraph. Being immortal would mean possessing some invariable aspect that does not underlie change – at least some "core" aspect of you – what religions call the soul. It is not clear what purpose immutable souls could serve except providing some kind of ID-function. I do not think that that is what religions have in mind. Our meaning lies in our becoming – in the way we relate to the world and continue to relate to the world. What we are – our value – is I think more about what we *have* become – given our heritage and our possibilities – and what we *can, will* and *want* to become; rather than what we are at a single snapshot at time t. So our very ability to dynamically relate to other beings – that which makes life ultimately worthwhile – is what also forces us to die – in one way or another.

Some people will not be satisfied by this – they want heaven badly. So let us have a look at what it could possibly mean to live for eternity (as is usually the case in the afterlife). Let's start with shorter time spans: imagine that you are a conscious entity and that you will live for billions of years – remember Mr. Smith above? After a while, things will become boring. Cognitive dissonance will arise; and you can either try to search for new challenges, which will be ever more difficult to attain; or choose a simpler path: forgetting. And complete forgetting simply means living anew. Enjoying the world and each experience for the first time *again*. There is of course also the option of expanding your cognition abilities beyond recognition and keep total recall; again, I think this could be equated with your death as a *human being*.

A return to forgetful bliss – what is forgetting but dying and being born again? It is the circle of life and death. The knowledge that there is no substance to the self, that there is nothing over and above current physical instantiations, combined with panqualicism, is a doctrine of hope. Death is not the termination of a soul irretrievably lost, but just a transformation, *a forgetting*, a deletion of certain causal correlations (especially in the brain of the deceased). Experience continues elsewhere. A completely scientific view of persons is liberating:

> When I believed the Non-Reductionist View, I also cared more about my inevitable death. After my death, there will no one living who will be me. I can now redescribe this fact. Though there will later be many experiences, none of these experiences will be connected to my present experiences by chains of such direct connections as those involved in experience-memory, or in the carrying out of an earlier intention. Some of these future experience my be related to my present experiences in less direct ways. There will later be some memories about my life. And there may later be thoughts that are influenced by mine, or things done as the result of my advice. My death will break the more direct relations between my present experiences and future experiences, but it will not break various other relations. This is all there is to the fact that there will be no on living who will be me. Now that I have seen this, my death seems to me less bad. (Parfit 1984, p. 281)

An ethical aside[291]: you may ask why we should care about our future states when all experiences are universal anyway – what would constitute a special relationship with our future states as opposed to the future states of someone else? First of all, we see in this kind of

291 These things are deeply interrelated.

consideration the rational core of ethics – that there is no *essential* difference between *you* and *me*. We should[292] not react by not caring about *our* future selves, but caring about the future (and present) selves of *others* too.

There is an answer to the question of why we think more about *our* future states than those of other beings: it is because we have a privileged connection to our own future experiences due to our agency and our privileged control of our future states. While our future is influenced by many factors *outside of our control*, it is undeniably the case that *agency* confers on us some power to actively participate in the universe. We do not need free will for this, only causal power, and that is what we have. The weighing of alternatives (plans) and the possibly deterministic decision for one or another of these plans suffices.

Take smoking: I can, through rational deliberation weigh alternative plans and compare outcomes and then choose the optimal course of actions, namely, for any situation X, not to smoke in situation X. I can not do this for another agent. Insofar as we *can* influence the future states of other people, we should assume moral[293] responsibility.

But back to eternity: eternity, that is quite a bit longer than a few billion years. It means living through cycles of *universal* death and creation – crunches and bangs or other more strange contortions of matter. Immortality, in a way, is already equivalent to Godhood – but if one has become a God – what is there to do but split up again into the myriad shards of becoming and restart the sport of Lila?

Up to now we have dealt with *physical* conceptions of immortality – ones still tied to this universe. But what about Heaven? What do people actually believe in when they want to become immortal and go to Heaven? Is it to be maximally happy forever? To live the lives they have *failed* to live *now*? I think people look for certainty in heaven; that things will be good; and that they will *not stop* being good. But certainty is stagnation. Certainty is absence of development, lack of surprise. Again, if you have eternity to spend, you have to be imaginative about what you are going to do[294]. But what is there to do in Heaven? Either Heaven presents a challenging (immortal) life, and thus will be quite similar to the physical world. To be challenging, it must encompass the mode

292 Hypothetically should – if you are at all interested in being an ethical creature.

293 I do not want to distinguish moral and prudential concern here. When we dissolve the concept of person, the two become very similar.

294 People in Egan (1994) who have achieved immortality react for instance by "turning off" certain reflective routines and go into blissful repetitive states, such as mountain climbing or wood carving. Persons become sphexish and go into infinite loops of happiness.

of failure. Thus Heaven need also contain *adverse* conditions, which seems a little contradictory to the notion of Heaven.

Or maybe, in Heaven, you merely enjoy a state of eternal bliss – a state which is of dubious desirability, as I will remark on in section 5.5.1.2.

Now, I am not blowing in the same horn as neo-luddites[295]:

> Raise the topic of cryonics, uploading, or just medically extended lifespan/healthspan, and some bioconservative neo-Luddite is bound to ask, in portentous tones:
>
> "But what will people do all day?"
>
> [...]
>
> That doesn't mean it's a bad question.
>
> (Yudkowsky 2008i)

As Yudkowsky admits, it isn't a bad question. Now, I support Yudkowsky's conception of *fun theory*:

> Fun Theory, then, is the field of knowledge that would deal in questions like:
>
> - "How much fun is there in the universe?"
> - "Will we ever run out of fun?"
> - "Are we having fun yet?"
> - "Could we be having more fun?" (Yudkowsky 2008i)

Yudkowsky (2009) presents thirty one possible laws of fun; aspects worth striving for in a better world. I agree with most points. But I believe that human fun space is limited, in the sense that *human* fun space does not scale to *immortal* fun space (and time scales), and to transcend human fun space we would need to engage in large scale personal/societal transformation – so as to, for

295 This whole thesis is about optimism about the future, the ability to change ourselves, and make the universe a better place – with the help of reason and the products of reason, especially technology.

instance, engage in ever increasing novelty, a method for generating fun (Yudkowsky 2008j). But what is large scale transformation other than personal death?

Why attach so much importance to a continuation of psychological experience, when the continuation will be *severed* along the way? The persons we are now, even if we could "live" forever, would "die" along the way. If someone wants to follow this path, then nobody should stand in her way. I just want to alleviate *angst* which could ensue from trying to reach “immortality regions”. It is, on reflection, not so important. I do not argue against the possibility of transformation per se. I argue against the opinion that this constitutes personal survival.

In other words, I question the importance of the goal of personal survival as opposed to survival of sentience per se (which is a quite different question). The extreme significance attributed to *personal* survival is of course ingrained in us by evolution, and traditional worldviews piggy-back on this. But the attitude stems from not having fully accepted what it means to be a material state in a material universe. That does not mean that you need embrace death. Only that it need not be feared.

The basic property of the cosmos – eternal change – is it's redeeming feature. Whatever suffering there is, it is not forever. Compared to the vastness of being, it is as nothing. That is the *regulatory principle* of the materialistic universe missing in certain religious conceptions of eternal reward or punishment. *Change* then is the *soteriology* of the material universe. Things are forgotten so that new things may arise. Maybe Leibniz was right. Maybe we do live in the best of all possible worlds (Leibniz 1710).

5.4 Religion and Ethics

The relationship between religion and ethics is very problematic. While religions often claim that they are necessary for ethics, this is neither backed empirically nor stands up to conceptual analysis. First of all, there are different ways religions can operate: either with concepts like guilt, sin and authority[296], like the monotheistic religions; or with concepts of compassion, unity with the universe and respect for other beings; at the core of many of the eastern religions (Daoism, Shintoism, Buddhism, Jainism etc). Here I will just highlight one problematic aspect of the former.

The main problem of monotheistic religion today is not that it clashes with science; but its *deficient* ethics; an ethics suitable for tribesmen of two thousand years past, but not fit for the

296 The way these religions operate can be devastating for the individual (see Buggle (1992) for psychological effects) and damaging for society as a whole. (See Deschner (1986) for a “criminal history” of Christianity in ten volumes, nine of which have appeared.)

current age of weapons of mass destruction and the global society. The deficiency of this ethics is elaborated in the literature and need not be repeated here (Deschner 1986; Buggle 1992; Harris 2004; Dawkins 2006). One example will suffice to illustrate the whole problematic nature of an authoritative God as conceived in the monotheistic tradition. Hartung examines religious moral belief, drawing on empirical results delivered by Tamarin around the 70s:

> The world's major religions espouse a moral code that includes injunctions against murder, theft, and lying — or so conventional 19th- and 20th-century Western wisdom would have it. Evidence put forth here argues that this convention is a conceit which does not apply to the West's own religious foundations. In particular, rules against murder, theft, and lying codified by the Ten Commandments were intended to apply only within a cooperating group for the purpose of enabling that group to compete successfully against other groups. In addition, this in-group morality has functioned, both historically and by express intent, to create adverse circumstances between groups by actively promoting murder, theft, and lying as tools of competition. Contemporary efforts to present Judeo-Christian in-group morality as universal morality defy the plain meaning of the texts upon which Judaism and Christianity are based. Accordingly, that effort is ultimately hopeless. (Hartung 1995)

The problematic empirical result is this:

> The Israelites' campaign to carry out their god's commandment to commit genocide against the native inhabitants of Canaan-cum-Palestine took several generations. It began with Joshua's massacre at Jericho. Contrary to the Christian song "Joshua Fought the Battle of Jericho," according to scripture there was no battle at all. It was a siege, at the end of which all of the city's inhabitants were killed except Rahab the prostitute (she and her family were spared in exchange for helping Joshua plan his strategy, Joshua 6:16-17, 19, 21, 24, RSV):
>
> Joshua said to the people, "Shout; for the LORD has given you the city. And the city and all that is within it shall be devoted to the LORD for destruction ... But all silver and gold, and vessels of bronze and iron are sacred to the LORD; they shall go into the treasury of the LORD." ... Then they utterly destroyed all in the city, both men and women, young and old, oxen, sheep, and asses, with the edge of the sword ... And they burned the city with fire, and all

> within it; only the silver and gold, and the vessels of bronze and of iron, they put into the treasury of the house of the LORD.
>
> The half-life and penetrance of such cultural legacies are often under-appreciated. Some 3,000 years after the fall of Jericho, Israeli psychologist George Tamarin (1966, 1973) measured the strength of residual in-group morality. He presented Joshua 6:20-21 to 1,066 school children, ages 8-14, in order to test "the effect of uncritical teaching of the Bible on the propensity for forming prejudices (particularly the notion of the 'chosen people,' the superiority of the monotheistic religion, and the study of acts of genocide by biblical heroes)." The children's answers to the question "Do you think Joshua and the Israelites acted rightly or not?" were categorized as follows: " 'A' means total approval, 'B' means partial approval or disapproval, and 'C' means total disapproval." Across a broad spectrum of Israeli social and economic classes, 66% of responses were "A," 8% "B," and 26% "C."
>
> ...
>
> As a control group, Tamarin tested 168 children who were read Joshua 6:20-21 with "General Lin" substituted for Joshua and a "Chinese Kingdom 3000 years ago" substituted for Israel. General Lin got a 7% approval rating, with 18% giving partial approval or disapproval, and 75% disapproving totally. (Hartung 1995, p. 10f)[297]

The Bible, it seems, should definitely not be a part of children's education. We also do not *need* religion to give us a basis for ethics. A naturalist conception of the universe lends itself to support an ethics of its own. And that is why we will now turn to matters of value, meaning, morality and ethics.

5.5 You: Universal Values

> For if moral norms don't reduce to norms of reason or rationality, then we must ask in what sense moral norms could be authoritative. (Smith 2005, p. 2009)

297 The references in the quote are to Tamarin (1966); Tamarin (1973). RSV refers to the Bible, Revised Standard Version.

Is there universal value, that is, value that can or must, due to reason, be accepted by everyone? Acceptability by reason is a basic condition for further discussion; otherwise we already pre-commit to power struggle and there is no need for argument (see already above the chapter on rationality). Speaking of universal value does not mean that I commit to categorical imperatives or objective ethics in the sense that there is something writ in stone. It is rather an examination of the universe we live in and how it feels like to be part of this all which will lead to *insight*. Insight that can be attained rationally. As naturalists, committed to reason, we are in an excellent position to tackle ethical questions[298]. Universal value is not identical with ethics, but intertwined with the topic.

I believe that there is universal value. But the issue is subtle. What is important is to first completely draw away from parochial moral assumptions. Greene (2002) highlights the many problems inherent in current ethical dialogue in his thesis, the most prominent problem being that everybody thinks that *their* morality is right:

> Our minds trick us into thinking that *we* are *absolutely right* and that *they* are *absolutely wrong* because, once upon a time, this was a useful way to think. It is no more, though it remains natural as ever. We love our respective moral senses. They are as much a part of us as anything. But if we are to live together in the world we have created for ourselves, so unlike the one in which our ancestors evolved, we must know when to trust our moral senses and when to ignore them. (Greene 2002)

So let us for the moment ignore our moral senses. Nothing in universal value will be related to *human concerns*. Universal value – value that should be acceptable by every rational agent – is above all *universal* value in a more literal sense: value inherent in the *physical* universe. When, above, I say that this universal value should be acceptable by everyone, I also mean to include reasoning alien species evolved elsewhere in this universe. I don't think that any single clear-cut ethics for human beings will follow from universal value. It is a guiding principle – a nourishing principle – standing behind *concrete* moral considerations which will be more related to the time, place and actors involved.

298 Moore's "open question argument" poses no problem to a naturalist ethics. Arguments can be looked up in Casebeer (2003); Sturgeon (2006).

What do I believe that universal value is? *Diversity*. Diversity is *the* universal value (or is it a meta-value? – I think it is both). Diversity in this sense is not only tolerated or accepted, but actively desired – diversity celebrates the multifarious aspects of material configurations[299].

I have come to this belief through a number of considerations:

- value as intrinsic to the universe
- the fragmentation of value; and the failure of hedonism to account for value alone
- evolved multi-agent systems
- the *boon of diversity* in evolutionary settings

All of these considerations are naturalistic. And while I can't exclude that there still is some basic, human prejudice in all of the above, I am at least optimistic that there is *very little* such prejudice present.

Let us have a look at the points above in turn.

5.5.1.1 Intrinsic Value

Values are built into the very fabric of the universe. From panqualicism and the identity theory concerning the dispositional and the qualitative, we know that all our feelings, our qualia states, are *real* in a powerful way. They are not epiphenomena or metaphysical surplus baggage. There is no metaphysical possibility for the physical world to be as it is without it also having the qualitative aspects we observe.

The feelings of love and value are as much real as the negative charge of an electron. Of course, feelings like love or, say, the appreciation of liberty, are complex states, building on more primitive values, such as simple sensations, or pleasure and pain. But when combined in a complex web of causal interactions, such that psychological states arise, *the* experienced value will ultimately derive from those qualities right there from the beginning. Of course, (complex) intrinsic value *need not be aligned* with universal value, nor with any concrete ethics. Intrinsic value is simply that: a state of a physical system experiencing value. It is a primitive building block, that does not *by itself* yield anything more. We can, at this moment, only assert that "there exists a physical state X containing a mind Y which values Z".

299 We may compare the sum of all material configurations with the universal form of Krishna, visva-rupa, if we are in a mythological mood.

We now have our grounding for value: together with other features of the physical universe – especially rational agency and evolution – we can build the conception of the universal value of diversity.

Note that here I differ fundamentally from Drescher, who writes:

> But a merely mechanical state could not have the property of being intrinsically desirable or undesirable; inherently good or bad sensations, therefore, would be irreconcilable with the idea of a fully mechanical mind. (Drescher 2006, p. 77)

Drescher is *implicitly* in the grip of Cartesian dualism, though he *explicitly* renounces it. In the metaphysical picture painted here – causal physical[300] states are necessarily qualitative. Why should a physical state *not* have intrinsic value? The Cartesian intuition is revealed in the next passage, where qualia are given this origination-story:

> Actually, though, it is your machinery's very response to a state's utility designation — the machinery's very tendency to systematically pursue or avoid the state — that implements and constitutes a valued state's seemingly inherent deservedness of being pursued or avoided. Roughly speaking, it's not that you avoid pain (other things being equal) in part because pain is inherently bad; rather, your machinery's systematic tendency to avoid pain (other things being equal) is what constitutes its being bad. That systematic tendency is what you're really observing when you contemplate a pain and observe that it is ''undesirable,'' that it is something you want to avoid. (Drescher 2006, p. 77f)

While the contortion to accommodate qualia in a purely mechanistic picture is admirable for intellectual ingenuity, it is not necessary[301]: the present account of dispositional and qualitative identity does not feel so „forced" – and explains intuitively how dispositional states with the „right"

300 Drescher speaks of mechanical states. I would like to avoid this. Drescher seems to equate „mechanical" with „computational" and „physical":

> This book seeks to integrate several lines of inquiry that attempt to reconcile the mechanical nature of the physical universe (Drescher 2006, p. xiii)

> ... our […] cognitive abilities and phenomena—are indeed implemented mechanically, computationally, by our neurons. (Drescher 2006, p. 40)

As we have seen above, we must keep the three notions logically – and most plausibly also practically – distinct.

301 And, I believe, ultimately fails: because why, again, should the *tendency* to avoid or pursue states *feel* like anything. It seems that mind, while being paid lip-service, is actually eliminated. There is – I believe – no way around a panqualicist theory for a naturalist committed to the existence of consciousness.

qualia are selected by evolution – because their dispositionality *can't be separated* from their qualitativity.

5.5.1.2 Fragmentation of Value and Orgasmium

> I do not believe that the source of value is unitary – displaying apparent multiplicity only in its application to the world. I believe that value has fundamentally different kinds of sources, and that they are reflected in the classification of values into types. Not all values present the pursuit of some single good in a variety of settings. (Nagel 1979, p. 131-132)

> But the topic of value is also one where many different views are at least initially attractive. Some of these views value competing states of human minds, such as pleasure, knowledge, and virtue; others value patterns of distribution across these states, such as equality or the proportioning of happiness to virtue; yet others compare or aggregate goods differently, while a final group values states of the nonhuman environment. The debate between these views is not easily resolved, but its sharpness, and the way the competing positions all make plausible claims, only underscores the importance and fascination of issues about intrinsic value. (Hurka 2006, p. 377)

There is a lot of value in the world; decisions are difficult because many trade-offs are involved.

The introductory quotes illustrate the fact that many philosophers believe that value *is* fragmented. And there are good reasons to believe why they do not seem fundamentally reducible to something simpler, such as pleasure or pain, as hedonists would have it. Value, it seems, is fundamentally fragmented by evolutionary necessity. Yudkowsky puts it well in this passage[302]:

> And when we finally learn about evolution, we think to ourselves: "Obsess all day about inclusive genetic fitness? Where's the fun in *that?*"
>
> The blind idiot god's single monomaniacal goal splintered into a thousand shards of desire. And this is well, I think, though I'm a

302 The blind idiot god of the passage below is of course evolution:

> ... Darwin discovered a strange alien God - not comfortably "ineffable", but *really genuinely different from us*. Evolution is not a God, but if it were, it wouldn't be Jehovah. It would be H. P. Lovecraft's Azathoth, the blind idiot God burbling chaotically at the center of everything, surrounded by the thin monotonous piping of flutes. (Yudkowsky 2007g)

> human who says so. Or else what would we do with the future? What would we do with the billion galaxies in the night sky? Fill them with maximally efficient replicators? Should our descendants deliberately obsess about maximizing their inclusive genetic fitness, regarding all else only as a means to that end?
>
> Being a thousand shards of desire isn't always fun, but at least it's not *boring*. Somewhere along the line, we evolved tastes for novelty, complexity, elegance, and challenge - tastes that judge the blind idiot god's monomaniacal focus, and find it aesthetically unsatisfying.
>
> And yes, we got those very same tastes from the blind idiot's godshatter. So what? (Yudkowsky 2007h)

We value certain states because of our evolutionary heritage. But evolution operates in a complex world, and this complexity is mirrored in the plentifulness of goals that have been correlated with pleasurable qualitative states. But just because our value is *grounded* in qualitative states – such as those of pleasure and contentment – does not mean that they *reduce* to those building blocks in any meaningful way beyond the grounding. If there were a single value – such as pleasurable qualia states – we should be able to achieve the best possible ethical state by maximizing this state.

For example, experiencing an orgasm is very pleasurable. If we could reduce value to qualitative states and assume for the time being that orgasm is the best possible qualitative state, then the best universe would be one where all matter is in a state of perennial orgasm (where matter in a state of orgasm will be called orgasmium). But this conclusion seems absurd.

It would take so much of value out of the lives that we know. It seems that in our lives we care more about just being "blissed out" in eternal orgasm. This does not seem to be a good life to live. Yudkowsky hones in on the same point:

> When I met the futurist Greg Stock some years ago, he argued that the joy of scientific discovery would soon be replaced by pills that could simulate the joy of scientific discovery. I approached him after his talk and said, "I agree that such pills are probably possible, but I wouldn't voluntarily take them."

And Stock said, "But they'll be so much better that the real thing won't be able to compete. It will just be way more fun for you to take the pills than to do all the actual scientific work."

And I said, "I agree that's possible, so I'll make sure never to take them."

Stock seemed genuinely surprised by my attitude, which genuinely surprised me.

[...]

It is an undeniable fact that we tend to do things that make us happy, but this doesn't mean we should regard the happiness as the only reason for so acting. First, this would make it difficult to explain how we could care about anyone else's happiness - how we could treat people as ends in themselves, rather than instrumental means of obtaining a warm glow of satisfaction.

[...]

The best way I can put it, is that my moral intuition appears to require both the objective and subjective component to grant full value.

The value of scientific discovery requires both a genuine scientific discovery, and a person to take joy in that discovery.

[...]

So my values are not strictly reducible to happiness: There are properties I value about the future that aren't reducible to activation levels in anyone's pleasure center; properties that are not strictly reducible to subjective states even in principle.

Which means that my decision system has a lot of terminal values, none of them strictly reducible to anything else. Art, science, love, lust, freedom, friendship...

And I'm okay with that. I value a life complicated enough to be challenging and aesthetic - not just the feeling that life is

> complicated, but the actual complications - so turning into a pleasure center in a vat doesn't appeal to me. It would be a waste of humanity's potential, which I value actually fulfilling, not just having the feeling that it was fulfilled. (Yudkowsky 2007i)

But why does happiness alone not suffice? Do we need the tension? The possibility of unforeseen change? The drama? Maybe we are simply lying to ourselves and it is only that we can't appreciate the full force of the argument – that, given evolution, all our valued states are indeed only instrumental, and that the perfect state of matter is to be orgasmium.

Maybe hypothetically. I do not know. *We can rule out such speculations in multi-agent scenarios*, because orgasmium will be unable to compete in evolutionary settings. In the real universe, we *do* have multi-agent scenarios. That is enough for the fragmentation of value to take hold, and for agents having been built by evolutionary algorithms in this universe to *genuinely* value this fragmentation[303].

5.5.1.3 Multi-Agent Systems

The basic building blocks of value and meaning are present in the fabric of the universe in the sense of qualia-ness, likeness. No qualia, no value. There is no value in a Zombie world. But value can constitute itself into quite different arrangements, dependent on the physical parameters of the system in question. Intrinsic value of physical states come into alignment with ethical and moral values via their selection in evolutionary settings, in agent-societies; indeed, it is only there that moral considerations apply in the first place. Above, we said: no qualia, no value. Here we can add: no agents, no morality. So, the two requirements for morality are the existence of qualia states and the existence of agents (P-Beings). We can imagine the possibility of "bad" alignments when imagining a human being who takes pleasure in seeing others suffer. Seeing others suffer has value for that person; of course, this kind of value is not tied to behavior or morality in an human society in any acceptable way. That is why this kind of behavior will not be tolerated and selected against.

So, evolution brings – at least, for most creatures in the population (there is always variation) – intrinsic pleasurable states into alignment with evolutionarily successful ones. We know from

303 Nowhere are we committing a naturalistic fallacy here. The naturalistic fallacy actually splits up into eight sub-cases – see Curry (2006). In fact, it is time to stress the other side of the equation – the anti-naturalistic fallacy:

> ...we must recognize that while not all natural facts are relevant to ethical or moral discourse, all facts that are relevant to ethical and moral discourse will nonetheless be natural facts. To hold that values are non-natural facts is to commit the anti-naturalistic fallacy. (Walter 2006)

simple games that altruism and cooperation can (and will) arise in evolutionary settings[304] (Axelrod 1984), see also Binmore (2005) and Nowak (2006). Drescher presents the following analysis for rational actors and (correctly) locates altruism in *subjunctive reciprocity* (Drescher 2006, p. 273f), thereby foregoing the need for strict causal links between reciprocal actions:

> The claim here is that rational moral regard reduces to subjunctive (not necessarily causal) reciprocity: roughly, you act as you want others to act toward you, because if that were the rational choice for you, it would like-wise be the rational choice for them; and if they are (more or less) rational choosers, they would (probably) make what is the rational choice for them, which would then be to your advantage. (Drescher 2006, p. 289)

The message of this section is summed up by the following passage:

> Rather than recommending particular solutions to problems, evolutionary game theory, coupled with the theory of bounded rationality and recent work bridging the gap between psychology and economics, provides what appears to be a radical restructuring of the foundations of moral theory. [...] the recommendations, constraints, and obligations imposed by moral theories are real and binding – but also somewhat arbitrary. If we were different kinds of creatures, and if our societies were structured differently, our lives would be composed of very different interdependent decision problems. Consequently, the moral theories which legislate certain actions as a means of solving those problems would also be different. This means that our moral beliefs are simultaneously relative to our evolutionary history and our cultural background, but at the same time objectively true. Insofar as our moral beliefs provide solutions to interdependent decision problems, we cannot say that any one solution is better than any other – in an abstract sense – because, detached from our preferences, there is no absolute standard from which to judge. Given our preferences, and from our own personal point of view, there can be an objective moral theory that prescribes the best way of satisfying those preferences. (Alexander 2007, p. 291)

304 The most well-known example of such a game is the Iterated prisoner's dilemma. Iteration is important, because cooperation can only evolve where agents interact over time. The tit-for-tat strategy is the most successful one this game, winning even against complex strategies: tit-for-tat always starts with cooperation, and never defects by itself. But it retaliates when another player defects.

We have experiences – qualia – and there are pleasant ones and not so pleasant ones. If we don't think about humans alone, but about all of possible qualia space (inhabited by super-minds etc) - then we will find that in the qualia landscape there are locations that are enjoyable and others that are less so. For a universe made up of agents – having to deal with "others" – there will be *pleasant qualia states attainable in harmony with other agents, that is, that do not lead to behavior that induces unpleasant qualia states in other agents*. And that is about as "objective" as the naturalist ethics endorsed in the present thesis will get. But it is a lot. And it is a far cry from relativism; albeit still being able to encompass a lot of diversity: the diversity found in evolved structures will be reflected in a diversity of values.

5.5.1.4 Evolution

I already commented on the boon of diversity for evolution in section 3.1.7 and immediately above; the basic point is simple: as we can't foresee future environmental changes, the greater the diversity in life-forms and behavioral habits the better. The principle of diversity applies everywhere where evolution is at work, not only in biological contexts.

5.5.1.5 You

I have put this section on universal value under the heading of "you" for a simple reason: while diversity comes in many material forms, not all of which are conscious in an interesting sense, when we speak of value and agent systems our interest immediately shifts to those Q-Beings which are our partners in this world. The *you* is then the basic recognition that diversity begins with the *other*. The *you* represents the universal value of diversity. *I value that you exist, and I value your difference from me*[305].

From the recognition that we are simply material patterns in motion, going through various transformations – indeed, from our elimination of the concept of person for all but practical purposes – we can arrive at the position of universal love[306]. We can call the structure – the concrete structure constituting a person – a cosmic perspective[307]. But no perspective can be given any reasonable privileged position. The feeling of being "yourself" does not correspond to a

305 That is not to exclude the importance of the many aspects that we share. But in ethical considerations, it is usually the differences that are more problematic.

306 Kolak for instance argues that there is only *one* person, making this the basis for ethics (Kolak 2004). Arguing for the no-person view or the one-person view is, from the concerns of this thesis, beside the point. The no-person view is more in line with the present metaphysics.

307 People who inquire into the nature of things are the universe discovering itself.

metaphysically interesting fact: it is only the presence of neural structure encoding an indexical in a brain.

The elimination of soul-essence and the insight into the nature of indexical being is a crucial step into the equality of being. The question: "why am I myself" and not someone else only makes sense in a soul-setting – the implicit belief that you are a soul and have been randomly assigned a "body" and a "life". When you eliminate the soul, the question "why am I myself" can be posed by any entity and the answer will always point back to the questioning entity itself, because there never was a combination of previously separate structures. What is experienced now – by you – is only one of a multitude of experiences in the universe. It is in no way privileged in time, in space or otherwise. All is equal.

As we begin to know our true nature – as the ultimate causal structure of the world reveals itself – and we try to reflect this knowledge on ourselves and our fellow living beings we find that we all underly the same conditions of suffering and joy; that we have the same desires, the same hopes, and the same wishes (at least among humankind). The ultimate instantiation of qualia states in us is beyond our control; so the most we can hope for is helping *each other* attain a better life. Compassion comes with knowledge. What's more, the rationalist knows that there is no one else except the agents in the universe who can change things for better or for worse. From this stems the urge to work for the happiness of agents, and quell suffering.

One may construct a more traditional argument for universal love, such as done by Flanagan:

1. If there is something I desire for its own sake and recognize that everyone else wants the same thing, then I ought to believe that everyone has a right to that thing.

2. Whenever I recognize that I ought to believe something, I believe it.

3. I desire to flourish (not suffer, be happy).

4. I recognize that everyone else wants to flourish (not suffer, be happy).

5. I ought to believe that everyone has a right to flourish.

6. I believe that everyone has a right to flourish.

> This argument is valid or can be made so. If it is, then everyone who believes the premises must believe that everyone has equal right to flourish. (Flanagan 2007, p. 214)

It is now time to move on to domains that are more down to earth than the present section – and see what we can derive for societies in a naturalistic setting.

5.6 We: The Polity and the Stars

5.6.1.1 A Polity of Toleration

> You can only protect your liberties in this world by protecting the other man's freedom. You can only be free if I am free. Clarence Darrow / Address to the court in People v. Lloyd (1920)

> I would defend the liberty of consenting adult creationists to practice whatever intellectual perversions they like in the privacy of their own homes; but it is also necessary to protect the young and innocent. (Clarke 1984, p. 265)

Given that our nourishing principle is diversity, toleration will play an important role in human culture. But toleration alone is not enough – one also needs a *Gemeinwesen*, a community where toleration can be practiced. Many people grow up and are socialized in parochial communities. In todays society, everyone sooner or later realizes that there is more than one way to view the world: this may come as a shock. Some react by becoming xenophobic, staving off other worldviews; others embrace relativism, and disillusioned by having their childhood "truths" destroyed acknowledge no truth at all. The mature reaction, I contend, is to realize that there are arbitrary human conventions, sometimes enforced as absolute truths but which are only historical contingencies (such as religion or local customs); and apart from that there are facts about the world which are quite independent of humans or other cognitive agents. These latter facts should be the foundation for our community.

To have a community, there must be shared experience and shared language. We share experiences because we live in the same universe and are very similar in physical structure. As to the language, we find a direct connection to the metaphysical aspect of the thesis, where we spoke of the process of *registering:*

> […] all of human communication lies in this middle region, between identical and incommensurable registration schemes. Not only that,

> when we get back to the proper metaphysical story,[...] it will emerge that, far from being undermined by it, registration is designed to cope with this middle region of partial commensurability.
>
> [...] developing a more adequate metaphysics is not only intellectually viable, but also politically urgent. The aim is to give metaphysical grounding to, and support for, communicative and political struggles among people whose experience of, and participation in, our world is different. (Smith 1996, p. 255)

Now as was argued above, science give us *robust* registries of the world – concepts that are invariant from many different points of view. Thus, the concepts of science can deliver the basis for a global language and thus the basis for global community. Science is about discovering the shared environment, making ever more features of reality available for communication. I would like to call the framework given by science the *polity*. It is the public space of quanta, corresponding to the communicable world; in the polity, people can interact, empathize, co-operate, help each other. The polity is a haven of discourse, a forum of discussion and consensus:

> Wissenschaft ist viel mehr als nur eine Ansammlung von Fakten. Sie ermöglicht Menschen, die durch Ozeane voneinander getrennt sind, in unterschiedlichen Dekaden leben, verschiedene Sprachen sprechen oder anderen Ideologien unterliegen, wechselseitig auf den Entdeckungen der jeweils anderen aufzubauen. (Wilson 2006)

The polity must be open to encourage diversity and guarantee the playful unfolding of events. This leads to More's proactionary principle and a positive view of technology[308]:

> People's freedom to innovate technologically is highly valuable, even critical, to humanity. This implies a range of responsibilities for those considering whether and how to develop, deploy, or restrict new technologies. Assess risks and opportunities using an objective, open, and comprehensive, yet simple decision process based on science rather than collective emotional reactions. Account for the costs of restrictions and lost opportunities as fully as direct effects. Favor measures that are proportionate to the probability and magnitude of impacts, and that have the highest payoff relative to their costs. Give a high priority to people's freedom to learn, innovate, and advance.

308 For more on the proactionary principle, see Appendix F: The Proactionary Principle.

…

> Most activities involving technology will have undesired effects as well as desirable ones. Whereas the precautionary principle is often used to take an absolutist stand against an activity, the Proactionary Principle allows for handling mixed effects through compensation and remediation instead of prohibition. The Proactionary Principle recognizes that nature is not always kind, that improving our world is both natural and essential for humanity, and that stagnation is not a realistic or worthy option. (More 2005)

The proactionary principle is contrasted with the precautionary principle – which comes in various degrees of strength and averseness to technology – prevailing today. The precautionary principle is detrimental to development:

> The precautionary principle has at least six major weak spots. It serves us badly by:
>
> 1. assuming worst-case scenarios
> 2. distracting attention from established threats to health, especially natural risks
> 3. assuming that the effects of regulation and restriction are all positive or neutral, never negative
> 4. ignoring potential benefits of technology and inherently favoring nature over humanity
> 5. illegitimately shifting the burden of proof and unfavorably positioning the proponent of the activity
> 6. conflicting with more balanced, common-law approaches to risk and harm.
>
> ...
>
> If the precautionary principle had been widely applied in the past, technological and cultural progress would have ground to a halt. Human suffering would have persisted without relief, and life would have remained poor, nasty, brutish, and short: No chlorination and no

> pathogen-free water; no electricity generation or transmission; no X-rays; no travel beyond the range of walking. (More 2005)

The idea, of course, is not to engage in Panglossian naiveté; but to restore symmetry to the process of weighing the pros and cons of innovation.

The polity is guarded by science; no other enterprise can stand in its place, because science is per definition the most refined methodology of studying the space of communicable facts – were better methods discovered, they would simply become part of science's arsenal.

And that is also where tolerance and liberalism must be weighed against the good of protecting the future of the polity. Toleration is a strange beast. We must distinguish between indifference/neutrality – things or situations we simply don't care about one way or another; and affirmation – states of affairs that we actively endorse. Then again there are those opinions and deeds which are so destructive to the life of a community – such as criminal behavior – that all communities which care to function forbid them.

Toleration is called for when we encounter behavior, opinions and deeds which we think are wrong; wrong in such a way that they endanger core values we hold. Of course, we need not stand passively by states of affairs that we oppose but tolerate – we can fight with arguments – but not with the method of censure and prohibition. Toleration stands on the verge of prohibition. But who is to decide what is still allowed and what prohibited? Frederick Schauer for instance says:

> Freedom of speech is based in large part on a distrust of the ability of government to make the necessary distinctions, a distrust of governmental determinations of truth and falsity, an appreciation of the fallibility of political leaders, and a somewhat deeper distrust of governmental power in a more general sense. (Schauer 1983)

And that is an important point. Who is to decide were tolerance ends and prohibition begins? This is not the place to solve this issue; I do not think that there is a solution in the abstract – those things will have to be negotiated by the actors at hand. I just want to highlight that this is the linchpin of success or failure for a society based on diversity – to find the right balance between tolerance of practices which may even *contradict* the goals of a society endorsing diversity, and the prohibition of opinions and practices which are too extreme and dangerous and endanger the continuity of diversity. While there are no solutions in the abstract, there are practical guidelines:

> that the practice of toleration has to be sustained not so much by a pure principle resting on a value of autonomy as by a wider and more mixed range of resources. Those resources include an active skepticism against fanaticism and the pretensions of its advocates; conviction about the manifest evils of toleration's absence; and, quite certainly, power, to provide Hobbesian reminders to the more extreme groups that they will have to settle for coexistence. (Williams 1996, p. 26f)

Things are not settled when we have constructed a society of toleration and formal liberal rights; because such a system can be undermined. Scanlon addresses this issue:

> I began by considering the paradigm case of religious toleration, a doctrine that seemed at first to have little cost or risk when viewed from the perspective of a secular liberal with secure constitutional protection against the "establishment" of a religion. I went on to explain why toleration in general, and religious toleration in particular, is a risky policy with high stakes, even within the framework of a stable constitutional democracy. The risks involved lie not so much in the formal politics of laws and constitutions (though there may be risks there as well) but rather in the informal politics through which the nature of a society is constantly redefined. I believe in tolerance despite its risks, because it seems to me that any alternative would put me in an antagonistic and alienated relation to my fellow citizens, friends as well as foes. The attitude of tolerance is nonetheless difficult to sustain. It can be given content only through some specification of the rights of citizens as participants in formal and informal politics. But any such system of rights will be conventional and indeterminate and is bound to be under frequent attack. To sustain and interpret such a system, we need a larger attitude of tolerance and accommodation, an attitude that is itself difficult to maintain. (Scanlon 1996, p. 238)

A free society is an on-going project. The best guarantee to repel the "frequent attacks" against a system of rights upholding a tolerant society is – apart from stable economic conditions – education and schooling in the art of rationality. Rationality as defined in this thesis is intrinsically inimical to intolerance.

5.6.1.2 The Stars

> There is hopeful symbolism in the fact that flags do not wave in a vacuum. A.C. Clarke as quoted in (Merchey 2004, p. 31)

The polity is one aspect of society. Another is a vision of the future. I have already mentioned this aspect above in section 2.9.7 on existential risks. It is important to realize that a scientific rational worldview does not preclude large-scale visions of the future of humanity; on the contrary, these thoughts are encouraged. The vision is, of course, space[309]; either to connect with other sentients, or, if we are alone, to ensure the continuation of life.

Intelligent life may once even be responsible for the very fate of the universe. That the universe will end under standard cosmological models is common knowledge (Adams & Laughlin 1997). But these calculations do not take into account the advent of intelligent agents. Agents are ways of matter organization which start engaging with other matter in feedback loops. From complex systems science, we know that new macro-behavior will most likely result (Waldrop 1992), thus changing the equations of the future development of the universe. In a scientific conception of things, sentients are not puppets in a theological game of good versus evil, but empowered actors shaping the fate of the universe[310]:

> Stars are born and die; galaxies go through their cycles of creation and destruction; the universe itself was born in a big bang and will end with a crunch or a whimper, we're not yet sure which.[...]The mindless mechanism of the universe is winding up or down to a I distant future, and there's nothing intelligence can do about it.
>
> That's the common wisdom. But I don't agree with it. My conjecture is that intelligence will ultimately prove more powerful than these big impersonal forces. [...]
>
> So will the universe end in a big crunch, or in an infinite expansion of dead stars, or in some other manner? In my view, the primary issue is not the mass of the universe, or the possible existence of antigravity, or of Einstein's so-called cosmological constant. Rather, the fate of the universe is a decision yet to be

309 While the colonization of space may be economically less attractive than first colonizing less hospitable parts of the Earth, such as Antarctica, the deep sea or diverse deserts, it is the visionary aspect which makes space the more important longterm choice; that, and the advantage of not limiting humanity to one celestial body, which is risky.

310 See also Dyson (1979); Hartung (1996).

> made, one which we will intelligently consider when the time is right. (Kurzweil 1999, p. 258f)

Maybe this sounds too much like science fiction. It is not important to decide on this issue now.

The polity can be seen as a gem in the void – an abode of sentients who have committed to the living universe. What then, is the Utopia of the polity? A society of eternal enlightenment and eternal disputation. A society which values life and consciousness, which recognizes how precious every living thing and every living moment is in and for this universe. The ethics of this society will be an ethics of diversity and possibility – of self-expression and it's sharing. The ethics of diversity will be restrained by compassion. Technology will be valued greatly, but not to enslave thinking beings, but to serve them: it will guarantee the survival of thought; enhance our full potential as creative and sentient beings, and make this universe a friendlier place to live in.

5.7 I: To Live

> There is no point telling people to promote the good without telling them what the good is.(Hurka 2006, p. 357)

> The primary experimental result in hedonic psychology - the study of happiness - is that people don't know what makes them happy.(Yudkowsky 2008i)

There are two aspects which I want to address in this last section. The first is metaphysical in nature: metaphors one can use to embrace the life in a completely naturalistic universe. The other is more practical: what conclusions to draw for one's personal life.

First to the metaphysical aspect: the vastness of the cosmos does not make us insignificant: on the contrary, the realization that the cosmos is so empty and that this earth and life are so rare shows the preciousness of this all. We are not what the universe was made for; but we are manifestations in this universe, valuable manifestations; no more, no less. We have a duty toward the universe, and this duty is: To Behold! To Think! To Live!

There are many ways to relish the cosmic perspective. Barbour expresses it most beautifully, having first established that the only theory of the cosmos worth taking seriously is one that sees it as a "heavenly vault" where the "music of the spheres" are sounding. He goes on:

> You will naturally ask why we do not hear this music of the spheres. Keats provides a first answer: 'Heard melodies are sweet, but those

> unheard Are sweeter'. But Leibniz may have given the true answer. In his monadology, he teaches that the quintessential you, everything you experience in consciousness and unconscious, is precisely this music. You are the music of the spheres heard from the particular vantage point that is you. (Barbour 1999, p. 326)

This is in tune with the vedantic "tat tvam asi" – "that thou art", a Mahavakya (Great Saying) from the Chandogya Upanishad[311]. It says what has in this thesis been called "no essential disconnection".

The biggest help in accepting a naturalist conception of the universe is lent to us by Nietzsche; who has embraced naturalism and all its consequences for persons *on* a personal level more than a hundred years ago. Here, finally, we arrive at the liberating message of materialism and science:

> If we affirm one single moment, we thus affirm not only ourselves but all existence. For nothing is self-sufficient, neither in us ourselves nor in things; and if our soul has trembled with happiness and sounded like a harp string just once, all eternity was needed to produce this one event – and in this single moment of affirmation all eternity was called good, redeemed, justified, and affirmed. (Nietzsche Will to Power, p. 532-533)[312]

We must understand this radical affirmation of life – of this earthly, material, bodily life, against the backdrop of the doctrine of the eternal return, articulated beautifully here:

> The greatest weight. — What, if some day or night a demon were to steal after you into your loneliest loneliness and say to you: "This life as you now live it and have lived it, you will have to live once more and innumerable times more; and there will be nothing new in it, but every pain and every joy and every thought and sigh and everything unutterably small or great in your life will have to return to you, all in the same succession and sequence—even this spider and this moonlight between the trees, and even this moment and I myself. The

311 The other Mahavakyas are:

- Prajnanam Brahma; "Consciousness/Knowledge is Brahman" (Aitareya Upanishad)
- Ayam Atma Brahma; "This Atman is Brahman" (Mandukya Upanishad)
- Aham Brahmasmi; "I am Brahman" (Brhadaranyaka Upanishad),

where (probably to the horror of scholars in Hinduism) I will equate Atman with the Self and, as above, Brahman with ultimate reality, the Dao.

312 I would like to dedicate this passage – especially "and if our soul has trembled with happiness and sounded like a harp string just once" – to Ainur, who has given me this moment.

> eternal hourglass of existence is turned upside down again and again, and you with it, speck of dust!" Would you not throw yourself down and gnash your teeth and curse the demon who spoke thus? Or have you once experienced a tremendous moment when you would have answered him: "You are a god and never have I heard anything more divine." If this thought gained possession of you, it would change you as you are or perhaps crush you. The question in each and every thing, "Do you desire this once more and innumerable times more?" would lie upon your actions as the greatest weight. Or how well disposed would you have to become to yourself and to life to crave nothing more fervently than this ultimate eternal confirmation and seal? (Nietzsche GayScience) (341)

Of course, the eternal return thus conceived is not terrible at all – the experiences would coincide if they where exactly the same; so, maybe we live our lives in eternal recurrence – we would not notice. Maybe we also live our lives in infinite variation, in case there is room for an even more benign soteriology than the materialist one presented above (Steinhart 2004). But that is not the point. The point is to *love every moment*; in such a way that you can love it even if it would recur eternally and you knew about these recurrences in the respective moments. It is a powerful imperative to *live one's life now* and not squander it – and especially not delegate it into an afterlife, be it of the religious or also scientific kind (the uploading scenario, or infinite life extension); it is the ultimate affirmation[313] (Reginster 2006).

After these rather heavy metaphors some more down to earth advice.

What does it mean to lead a good life? To lead a good life, we must[314] first consider what is the case – we must try to form a correct world model; that is where rational and scientific processes step in. This includes evaluating our current agent situation – our embodiment as human beings. Then we can ask what goals we *want* to achieve – and I am sure that many possible answers can be

313 There is of course also the option of viewing life and the universe purely as an *aesthetic phenomenon:*

> For we need to be clear on this point, above everything else, to our humiliation *and* ennoblement: the entire comedy of art does not present itself for us in order to make us, for example, better or to educate us, even less because we are creators of that art world. We are, however, entitled to assume this about ourselves: for the true creator of that world we are already pictures and artistic projections and in the meaning of works of art we have our highest dignity — for only as an *aesthetic phenomena* are existence and the world eternally *justified* — while, of course, our consciousness of this significance of ours is scarcely any different from the consciousness which soldiers painted on canvas have of the battle portrayed there. (Nietzsche Birth of Tragedy)

314 This is only valid for those who are interested in a good life, of course. Only hypothetical imperatives in this world.

given; after all, value is fragmented. As rationalists we can see that there are states worth attaining, and states that are worth avoiding. We can then plan to attain states that are worth attaining.

For all this, we need knowledge. What purpose serves knowledge? First, to unlearn. We must learn to unlearn everything false that we have been imprinted with by our culture and our elders. We must observe and think and reflect to eliminate all superstition, spurious correlations and false causal models in our minds. The knowledge that most of our knowledge is false – and will be false in the future – is the starting point for abandoning our "self-imposed immaturity"[315]. So learning first of all means: unlearning and forgetting.

The next step is to acquire positive knowledge; knowledge which reflects the states of the world and their transitions; that is the process of growth. Knowing is essential for acting; calling something knowledge means that the models such described have enough correlation with the world to enable successful action. The more in tune with the world the models are, the more potential there is for action; this is vividly seen by considering technology as "embodied" knowledge. Ignorance condemns one to inaction and passivity; and make plausible appeals to fate and to the conception of being helpless puppets in either an inimical universe or at the mercy of a God.

Growth of knowledge on the one hand is increased potential for action. But this must be combined with ethical diligence, lest one become a tyrant. In psychological terms, growing means above all: letting the *ego* become small. Not taking the self too seriously; to see how we are dependent on everything around us; how we are caused by past events on which we had no influence. And how future events will be influenced by us; and our responsibility for other agents when we act.

Now science seems well apt to fill both roles: that of furthering the goal of increased action potential but also of making us small (think of the Copernican, Darwinian and Freudian turns in our conceptions of ourselves as human beings). That is why we should heed the results of science in our personal lives. While knowledge alone does not make one a better human being, there is certainly no growth without knowledge. Science is the only reliable method to obtain knowledge of the world and of ourselves. Science shows us our place in the world – our potential and our limits. Both science and ethics are principally about a curtailing of the "self" and a recognition of the other. Without the other, the self becomes empty. Recognizing this is the precondition for further development.

315(Kant 1784)

This leads to the concept of wisdom. Wisdom, indeed, is even more broad than the age-old question of a happy life, eudaimonia. Wisdom is rationality bent back on itself; assessing one's goals and thinking about which ones are worth pursuing and which ones are not. Goals are available for updating. Are there physical constraints? Can I change them? We have to accept environment and agent body as given; indeed, we saw that without these, it is hard to see how one could be rational at all. But, then, *given these*, we can find values through rational reasoning which apply us humans.

There is a lot of merit to be found in the Buddhist way (although I do not endorse the more religious aspects of Buddhism). Given that it is impossible to invent a workable ethics from scratch, it is rational to draw on ethical traditions which are close to a scientific ethics:

> Buddhist values are rooted in the project of overcoming greed/attachment, hatred and delusion, which are seen as the roots of unwholesome actions and the key causes of suffering. Greed is to be overcome by generosity and sharing, combined with restraint from theft and cheating, with subtler forms of attachment overcome by monastic training and meditative training. Hatred and anger are to be dealt with by restraint from behaviour harming others, cultivation of loving kindness and compassion, and insight into the distorted vision that makes hatred possible. Delusion is to be overcome by avoiding intoxication, and cultivating the mental clarity that allows one to see things directly 'as they really are'. This project begins with moral virtue, but also entails the other aspects of the Buddhist path: meditative development and the cultivation of insight. It has implications for individual conduct as well as inter-personal relationships and social ethics. (Harvey 2000, p. 122)

In the same way that the path of rationality is not easily embodied, ethics is neither. (Schwitzgebel & Rust Forthcoming) – ethics has to be grown into behavioral structures. An ethical way of living must be cultivated and shaped every day by concrete actions.

While above we discussed value in a very abstract sense, we humans are faced with the challenge of finding meaning in our lives. Advice can be found in the excellent book of Flanagan (2007):

> Everything [...] is compatible with the picture of persons that emerges from neo-Darwinian theory and from the best current mind science. According to that picture, we are fully embodied thinking-

> feeling animals who live and achieve meaning — if we do — in a world that is fully natural. We are agents, and we act freely. But we do not possess any non-natural faculty of free will that permits circumvention of natural law. When we die, our career as a conscious being is over. But we leave effects. Our karma, good or bad, carries on. This matters. So it is wise to live well, in a way that makes meaning and sense in a manner that alleviates suffering and equips others to pursue what our common humanity makes us seek. If you live with your eye on the prize, then when you die, although you won't go to heaven, you'll have lived in a worthy way and have something to be proud of. (Flanagan 2007, p. 61)

When asking about what to *actually* do with your life, it is good to extract subquestions from the "big" question of the meaning of life:

> Instead of asking "What is the meaning of life?" we can ask such tame but very difficult questions as these:

- How shall I (or we) live?
- What ways of being and living produce fulfillment and meaning?
- What attitudes and beliefs about such matters as my place in the universe is it sensible to adopt?
- How can I understand my life's meaning, given that I am mortal?
- Given what I know about my talents, aspirations, hopes, and expectations, and given what I know about the existing network of social support, what sort of sensible plan can I make about how to live?

> I will not spell out my own answers. I can't do so fully; they are unfolding — works in progress. We are each, with social support, supposed to find our own way to the answers. Spiritual and philosophical traditions often do the following work: They give us a head start in asking and answering these questions by being repositories of past "good answers." Aristotle, Confucius, and Buddhism, all in different ways, see productive and respectful social relations as time-tested ingredients of meaningful and fulfilling lives. (Flanagan 2007, p. 202f)

And death, finally, need not be feared. I am aware that fear will not subside by reading the paragraph below; but I can assure the reader that fear will subside *after* deep meditation on this question, and then the below will read like the most natural thing in the world:

> Many spiritual traditions moralize death by embedding a life story in what […] I called a karmic eschatology. There is a payoff system that kicks in after you die. Your fate is determined by the moral quality of your earthly life, and in a surprising number of cases, by whether you believed in the ''true God.'' This last idea can be said, but it can't be asserted.
>
> Here is the way I think about my own death, given that I think karmic eschatology(ies) makes no sense (I don't even get why anyone would find the idea of living forever, even blissfully, very appealing). I recently heard a wise Buddhist friend say that ''death is the ultimate absurdity, you lose everything you care about.'' This, it seems to me, is not true. Furthermore, it is not a particularly Buddhist way (even for a secular Buddhist) to see things. Here is a better way: If you live well, then when you die you lose nothing you care about. Why? Because you are no longer there. You are just gone. That which is gone has nothing to lose. That which was once something, but is now nothing, cannot suffer any loss. But assuming the world and the people in it, including your loved ones remain, then your good karmic effects continue on. This is something to be proud of and happy about while alive. Your goodness, your presence, your worth are why the living feel your loss, and are sad, possibly very sad. But you are not sad, you neither suffer nor experience any loss because you are gone. Nothing absurd has occurred. True, dying could be miserable, but your own death is nothing to worry about. (Flanagan 2007, p. 203f)

There is but one fear we should entertain, and that only in the fashion that it spurns us to right action: that of not living to our full potential. And when we notice that we have failed to meet our potential in the past, we should not be filled by remorse, but look forward with a smile on our face and say: *from today on, I will do better.*

So, from rationality flows affirmation of every living moment. Ultimate rationality means being completely in tune with reality. Processes in the finite brain model as optimally as possible processes in the world; thus, all actions are in tune with the Dao. Automatically, no evil is done – because the fundamental unity of all being is seen; love of life and consciousness guide your

behavior. If we see ourselves as manifestations of the Dao – as necessary beings in tune with reality that have never been separate from the universe, as *instantiations of playful physical abandon* – we can be comfortable with simply *being* now and here. A person is a bundle of contingencies, produced by the universe to enjoy its fullness.

And if, in some hour of despair, the transvaluation of values proposed here seems to be too much to bear – never forget Egan's Law:

> *"It all adds up to normality"* (Egan 1992, last page)

However we come to view the world – it is as it always has been. It is the world in which you grew up in as a child. It is the world which has given you happiness, joy, grief; friendship and love. There is nothing to fear. From the beginning, you have been at home.

And whatever else you take with you after reading this thesis, I hope to have inspired you to this attitude:

> This doctrine contributes to the welfare of our social existence, since it teaches us to hate no one, to despise no one, to mock no one, to be angry with no one, and to envy no one. (Spinoza 1677)

Appendix A: Terminology of Mario Bunge

It will be helpful to introduce some terminology from the philosophical system of Mario Bunge. Overviews can be found in Mahner & Bunge (1997); Mahner & Bunge (2000); Bunge & Mahner (2004); Bunge (2006). The portions reproduced here have been excerpted from Mahner & Bunge (1997). The system of Mario Bunge is to be seen as a starting point, not a dogmatic position. It provides a clear basis for venturing into more difficult philosophical territory. The postulates, definitions, theorems and corollaries are presented in the order as they appear in cited book. Some comments are interleaved, identifiable by being in standard font. I have only excerpted those passages which are of immediate relevance to the present work, but have kept the numbering of the cited book in place; apart from direct quotes, some parts are paraphrased.

> Postulate 1.1. The world (or universe) exists on its own (i.e., whether or not there are inquirers).
>
> Postulate 1.2. Every object is either a thing or a construct, i.e., no object is neither, and none is both.

Postulate 1.2 asserts methodological dualism, not ontological dualism; constructs have no independent existence from material processes, but it is well to distinguish them conceptually. See also immediately below.

> Postulate 1.3. The world is composed exclusively of things (i.e., concrete or material objects).
>
> Definition 1.1. Let x represent a bare substantial individual and call P(x) the collection of all the (known and unknown) properties of x. Then the individual together with its properties is called the *thing* (or *concrete* or *material* or *real object* or *entity*) X; i.e., $X =_{df} \langle x. P(x) \rangle$.
>
> Definition 1.2. The *scope S* of a property is the collection of entities possessing it.

Scope is an ontological concept defined on properties, not to be confused with the semantical concept of *extension* defined on predicates.

> Definition 1.3. If P and Q are (essential) properties of things, then P and Q are said to be lawfully related if, and only if, $S(P) \subset S(Q) \text{ or } S(Q) \subset S(P)$.
>
> Postulate 1.4. Every essential property is lawfully related to some other essential property. That is, for any two essential properties, P and Q, either $S(P) \subset S(Q) \text{ or } S(Q) \subset S(P)$.

Bunge calls this the *ontological principle of lawfulness*; where laws are constant relations between properties, and not propositions about these relations. Bunge takes *essential* properties to be those without which a thing would be a different thing.

> Postulate 1.5. Every thing changes.
>
> Theorem 1.1. Every thing can undergo only lawful changes (i.e., events or transformations).

Note that Bunge does not restrict lawful behavior to strict causality – stochastic lawfulness suffices. Conceptualized somewhat abstractly, things can change in the nomological state space spanned up by the axes representing properties they possess. The theorem is inferred from postulate 1.4.

> Corollary 1.1. There is no total disorder, and there are no miracles.
>
> Definition 1.4. A complex event, i.e., one formed by the composition of two or more events, is called a process.

An event is an ordered pair of states.

> Definition 1.5. For any x: x is a *concrete* (or *material*, or *real*) thing (or entity) $=_{df}$ x is changeable.
>
> Definition 1.6. For all x: x is an *ideal* (or *abstract* or *conceptual*) object (or *construct*) $=_{df}$ x is neither unchanging nor changeable.

Bunge is clear to stress that this definition does not imply the autonomous existence of abstractions – it is, again, just a conceptual classification.

> Postulate 1.6. Every concrete thing is either a system or a component of one.
>
> Postulate 1.7. Every system, except the universe, is a subsystem of some other system.
>
> Postulate 1.8. The universe is a system, namely the system such that every other thing is a component of it.
>
> Definition 1.7. A relation between a thing x and a thing y is a *bonding* relation if, and only if, the states of y alter when the relation to x holds.

Bunge describes a method of analyzing systems, the so called CES Analysis, where C stands for composition, E for environment and S for structure:

- The collection of all parts of a system s – its composition – is designated by C(s).
- The environment of the system s is called E(s). The environment is always relative to the system. Of interest is the proximate environment, not the whole universe.
- A system has an internal structure, that is, relations holding among the elements of C(s) – the endostructure – and also possesses external relations that elements of C(s) undergo with elements of E(s). The union of the two is called the structure S(s). The relations can be bonding and non-bonding relations, where bonding relations are usually more interesting.
- One can then qualitatively define a system *m* by the following ordered triple:

  m(s) = <C(s), E(s), S(s)>

m is called the CES model of a system; with this definition, one can also easily define subsystems of systems etc.

> Definition 1.9. Let P represent a property of a thing b. P is an *emergent* property of b if, and only if, either
>
> (i) b is a complex thing (a system), no component of which possesses P; or
>
> (ii) b is a thing that has acquired P by virtue of becoming a component of a system (i.e., b would not possess P if it were an independent or isolated thing).

This conception of emergence is compatible with the one endorsed in this thesis, that is, weak emergence coming about through manifestations of mutual dispositions. The dual notion of *emergence* is *submergence*.

> Postulate 1.9. All processes of development and evolution are accompanied by the emergence or the submergence of (generic) properties.

Now on to the important distinction between fact and phenomenon. A *fact* is the *being of a thing in a given state* or *an event occurring in a thing*. Neither constructs, theories, data or propositions are facts. Phenomena are perceptions, occurring in neural systems, and as such are also facts; facts

designating a subject-object relation. The same fact may appear differently to different perceiving agents, and thus one fact may appear in many subject-object fact relations, giving rise to different phenomena in different agents.

> Postulate 1.10. Let Φ designate the totality of possible facts occurring in an animal b and its (immediate) environment during the lifetime of b, and call Ψ the totality of possible percepts of b (or the phenomenal world of b) during the same period. The Ψ is properly included in Φ, i.e. $\Psi \subset \Phi$.

Causation is defined by Bunge as *event generation by energy transfer from one entity to another*, energy being the single property which all material things possess (this conception would need some clarification in regard to the ontology of the present thesis; space does not permit this, but nothing fundamental is at stake).

He distinguishes strong energy transfer from weak energy transfer (signaling), which corresponds roughly to structural causes versus triggering causes. While Bunge construes the scientific world view as deterministic, he does not require strict causation for determinism to hold. Determinism only requires lawful behavior – which can also include probabilistic laws – but excludes the occurrence of miracles and posits the principle of *ex nihilo nihil fit*.

Can reasons be causes? Reasons are constructs, but the process of reasoning is a material brain process, which can be a cause.

Now on to philosophical semantics.

Intension: the intension, or sense, is practically prior to the extension, as you need a predicate to build the extension. The proposition "b is an F" corresponds to 'Fb' where F is a function. The sense of a proposition is the set of all propositions it entails or is entailed by.

Extension: predicates define extensions over the properties they designate. Example: the predicate M defines the following extension: E(M) = { x e S | Mx} where S is an appropriate domain. It is possible that predicates have empty extensions, such as the predicate G: "is a ghost": E(G) = {}

Reference class: from the extension one must distinguish the reference class; the reference class is the domain over which the x or b range. The reference class is *conceptual*, whereas the extension refers to actually *existing material entities*; to build the extension, one needs the concept of truth,

because only those elements for which M holds can be part of the extension. On the other hand, the reference class may consist of purely theoretical or even fantastic entities – like in the case of ghosts – the reference class is open to all kinds of possible entities being able to pass for a ghost. Another difference between *reference classes* and *extensions* is that in n-ary predicates, where n>1, the extension is a set of ordered n-tuples, while the reference class continues to be composed of individuals; finally, the reference class does not change when applying negation, disjunction etc.

The meaning of a *construct c* is the sense together with its reference.

> Definition 3.3. The knowledge of an animal at a given time is the set of all items it has learned and retained up until that time.

where

> Postulate 3.3. Learning is the specific function of some plastic neuronal system.
>
> Definition 3.6. An animal b has acquired some (partially true) perceptual knowledge of some items in its environment E if, and only if, b possesses a plastic neuronal system n such that some events in E are mapped into events in n.
>
> Postulate 3.5. Any knowledge of factual items is not direct or pictorial but symbolic.

From this follows that knowledge is not an abstract platonic ideal, but always some process in a cognitive system. This completely naturalistic conception of knowledge as consisting of certain neural processes makes it independent of either objectivity or truth. Also, books etc do not contain knowledge per se, but encode the knowledge of one cognitive entity so that other cognitive entities may reconstruct that meaning with their own neural processes.

> Successful communication consists in the construction or (re)creation of similar processes in the brains of the animals involved in the interaction.

Concerning knowledge, the following distinction is useful:

> ...if subject s knows p, then (a) s has *explicit* knowledge of p iff s also knows that s knows p or knows how to express p in some language; (b) otherwise s has *tacit* knowledge of p.

Public or intersubjective knowledge is shared by more than one agent. But intersubjective is not equivalent with objective:

Definition 3.4. Let p designate a piece of knowledge. Then p is *objective* if, and only if,

(i) p is public (intersubjective) in some society, and

(ii) p is testable either conceptually or empirically.

Bunge goes on to define *belief*:

Definition 3.5. Let s denote a subject and p a piece of knowledge. Then

(i) s believes p $=_{df}$ s knows p and gives assent to p

(ii) s is justified in believing p $=_{df}$ s knows p and s knows that p is reasonably well confirmed;

(iii) s is justified in disbelieving p $=_{df}$ s knows p and s knows that p has been disconfirmed.

Now on to *truth*. While for the formal sciences consistency is enough, for the factual sciences this is not so. When you state propositions, if they are to be true, they need to stand in some relation to the real world; it calls for a correspondence theory:

Consider a thing ϑ internal or external to an animal a endowed with a brain capable of learning. Call e an event occurring in thing ϑ, and e* the corresponding perceptual or conceptual representation of e in the brain of a. Then we say that a has gained *true partial knowledge* of fact e if, and only if, e* is identical to the perception or conception of e as a change in thing ϑ (rather than as a nonchange or as a change in some other thing). The true (though partial) knowledge that a has acquired of event e is the neural event e*; and the *correspondence* involved is the relation between the events e and e*.

We have so far been talking about thoughts, i.e., concrete events, not constructs, which we have defined as equivalence classes of thoughts. To arrive at propositions, we form the equivalence class of thoughts e* constituting true (though usually partial) knowledge of e: [e*]. Note that no two members of the class [e*] are likely to be identical, for they are thoughts of a given animal at different times, or thoughts of different animals, and in either case they differ in some respect or other. However, they are all equivalent in that every one of them constitutes true partial knowledge of e; that is, for every member e* of [e*], e* happens if, and only if, e is (or has been or will be) the case. We identify the proposition p = "e is the case"

> with that equivalence class of thoughts, i.e., we set p = [e*]. And we stipulate that p is true if, and only if, e happens or has happened. Thus, the correspondence relation holding between a mental fact and some other (mental or nonmental) fact carries over to propositions in relation to facts. Accordingly, truth and falsity are *primarily* properties of perceptions and conceptions (e.g. propositional thoughts), and only *secondarily* (or derivatively) attributes of those equivalence classes of thoughts we call 'propositions'. (p. 130)

Now, how to determine if our [e*] are in fact in correspondence with reality? That is where truth indicators enter the stage, such as:

- empirical confirmation or disconfirmation, that is, empirical adequacy;
- conceptual indicators such as theory-internal and external (in relation to other knowledge) consistency;
- unifying power and
- predictive power;
- heuristic power (in the sense of furthering scientific advancement);
- stability (not toppled by every new incoming datum) and
- depth (of theoretical entities and mechanisms);
- simplicity.

When all these tests are applied, propositions and theories can be assigned a qualitative degree of confirmation in the sense of "very strong, strong, indecisive, weak or very weak".

> Postulate 3.4. We can get to know the world, although only partially, imperfectly (or approximately), and gradually.

This is the postulate of epistemological (critical) realism which together with ontological realism is the cornerstone of scientific realism. Epistemological constructivism is counterbalanced by epistemological naturalism and evolutionism, which defeats radical constructivism. The position is both fallibilist and meliorist. Finally, scientism is advocated, the thesis that

> anything knowable and worth knowing can be known scientifically, and that science provides the best possible factual knowledge, even though it may, and does, in fact, contain errors. (p 135)

Appendix B: On the Origin of Objects

Below I have excerpted important passages from (Smith 1996). They are intended as appetizers to reading the whole book. All page numbers refer to (Smith 1996).

Particularity:

> By particularity, first, I mean simply that our everyday notion of an object – the base case, on which any more abstract versions rest – is of a located patch of metaphysical flux. By 'particular,' that is, I mean something like 'occurrent': something that is located or that happens, something that is embodied, something for which there is a Steinian "there there." (p. 117)

Individuality:

> By individuality, on the other hand, I mean whatever it is about an entity that supports the notion of individuation criteria – something that makes 'object' a count noun, something that makes objects discrete. Somehow or other, an individual object is taken to be something of coherent unity, separated out from a background, in the familiar "figure-ground fashion. (p. 119)

> to separate the sense of the very specific or local or 'peculiar," to be associated with particularity, from the quite different sense of being discrete or chopped up into distinct units or wholes, to be associated with individuality. Thus the toys strewn around on the lawn outside, as I write, have both properties: they are particular individuals; whereas the water lapping on the rocks far below them remains particular - even exquisitely so-without thereby requiring any such "division" into discrete individual parts. (p. 120)

The deviation from traditional ontological bare particulars is illustrated on p. 123 in Smith (1996).

Nature and Naturalism:

> From here forward, that is, 'nature' will be taken to be another name for the unnameable world, 'natural' to mean part of this unnameable world, and 'naturalism,' to shift back to its epistemic or metatheoretic sense, to be our mandate, as theorists, to show how everything is part of this world-i.e., to show how the world is One, in chapter 3's sense of being entire or complete. In sum:

Naturalism is realism's methodological correlate.

Note that this construal does not give the physical any logical or metaphysical pride of place. Nor does it imply the existence of, or probability of, or even the commitment to look for, anything that would warrant being called a "unified theory of science." To think that would be to commit an error of scope. Naturalism as I understand it, that is, is not so much a desire for an integrated understanding of nature, as a desire for *an understanding of an integral nature*. To see how the world is One, however that is accomplished. (p. 140)

Criterion of Ultimate Concreteness:

No naturalistically palatable theory of intentionality - of mind, computation, semantics, ontology, objectivity - can presume the identity or existence of any individual object whatsoever.

The name "Criterion of Ultimate Concreteness" is appropriate because, as will soon be evident, one of the things that individuals have, that physical phenomena lack, is concreteness's opposite: abstraction. (p. 184)

Given … the commitment to honor the Criterion of Ultimate Concreteness, that translates into the following more specific goal: to understand how a conception of objects can arise on a substrate of infinitely extensive fields of particularity. (p 191)

Except of course that this is an untenable way to phrase it. To say "a conception of objects" makes it sound as if the achievement is the subject's, by assuming a split between conception and what is conceived of. It also fails by making it sound as if the achievement is cognitive. Nor is anything gained by striking a more traditionally realist stance, and asking "how objects can arise on a substrate of infinitely extensive fields." That puts the achievement too squarely on the object. Both ways of putting it violate the mandate of avoiding an a priori subject-world split. In place of these dichotomous formulations, therefore, I will speak, unitarily, of registering the world. (p 191)

Register:

> By 'register' I mean something like *parse, make sense of as, find there to be, structure, take as being a certain way- even carve the world into*, to use a familiar if outmoded phrase. (p. 191)

Flex and Slop:

> The world is fundamentally characterized by an underlying flex or slop - a kind of slack or "play" that allows some bits to move about or adjust without much influencing, and without being much influenced by, other bits. (p. 199)

> Even if we were to discover that every macroscopically observable regularity was the product of such amplified long-distance effects of microdisturbances, it would still be true that far and away the majority of microdisturbances quickly die away. In our world, especially to the extent that we find it coherent, effects by and large dissipate. Think 1/r^2. (p. 200)

Smith sets out to put *intentionality* on a good footing:

> Overall, my aim in this book is to show that the world's primordial flex or play does two crucial things: (i) establishes the problem that intentionality solves; and (ii) provides the wherewithal for its solution. (p. 200)

> Because of the dissipative nature of the playing field, an enduring entity cannot, at any given moment, be affected by things that, at that same moment, are beyond what I will call *effective reach*. Effective reach is not a yes/no affair; rather, this is essentially the gradual falling off or dissipation of influence familiar from physics. (p. 201)

> In all these situations, what starts out as effectively coupled is gradually pulled apart, but separated in a such a way as to honor non-effective long-distance coordination condition, leading eventually to effective reconnection or reconciliation. There is a great deal more to intentionality than that, and a great deal to say about what constitutes coupling, coordination, and so on, but in various forms these notions of connection, gradual disconnection, maintenance of coordination while disconnected or separated, and ultimate

reconnection or reconciliation permeate all kinds of more sophisticated example. There is nothing more basic to intentionality than this pattern of coming together and coming apart, at one moment being fully engaged, at another point being separated, but separated - this is the point - in such a way as to stay coordinated with what at that moment is distal and beyond effective reach. (p. 206)

Intentionality is a way of exploiting local freedom or slop in order to establish coordination with what is beyond effective reach (p. 208)

The underlying spatio-temporal extended fields of particularity throw tufts of effective activity up against each other, and let them fall apart, fuse them and splinter them and push them through each other, and generally bash them around, in ways governed by the pervasive underlying (physical) laws of deictic coupling. For a subject to begin to register an object as an individual is, first, for a region of the fields (the s-region) not to be connected to another region (the o-region), but in the appropriate way to *let go* of it. Not in the sense of dropping connection forever, but in a way that maintains an overall pattern of coordination-a pattern that in all likelihood will allow it, among other things, to come back into connection with it again, at other times and places, and perhaps in other ways. The coordination requires establishing appropriately stable (extended in the s-region) and abstract (extended in the o-region) focus on the o-region, while remaining separate. The separation helps in maintaining a somewhat abstract focus on the o-region, by insulating the s-region from being buffeted by every nuance and vibration suffered by the o-region. (p. 241)

Commensurability:

As should be clear from even as much of the metaphysical picture as has been suggested so far, a*ll of human communication lies in this middle region*, between identical and incommensurable registration schemes. Not only that, when we get back to the proper metaphysical story, as was promised earlier, it will emerge that, far from being undermined by it, registration is designed to cope with this middle region of partial commensurability. (p. 255)

In sum, it is important that shifting registers is hard. But it is not hard so much in the sense of there being metaphysical limits. What is important, rather, is that communicating across registers is *hard work*. This is what will make the metaphysical picture substantive. It is also a conclusion that underwrites a claim made in chapter 3: that developing a more adequate metaphysics is not only intellectually viable, but also politically urgent. The aim is to give metaphysical grounding to, and support for, communicative and political struggles among people whose experience of, and participation in, our world is different. (p. 255)

Connection and Encounter:

So the story must be broadened and made more symmetrical: to include patterns of connection and encounter as well as coordinated patterns of disconnection and separation. (p. 292)

Connection or encounter is important, first, because this is where things happen; this is even what it is to happen-the locus of all struggles and trials and engagements and meetings, the pure and unvarnished bumping and shoving of the world. This is the realm of the effective, of which so much of computer science (or so at least I claim) is a nascent theory. This is what has to be *implemented* if you want to build something that plugs into the wall and gets something done. It is what was lacking in the properties Searle attributed to the wall in his office. And it is the locus of what is right about physicalism (though physicalism gets its own intuition wrong by trying to formulate it ontologically). The connected is the realm of force, struggle, energy, encounter. Connection or encounter is how the whole thing works. (p. 292)

The virtues of Connection and Disconnection:

If abstraction is virtuous withdrawal we equally need virtuous re-engagement. (p. 311)

Formality is discreteness run amok:

Formality is discreteness run amok. On the picture being painted here, in contrast, the world is not presumptively discrete - indeed,

it is as completely opposite of formal as it is possible to imagine. It is instead permeated by:

1. Indefiniteness *at the edges of given objects*, such as the boundaries of the region on the wall where I ask you to write your name (there being no metaphysical need for determinate edges);

2. Indefiniteness *between and among objects of the same type*, such as whether you are standing on this sand dune or the neighboring one; or whether the massif above our campsite consists of three mountains or four;

3. Indefiniteness *among different types*, such as among chutzpah, bravado, ego, self-confidence, and brashness;

4. Indefiniteness among the notions *'concept,' 'type,' and 'property'* - as for example in debates between philosophers and psychologists on the nature of concepts: about whether they are mental or abstract, and about what it is that people can and cannot share (do we share a concept of red? do we each have private concepts that represent the same abstract property? or do we all have different concepts?);

5. Indefiniteness *between objects and the types they exemplify*, implying that the "instance-of" relation is itself approximate, contested, and potentially unstable - as for example in whether the headache you have this morning is the same one you had last night, or a different one of the same type; and similarly for patches of color, fog, and "the rain"; and

6. Indefiniteness *between and among different realms of human endeavor*, such as the political, the social, the technical, the religious, the esthetic, the psychological, etc.

The ubiquity of this gradualism shows once again why it was so important to avoid making sharp theoretic distinctions in advance. This was especially true in the case of the classical dualisms: between subject and world, mind and body, abstract and concrete, nature and society. I initially motivated avoiding these binarisms for two reasons: in order to avoid making inscription errors, and in order

to keep theoretician's and subject's ontology (registration) distinct. Third, I avoided them because of their expense. (p. 324f)

The relationship of ontology, registration, world, representation and subject:

Ontology is the projection of registration onto the world. Representation is the projection of registration onto the subject or vehicle. (p. 349)

Remarks on Mathematics:

Compared to the partial merger of representation and ontology, and the attempt to steer a tenable course between constructivism and realism, the third consequence of the metaphysical lack of categorical decisiveness may seem less important. But it is still dramatic enough. Mathematics will need to be overhauled.

To see why, note first how present-day (i.e., modernist) mathematics orders its explanations:

I. Discreteness is assumed to be primitive and absolute, exemplified for example by sets, natural numbers, and many other

such properties (being even, being irrational, etc.);

2. Continuity is then defined in terms of discreteness, with the usual apparatus of Dedekind cuts, convergent Cauchy sequences, and the like; and

3. Finally, if at all, vagueness, or at least a little bit of vagueness, is modeled (as for example in the current fashion for fuzzy logic[4]).

This is the world view captured in Kronecker's famous dictum: that "God made the integers; all else is the work of man."[5] If my metaphysical picture is right, Kronecker's order of explanation is *close to backwards*. Metaphysical indefiniteness is the base case, continuity needs to be extruded from the flux, and then discreteness won, at a very high price, from that.

That is to put it metaphysically. It may be more revealing to approach it epistemically, however - or at least we should look at it

from that angle as well, given that the two are never wholly separable. At a minimum, the proposal will require exhuming mathematical practice, as recommended for example by intuitionists, and explicating this *achievement* of mathematical results by giving mathematicians partial ontological as well as partial epistemic responsibility for their acts - some ontological responsibility, rather than none; some epistemic responsibility, rather than all. This is not to give unrestricted license to idealism or formalism, because of realism's second constraint: mathematicians themselves, the very ones to which this metaphysical respect is to be granted, must be recognized as part of the same reality as the numbers they extrude.

This recognition that mathematicians are as much part of the world as the numbers they study puts the lie to the sharpness, and perhaps even to the coherence, of the distinctions among three allegedly alternative ways in which mathematics is traditionally understood:

1. *Empirical:* true of the physical or material world, even if at a

relatively abstract level or high order;

2. *Platonist*: true of an independent mathematical realm; and

3. *Intuitionist*: characteristic of our native mental or cognitive capacities.

At a minimum, on the present metaphysics, the first and third positions, empirical and intuitionist, begin to merge. For suppose that the empirical view is right: that (what we come to register as) mathematical properties are high-level abstractions of ordinary material situations. Suppose, that is, that "threeness" is first and foremost a property of those worldly states of affairs that we register as consisting of three individuals. This essentially empirical view is compatible with the intuitionist's claim, to put it into current language, that our ability to register situations as exemplifying threeness depends, inexorably, on architectural facts about our native registrational capacities. (p. 354f)

The Middle Ground:

...that material objects - and the material world more generally - occupy what we might call a *middle ground* [Footnote 12], halfway between (the predecessor era's notion of) the physical world, and (the predecessor era's notion of) the intentional world. The resulting *median nature of materiality* has numerous theoretical consequences, only a few of which have even been touched on here. For example, it undergirds the fact that objects themselves, not just their representations, are culturally, historically, and socially plural - and yet not just products of the imagination or intentional whim of a person, society, or community, either, but made of the stuff of the world, as resistant and wily and obstreperous as the rest of us. (p. 363)

But the proposal is not to get rid of syntax, and to leave the two realms unconnected. On the contrary, I am arguing that *all of ontology* lives in the intermediate realm. You can see this in the moves I have made. On the one hand, I have "lifted" material objects up from the bottom, claiming that they depend inherently on intentional (registrational) practices of subjects. At the same time I have driven semantics and content down, claiming that thought is intrinsically material, giving priority to non-conceptual content, arguing that connected practices are a constitutive part of intentionality, and the like. It is the thick participatory mix to which I have given the label "middle ground." (p. 365)

One World:

For the account has supplied what would otherwise have been impossible: *ontological pluralism sustained by metaphysical monism.*

There is only one world - that is what was important about realism. But its unity transcends all ability to speak. (p. 375)

Appendix C: The Self

Here I will sketch a very bare-bones variant of Metzinger's model. All quotes are from (Metzinger 2003). An intriguing account for the illusion of the self from a Buddhist perspective is given by Albahari (2006).

First of all we will need the concept of mental representation:

Mental Representation: Rep_M (S, X, Y)

- S is an individual information-processing system.
- Y is an aspect of the current state of the world.
- X represents Y for S.
- X is a functionally internal system state.
- The intentional content of X can become available for introspective attention. It possesses the potential of itself becoming the representandum of subsymbolic higher-order representational processes.
- The intentional content of X can become available for cognitive reference. It can in turn become the representandum of symbolic higher-order representational processes.
- The intentional content of X can become globally available for the selective control of action. (p. 21)

Important is Metzinger's concept of transparency of phenomenological properties:

> Transparency is a special form of darkness. With regard to the phenomenology of visual experience transparency means that we are not able to see something, because it is transparent. We don't see the window, but only the bird flying by. *Phenomenal* transparency in general, however, means that something particular is not accessible to subjective experience, namely, the representational character of the contents of conscious experience. This analysis refers to all sensory modalities and to our integrated phenomenal model of the world as a whole in particular. The *instruments* of representation themselves cannot be represented as such anymore, and hence the experiencing system, by necessity, is entangled in a naive realism. (p. 169)

Metzinger then introduces a phenomenal self-model (PSM):

> The content of the PSM is the content of the conscious self: your current bodily sensations, your present emotional situation, plus all the contents of your phenomenally experienced cognitive processing. They are constituents of your PSM. Intuitively, and in a certain metaphorical sense, one could even say that you *are* the content of your PSM. All those properties of yourself, to which you can now direct your attention, form the content of your current PSM. Your self-directed thoughts operate on the current contents of your PSM: they *cannot* operate on anything else. When you form thoughts about your "unconscious self" (i.e., the contents of your *mental* self-model), these thoughts are always about a *conscious* representation of this "unconscious self," one that has just been integrated into your currently active PSM. If you want to initiate a goal-directed action aimed at some aspect of yourself—for example, brushing your hair or shaving yourself—you need a conscious self-model to *deliberately* initiate these actions. (p. 299)

The last theoretical entity we need is a phenomenal model of the intentionality relation (PMIR):

> What is the phenomenal model of the intentionality relation? It is a conscious mental model, and its content is an ongoing, episodic *subject-object relation*. Here are some examples, in terms of typical phenomenological descriptions of the class of phenomenal states at issue: "I am someone, who is currently visually attending to the color of the book in my hands"; "I am someone currently grasping the content of the sentence I am reading"; "I am someone currently hearing the sound of the refrigerator behind me"; "I am someone now deciding to get up and get some more juice." (p. 411)

And:

> The overall picture that emerges is that of the human self-model continuously integrating the mechanisms of attentional, cognitive, and volitional availability against a stable background formed by the transparent representation of the bodily self.
>
> Please note how the PMIR has a phenomenally experienced direction: PMIRs are like arrows pointing from self-model to object component.(p. 413)

With the above in place, the paragraph below becomes decipherable and gives us a first scientific approximation of how the illusion of a "substantial" self comes about:

> Subjectivity, in the theoretically interesting sense of being bound to an individual, consciously experienced first-person perspective, is something that can only be conceptually analyzed and turned into an *empirically* tractable feature of consciousness by introducing the two new theoretical entities I presented in this chapter, namely, the transparent PSM and the transparent PMIR. We can now see how 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. Call this the "self-model theory of subjectivity" (p. 427)

Appendix D: A Question of King Milinda

Now Milinda the king went up to where the venerable Nâgasena was, and addressed him with the greetings and compliments of friendship and courtesy, and took his seat respectfully apart. And Nâgasena reciprocated his courtesy, so that the heart of the king was propitiated.

And Milinda began by asking, [Footnote 1] 'How is your Reverence known, and what, Sir, is your name?'

'I am known as Nâgasena, O king, and it is by that name that my brethren in the faith address me. But although parents, O king, give such a name as Nâgasena, or Sûrasena, or Vîrasena, or Sîhasena, yet this, Sire,- Nâgasena and so on – is only a generally understood term, a designation in common use. For there is no permanent individuality (no soul) involved in the matter. [Footnote 2].'

Then Milinda called upon the Yonakas and the brethren to witness: 'This Nâgasena says there is no permanent individuality (no soul) implied in his name. Is it now even possible to approve him in that?' And turning to Nâgasena, he said: 'If, most reverend Nâgasena, there be no permanent individuality (no soul) involved in the matter, who is it, pray, who gives to you members of the Order your robes and food and lodging and necessaries for the sick? Who is it who enjoys such things when given? Who is it who lives a life of righteousness? Who is it who devotes himself to meditation? Who is it who attains to the goal of the Excellent Way, to the Nirvâna of Arahatship? And who is it who destroys living creatures? who is it who takes what is not his own? who is it who lives an evil life of worldly lusts, who speaks lies, who drinks strong drink, who (in a word) commits any one of the five sins which work out their bitter fruit even in this life? If that be so there is neither merit nor demerit; there is neither doer nor causer of good or evil deeds; there is neither fruit nor result of good or evil Karma - If, most reverend Nâgasena, we are to think that were a man to kill you there would be no murder, then it follows that there are no real masters or teachers in your Order, and that your ordinations are void.- You tell me that your brethren in the Order are in the habit of addressing you as Nâgasena. Now what is that Nâgasena? Do you mean to say that the hair is Nâgasena?'

'I don't say that, great king.'

'Or the hairs on the body, perhaps?'

'Certainly not.'

'Or is it the nails, the teeth, the skin, the flesh, the nerves, the bones, the marrow, the kidneys, the heart, the liver, the abdomen, the spleen, the lungs, the larger intestines, the lower intestines, the stomach, the faeces, the bile, the phlegm, the pus, the blood, the sweat, the fat, the tears, the serum, the saliva, the mucus, the oil that lubricates the joints, the urine, or the brain, or any or all of these, that is Nâgasena?'

And to each of these he answered no.

'Is it the outward form then (Rûpa) that is Nâgasena, or the sensations (Vedanâ), or the ideas (Saññâ), or the confections (the constituent elements of character, Samkhârâ), or the consciousness (Vigñâna), that is Nâgasena?'

And to each of these also he answered no.

'Then is it all these Skandhas combined that are Nâgasena?'

'No! great king.'

'But is there anything outside the five Skandhas that is Nâgasena?'

And still he answered no.

'Then thus, ask as I may, I can discover no Nâgasena. Nâgasena is a mere empty sound. Who then is the Nâgasena that we see before us? It is a falsehood that your reverence has spoken, an untruth!'

And the venerable Nâgasena said to Milinda the king: 'You, Sire, have been brought up in great luxury, as beseems your noble birth. If you were to walk this dry weather on the hot and sandy ground, trampling under foot the gritty, gravelly grains of the hard sand, your feet would hurt you. And as your body would be in pain, your mind would be disturbed, and you would experience a sense of bodily suffering. How then did you come, on foot, or in a chariot?'

'I did not come, Sir, on foot. I came in a carriage.'

'Then if you came, Sire, in a carriage, explain to me what that is. Is it the pole that is the chariot?'

'I did not say that.'

'Is it the axle that is the chariot?'

'Certainly not.'

'Is it the wheels, or the framework, or the ropes, or the yoke, or the spokes of the wheels, or the goad, that are the chariot?'

And to all these he still answered no.

'Then is it all these parts of it that are the chariot?'

'No, Sir.'

'But is there anything outside them that is the chariot?'

And still he answered no.

'Then thus, ask as I may, I can discover no chariot. Chariot is a mere empty sound. What then is the chariot you say you came in? It is a falsehood that your Majesty has spoken, an untruth! There is no such thing as a chariot! You are king over all India, a mighty monarch. Of whom then are you afraid that you speak untruth? And he called upon the Yonakas and the brethren to witness, saying: 'Milinda the king here has said that he came by carriage. But when asked in that case to explain what the carriage was, he is unable to establish what he averred. Is it, forsooth, possible to approve him in that?'

When he had thus spoken the five hundred Yonakas shouted their applause, and said to the king: Now let your Majesty get out of that if you can?'

And Milinda the king replied to Nâgasena, and said: 'I have spoken no untruth, reverend Sir. It is on account of its having all these things - the pole, and the axle, the wheels, and the framework, the ropes, the yoke, the spokes, and the goad - that it comes under the

generally understood term, the designation in common use, of "chariot."'

'Very good! Your Majesty has rightly grasped the meaning of "chariot." And just even so it is on account of all those things you questioned me about - the thirty-two kinds of organic matter in a human body, and the five constituent elements of being - that I come under the generally understood term, the designation in common use, of "Nâgasena." For it was said, Sire, by our Sister Vagirâ in the presence of the Blessed One:

'"Just as it is by the condition precedent of the co-existence of its various parts that the word 'chariot' is used, just so is it that when the Skandhas are there we talk of a 'being.'"'

'Most wonderful, Nâgasena, and most strange. Well has the puzzle put to you, most difficult though it was, been solved. Were the Buddha himself here he would approve your answer. Well done, well done, Nâgasena!' (Davids 1890)

Appendix E: Maudlin's Olympia

Here I will take a look at Maudlin's Olympia (Maudlin 1989) which will bring additional force into the argument for the incompatibility of materialism and computationalism, especially for those who still adhere to some levelist conception of reality. A rough sketch of Maudlin's paper is in order; as the point is important but not well known.

Central in the following discussion is the concept of *supervenience of consciousness on physical and computational states*. Maudlin assumes that the computational supervenes on the physical, and consciousness – mind – supervenes on the computational. But he derives a contradiction by showing that supervenience of mental states on the physical depends on physical *activity*; while supervenience of mental states on the computational depends on *counterfactual structure alone*, with no additional physical activity necessary. Both concepts of supervenience can't be held at the same time.

What is supervenience exactly?

> A set of properties *A* supervenes upon another set *B* just in case no two things can differ with respect to *A*-properties without also differing with respect to their *B*-properties. In slogan form, "there cannot be an *A*-difference without a *B*-difference".(McLaughlin & Bennett 2005)

A simple example: imagine the predicate *even/odd*; it supervenes on the natural numbers.

1 = odd; 2 = even; 3 = odd; 4 = even

You can't have an A-difference – changing *even* to *odd* – without also changing the underlying number. In regard to mental states, it works this way: once a physical state is fixed, the mental content (mind state) – your thoughts – are fixed. To get at another mind state, you need to change the physical state. The mind state is determined by a certain physical state. Physical supervenience means that mind supervenes on this *current and actual physical activity*.

Now on to computationalism. We will – actually, *must*, to not make the claim vacuous and arrive at pancomputationalism – assume that the kind of computationalism advocated here is operating on a higher abstraction level than naked, finest-grained reality[316]. We will make the assumption that we can abstract away from material properties and *still* get cognition. We are

316 In this sense, the following argument can also be seen as deriving a straight contradiction form a *levelist* conception of reality.

entering the levelist conception of reality, the Picture Theory. Maudlin is quite explicit with what computationalism actually means:

> the computational structure of the brain is what bestows mental properties. We must abstract away the particular physical, biochemical, and neural features of brains (Maudlin 1989)

Maudlin also differentiates between *functionalism* and *computationalism,* in a somewhat different way than we have above. Computationalism is directly defined via Universal Turing Machines (UTM). UTMs consist of machine tables with transition rules that subjunctively govern input/output relations and are capable of having internal states. Maudlin requires that the computation be *nontrivial,* that is, it really is a computation – giving different outputs for at least some different inputs. *Functionalism*, as described by Maudlin, would be more in line with standard materialism as presented in this thesis, because he requires a strong physical connection for functional relations to hold. He brings the example of a valve lifter: for a valve lifter to be functional, it needs to posses certain material properties, such as hardness, for instance. I will call it *M-functionalism* (M for Maudlin) to avoid confusion with other kinds of functionalism.

Maudlin now seems to endorse the following hierarchy:

- *physicalism*, which says that ultimately material properties count; I equate this with materialism as used in this thesis.
- *M-functionalism,* which says that matter *and* function count; but different matter *may* support the same function[317].
- *Computationalism*, which asserts that for all practical purposes *only* abstract properties count, but these even in their counterfactual implications. The connection to physical systems is very weak: it suffices that the physical system is able to support states, machines tables and account for I/O.

The assertion of both *physicalism* and *computationalism* combined then is this:

- a Turing machine running a non trivial program is
  - *necessary* for consciousness, and
  - *sufficient* for consciousness and

317 This view is actually quite acceptable, if one rejects the Picture Theory and its corollary that linguistic predicates must have real-world counterparts.

- consciousness also supervenes on the physical.

The problem is that these three propositions are inconsistent[318]. Maudlin shows this with an ingenious machine, Olympia[319]. Olympia is a machine with *minimal physical activity* that still supports consciousness – a phenomenal state PHI. Maudlin is clear that he does not want to cheat by deploying tricks of philosophy: for instance, *state transitions must conform to physical activity*. Searle's wall implementing a word processor is ruled out – it employs physically meaningless (Goodmanesque) predicates (Goodman 1983; Searle 1992). Maudlin's refutation is also not related to either Block's Chinese People computer (Block 1980) or Searle's Chinese room argument[320].

Three conditions must be satisfied by the machine to guarantee this:

- a machine must run through the states from s1 to s1000 and
- read a tape structure – the input – and
- it must have counterfactual structure; that is, react differently under different inputs, otherwise the program would be trivial.

Imagine we already have machines which instantiate a certain program PI, and this program PI supports phenomenal consciousness PHI for every input; that is, after all, what computationalism asserts. Let us call a machine of this type *Klara*. They are very simple and mechanical machines. Now Maudlin proceeds to construct the machine mentioned above, Olympia. Olympia consists of water, troughs, and other simple mechanical contraptions; and the states of Olympia correspond to different configurations of these mechanical pieces. Olympia, at the start, always transits from state s1 to s1000 on a fixed input TAU. Not more. That is, for a different input TAU_DIFF, it would probably compute the wrong result, because it would again simply transit from state s1 to s1000.

Olympia needs counterfactual structure so as not to be computationally trivial. So Maudlin proceeds to attach other machines – the Klaras – to Olympia. In fact, one Klara is attached at every mechanical point where a state transition can occur. But they are set up in a way that they only spring into action *if the input of the tape on that state varies from TAU*. That is, if the tape contains TAU, then Olympia will *not* need the attached machinery, but simply run on its own predetermined course and make the state transitions from s1 to s1000. Only if the tape *differs* from TAU, will one

318
319 The name is inspired by one of the figures in E.T.A Hoffmann's "The Sandman".
320 Incidentally, why always the Chinese?

of the attached Klaras take over – namely, the one where the tape first varied – and this Klara will continue to perform the rest of the computation.

The situation is now as such: Olympia corresponds, computationally, to a single but constructionally overly complex Klara. Olympia is originally a very simple machine, which, when given input TAU will only run on its original parts, and, by assumption, this will lead to phenomenal state PHI; she only needs the more complex additional machinery in case of deviation from TAU. Olympia clearly instantiates PI: for every tape input, she will give the correct output; not matter that for all but one tape input she delegates activity to the Klaras inside her.

And here we get the inconsistency: we can add the counterfactual structure by adding material, which, for a *concrete run on tape TAU*, is *physically inert*. We already have all the physical activity we need by the running of the simple Olympia in the first place.

Where is the inconsistency located exactly? Computationalism says that a computation is sufficient for PHI: Olympia computes PI for every TAU, so has sufficient power to support PHI. The computationalist also requires that the counterfactual structure be present: the computational power is necessary for PHI. But supervenience on the physical requires that consciousness supervenes on physical *activity* and not on *causally inert matter*. Thus, in the case where Olympia runs on the tape TAU, consciousness supervenes already on the *simple* Olympia without need for the inert Klaras. These positions contradict each other: according the computationalism, Olympia *with* the inert Klaras would be conscious, but Olympia without the inert Klaras would not be; which contradicts the *physical* supervenience principle – a physical state which was unconscious before would be made conscious by adding material that does not contribute physically to the computation. That, then, is the difference between physical (*causally active*) and computational (*counterfactual, inactive*) supervenience space. The physical activity in both cases is exactly identical.

To be especially clear on this point: consciousness, in case of computationalism would *supervene on physically inert material*. A computationalist needs both sufficiency and necessity – otherwise it isn't computationalism any more. So, Maudlin concludes, something must be wrong either with physical supervenience; or, with computationalism. Giving up computationalism is certainly more conservative and in line with current scientific evidence; and given the ontology advocated here, mandatory[321].

321 Counterfactuals enter the physical picture in quantum mechanics. One appetizer:

> In interaction-free measurements [...], an object is found because it might have absorbed a photon, although actually it did not. This idea has been applied to "counterfactual computation" [..], a setup in which the outcome of

It gets worse. Maudlin adds further fine points, such as the " argument by subtraction" – the mechanisms that *would* activate the Klaras in case of a diverging tape input could be mechanically weakened by attrition – rust for instance; the activation of a relevant Klara would then fail, because the mechanical link would break. He also presents an "argument by addition" – adding mechanical blocks that hinder an activation of a relevant Klara – the blocks need not even touch the rest of the machinery. With these methods, the physical difference between a "normal" machine instantiating a program PI and the mechanical Olympia can be made *arbitrarily* small[322].

Barnes (1991) raises objections to Maudlin's argumentation. He thinks that the rigging of Olympia to a certain input tape state – the TAU – makes her too “introspective”. Consciousness only appears if there is *real* interaction with the environment; that would make Barnes an externalist.

Barnes first defines "*generally appropriate activity conditions*" (GAA) for computations, which coincide with the ones given by Maudlin (that is, a machine must have states, a transition table, and perform the correct computation for different inputs). Barnes now objects that while Olympia performs a GAA computation, she does not do so because of *active causes* – that is, the concrete tape input TAU. She is rigged to run on TAU solely on causes *internal* to her.

Barnes constructs a dreamer/bell system to drive home his point:

- We have a text in a book, a dreamer and a bell.[323]
- The dreamer dreams with open eyes, but he is asleep. He dreams of the text and verbalizes his dream. His eyes trace the text – to an onlooker, it would seem as if the dreamer were awake and reading from the book; but this is not the case.

---

a computation becomes known in spite of the fact that the computer did not run the algorithm (Vaidman 2008).

That is why the universe being computational at the fundamental level would not encounter similar objections as the ones presented above – but maybe different ones. I can not explore this here; a full discussion must wait for another day and paper. But what is at stake here is *computationalism as a theory of the mental;* not computationalism as a theory of the physical. That is why it is of no concern here.

322 Another problem which computationalists face is the problem of clock slowdown: Imagine a computer going through states s1 to s1000, and this is necessary and sufficient for the arising of phenomenal states. We can now do something perverse: we let the computer run s1, then make a backup, disassemble the computer, wait a thousand years, build a new computer, install the backup, run to s2 etc and finally reach s1000 after a billion years ; it would still support PHI if computation is *sufficient* for consciousness.

Computationalism is actually even weirder: a computationalist should have no problem if the *successive states are mixed in their order of appearance*; here, the platonic anti-materialist roots of computationalism begin to show visibly. An exploration of these ideas can be found in Egan (1994).

323 The setting does have a Buddhist flavour about it.

- There is a bell in the room – most probably magical[324] – that rings and wakes up the dreamer just in case he dreams of a word that is not in the text as his eyes are upon the relevant section. The dreamer wakes up because of the bell, and then carries on reading the text with his normal reading abilities.

The dreamer will always read the whole text; if he dreams "correctly" without any *active* causal connection to the text whatsoever; all the active causes are *internal* to the dreamer. But can we call the activity the dreamer performs *reading*? Barnes does not think so. Barnes claims that, to speak of cognition, the object which is the focus of cognition must play an active causal role leading the subject's thinking of the object. He labels this condition for cognitive activity as the "*causally active object condition*" (CAO).

He concludes that if a subject/object system meets GAA and not CAO, this does not suffice for cognition. But maybe GAA and CAO together do – he does not elaborate, but remarks that CAO probably has to be fulfilled in a *particular* way. Computation is, according to Barnes, "intrinsically a reciprocal, active causal interaction". Barnes concludes:

> Olympia, while she succeeds in performing a GAA computation, does not succeed in performing a GAA/CAO computation. Hence Maudlin's argument entails the following conclusion: the status of a system as conscious cannot supervene on a system simply in virtue of its GAA computational structure. There is nothing in his argument which counts against the possibility that a system is conscious simply in virtue of its GAA/CAO computational structure. For those of us who take 'GAA/CAO computation' as simply equivalent to the ordinary notion of "computation," this amounts to Maudlin's failure to refute a computationalist theory of consciousness.

Well; whatever the "correct" notion of computation is – we have left the abstract behind us and have entered the domain of the physical world and its metaphysical categorization. And rightly so, because that is where solutions to worldly problems should be sought. Taken this way, the above results can be seen as defeating a *purely abstract conception of cognition as computation* – performing the right computations is *not sufficient* to be conscious.

324 Barnes does not say.

The problem above, incidentally, is reminiscent of Swampman. Olympia performs a GAA computation, but not a GAA/CAO computation. The same holds for Swampman, computationally construed: he computes GAA-like, but not GAA/CAO-like.

In any case, a causal connection to the world seems to be required; better: the causal connection ensures that in truth there is no real separateness of world and cognitive entity, only one world which contains a cognitive subsystem. The computationalist may reach this conclusion via the requirement of CAO. The metaphysical monist may locate this connection in dispositionality.

Appendix F: The Proactionary Principle

More distinguishes nine principles underlying the proactionary principle:

1. Freedom to innovate: Our freedom to innovate technologically is valuable to humanity. The burden of proof therefore belongs to those who propose restrictive measures. All proposed measures should be closely scrutinized.

2. Objectivity: Use a decision process that is objective, structured, and explicit. Evaluate risks and generate forecasts according to available science, not emotionally shaped perceptions; use explicit forecasting processes; fully disclose the forecasting procedure; ensure that the information and decision procedures are objective; rigorously structure the inputs to the forecasting procedure; reduce biases by selecting disinterested experts, by using the devil's advocate procedure with judgmental methods, and by using auditing procedures such as review panels.

3. Comprehensiveness: Consider all reasonable alternative actions, including no action. Estimate the opportunities lost by abandoning a technology, and take into account the costs and risks of substituting other credible options. When making these estimates, carefully consider not only concentrated and immediate effects, but also widely distributed and follow-on effects.

4. Openness/Transparency: Take into account the interests of all potentially affected parties, and keep the process open to input from those parties.

5. Simplicity: Use methods that are no more complex than necessary

6. Triage: Give precedence to ameliorating known and proven threats to human health and environmental quality over acting against hypothetical risks.

7. Symmetrical treatment: Treat technological risks on the same basis as natural risks; avoid underweighting natural risks and overweighting human-technological risks. Fully account for the benefits of technological advances.

8. Proportionality: Consider restrictive measures only if the potential impact of an activity has both significant probability and severity. In such cases, if the activity also generates benefits, discount the impacts according to the feasibility of adapting to the adverse effects. If measures to limit technological advance do appear justified, ensure that the extent of those measures is proportionate to the extent of the probable effects.

9. Prioritize (Prioritization): When choosing among measures to ameliorate unwanted side effects, prioritize decision criteria as follows: (a) Give priority to risks to human and other intelligent life over risks to other species; (b) give non-lethal threats to human health priority over threats limited to the environment (within reasonable limits); (c) give priority to immediate threats over distant threats; (d) prefer the measure with the highest expectation value by giving priority to more certain over less certain threats, and to irreversible or persistent impacts over transient impacts.

10. Renew and Refresh: Create a trigger to prompt decision makers to revisit the decision, far enough in the future that conditions may have changed significantly.

(More 2005)

Appendix G: Abstracts

Abstract

The thesis consists of three parts: the first being on rationality and its import in tackling all questions facing us in our lives, not only a reduced domain of scientific investigation; the second, metaphysical in nature, forming an essay on the nature of the world, especially as informed by the principles of rationality sketched in the previous part; and the third, applying the findings of the previous sections to sentient agents – among them humans – in this universe.

I argue that the rational approach is the best way to approach all questions facing us in our lives. Only by being rational can we extract as much information from our environment as possible. Our best way of gaining knowledge should quite naturally also influence our spirituality, our ethics, our view of the meaning of life and so on. It is important to apply the open standards of rationality to all areas of interest to humans. The agent centric approach is central to the thesis. For individuals, knowledge means having a good mental model of the world: the closer to the actual effective factors in the world, the more potential there is for action leading to achievement of goals. Ignorance condemns one to inaction and passivity. The litmus test for knowledge – and philosophy – is this: does it change the way we view the world, our life, and, ultimately and most importantly, the way we act?

Rationality quickly leads to two metaphysical commitments: to naturalism, and to monism. Naturalism, in any non-trivial or question-begging sense, must mean that there are no separate realms – neither for spirits, gods, or minds. Monism, the doctrine that there is certain sense of oneness about seemingly differing subject matters, is here proposed in a radical form: the oneness of the world, indifferent to categories of human inquiry and categorization. Central to the thesis is mind-body monism, the idea that mind and matter are not distinct, but different aspects of the same thing. The core metaphysical assumption of the thesis is that ontological stratifications can be variously drawn and always come at a price; that the adoption of ontic structural realism reduces this price; and that levelism is false and lies at the heart of many philosophical errors. Dispositional and causal structure are presumed to be fundamental. Considerations from philosophy of mind lead us to the astonishing conclusion that neither meaning nor qualia are in the head, but arise from causally induced representations of the world.

Humans are only a subset of possible persons and this forces us to rethink our place in the universe. Consciousness and qualia are not a prerogative of humans, but result from certain matter configurations. Furthermore, ethics will not require persons as subjects, but qualia bearers. Also,

there is no “essential” self to a person, but the self arises as a matter of representation. Finally, the concept of free will, where the word “free” actually means anything, is shown to be largely incoherent: it presupposes the disconnection of humans in regard to the universe. I will offer a different concept, optimal will, to replace “free will”. Achieving optimal will is a difficult process; but one worth pursuing and which will achieve (in the limit) what philosophers probably have in mind when speaking of “free” will.

In the universe meaning arises only in relation to the environment. That does not make it relative in a nihilistic way, but in an exploratory way. We know that here and now, certain sets of values befit us and our situation. But it is also possible to transcend them, and the more we transcend our current limitations the different our values will become. In closing, a deeply rational, naturalistic view of the universe leads not to despair, but to hope and to a vision for the future.

Kurzfassung

Die Dissertation besteht aus drei Teilen: der erste behandelt Rationalität und deren Bedeutung für alle Fragen des Lebens, nicht nur für einen reduzierten – für wissenschaftliche Untersuchungen reservierten – Teilbereich der Welt; der zweite Teil ist metaphysischer Natur und skizziert den postulierten Aufbau der Welt anhand der im ersten Teil erläuterten Prinzipien. Im dritten Teil frage ich – und gebe vorläufige Antworten – wie die Ergebnisse der vorherigen Teile unseren Blickwinkel auf empfindsame Wesen des Universums – darunter Menschen – ändern.

Nur wenn wir rational sind können wir das Maximum an Information aus unserer Umwelt extrahieren. Unser beste Weg zum Erkenntnisgewinn sollte auch Leitfaden für unsere Spiritualität, Ethik, unsere Ansichten über den Sinn des Lebens etc sein. Es ist wichtig die sich mit unserem Erkenntnisstand ändernden Standards der Rationalität auf alle menschlichen Unterfangen anzuwenden. Für Individuen bedeutet Wissen gute mentale Modelle der Welt zu besitzen: je genauer effektive Faktoren in der Welt gespiegelt werden, umso besser können angestrebte Ziele erreicht werden. Unwissenheit führt zu Inaktivität und Passivität. Die Bewährungsprobe für Wissen und Philosophie ist die: werden durch sie die Art und Weise wie wir die Welt, unser Leben, und – letztlich am Wichtigsten – die Art und Weise wie wir handeln verändert?

Rationales Denken führt schnell zur Adoption zweier metaphysischer Annahmen: Naturalismus und Monismus. Ein nicht-trivialer Naturalismus bedeutet, dass es keine verschiedenen

„gesetzlichen“ Ebenen für Geister, Gott, das Bewusstsein etc gibt. Monismus in der gegenständlichen Arbeit bedeutet das Postulat der Einheit der Welt, unabhängig von menschlichen Kategorisierungen. Von besonderer Bedeutung ist die Einheit von Geist und Körper: Geist und Körper sind nur verschiedene Aspekte derselben zugrunde liegenden Sache. Die zentrale metaphysische Annahme der Dissertation ist, dass ontologische Gliederungen auf verschiedene Arten und Weisen gezogen werden können, aber immer einen Preis haben; dass der ontologische strukturelle Realismus diesen Preis minimiert, und dass ontologisches Ebenen-Denken falsch und verantwortlich für viele philosophische Fehler ist. Die fundamentalen ontologischen Kategorien sind dispositionale und kausale Strukturen. Erwägungen aus der Philosophie des Geistes führen zu dem erstaunlichen Schluss dass weder Bedeutung noch Qualia im Kopf verankert sind, sondern durch kausal induzierte Repräsentationen der Welt entstehen.

Menschen sind nur eine Teilmenge möglicher Personen – dies zwingt uns die Stellung des Menschen im Kosmos zu überdenken. Bewusstsein und Qualia sind ebenfalls kein Privileg der menschlichen Spezies, sondern resultieren aus spezifischen Materiekonfigurationen. Ethik wird zudem nicht auf Personen beschränkt sein, sondern auf alle Qualia-Träger ausgedehnt werden. Auch gibt es keine “Essenz” einer Person – das gefühlte “Selbst” ist eine spezifische Eigenrepräsentation. Schlussendlich wird gezeigt dass das Konzept des freien Willens inkohärent ist – es setzt eine Spaltung von Mensch und Umwelt voraus die schwer aufrecht zu erhalten ist. Ich entwickle ein anderes Konzept – das des *optimalen Willens* – das den freien Willen ersetzt. Optimalen Willen zu erreichen ist ein schwieriger Prozess – aber ein Prozess auf den es Wert ist sich einzulassen, und an dessen Ende das steht was Philosophen gemeinhin mit “freien” Willen zu umschreiben versuchen.

Im Universum entsteht Sinn nur in Relation zur Umgebung. Das macht den Sinn nicht relativ in einem nihilistischen Sinne, sondern in einem explorativen Sinne. Wir wissen, dass hier und jetzt gewisse Werte uns und unserer Situation angemessen sind. Aber es ist möglich diese Werte zu transzendieren, und je mehr wir unsere jetzigen Limitationen überschreiten, je mehr werden sich unsere Werte weiterentwickeln. Abschließend bleibt zu sagen, dass ein tief gehend rationaler, naturalistischer Blick auf das Universums nicht zu Sinnleere führt, sondern zu Hoffnung und zu einer Vision für die Zukunft.

Appendix H: Curriculum Vitae

CONTACT INFORMATION

mail: guenther.greindl@gmail.com

www: http://www.complexitystudies.org

Affiliation: University of Vienna, Department of Philosophy

Address: Institut für Philosophie Forschungsbereich Wissenschaftstheorie: Kulturen und Technologien des Wissens. A - 1010 Wien Austria Universitätsstr. 7

EDUCATION

Oct 2006 - 01/2010 Doctoral Studies in Philosophy of Science.

2002 – 2006: Software and Information Engineering; Specialization in Artificial Intelligence. (TU Vienna); B. Sc. Excellent Grades.

1995 - 1999: Law (University of Vienna); Mag. iur. Ranking: Top 1%

EMPLOYMENT HISTORY

2007 – present: Lecturer at University of Vienna (Introductory Computer Science and Artificial Intelligence)

2006 – 2007: Tutor at University of Vienna

2001 – 2006: Lexisnexis Austria (IT Department)

2000: Nine months legal intern at court (to complete legal training)

1999 – 2000: Legal Assistant at Dr. Scheiber & Wessely, Attorneys at law

1994 – 1995: Military Service. One year voluntary service in Austrian Military. Training as Officer. Present: Officer of the Reserve with the rank of Lieutenant.

TECHNICAL QUALIFICATIONS

Web: (X)HTML, CSS, XML, CMS and Blogging Software configuration

Programming: Python, JAVA, PHP, Perl, C

Databases: MySQL, PostgreSQL

System Administration: Linux (Suse, Debian and derived), Windows 2000 und XP

Other: Security and Networking (Firewalling, Service vulnerabilty, TCP/IP)

LANGUAGES

German; fluent - mother tongue

English; fluent - bilingual upbringing

French; 6 years in school

PERSONAL

Titles: Mag. iur. B. Sc. techn.

Interests: Science (Math, Physics), Philosophy, Science Fiction, Net Culture, Dancing (Salsa)

Appendix I: References

Adams, F. C. & Laughlin, G. (1997) A Dying Universe: The Long-term Fate and Evolution of Astrophysical Objects, Reviews of Modern Physics 69: 337.

Aerts, D.; Apostel, L.; Moor, B. D.; Hellemans, S.; Maex, E.; Belle, H. V. & Veken, J. V. d. (**1994**) *World Views. From Fragmentation to Integration*. VUB Press, Brussels (2007 Internet Edition).

Ainsworth, P. M. (**2009**) *Newman's Objection*, British Journal for the Philosophy of Science 60: 135-171.

Albahari, M. (**2006**) *Analytical Buddhism: The Two-Tiered Illusion of Self*. Palgrave Macmillan, New York, NY.

Albert, D. Z. (**1992**) *Quantum Mechanics and Experience*. Harvard University Press, Cambridge, MA.

Albert, H. (**1996**) *Der Mythos des Rahmens und der Moderne Antirealismus. Zur Kritik des idealistischen Rückfalls im gegenwärtigen Denken*. In: Gadenne, V. und Wendel, H. (Ed.), *Rationalität und Kritik*, Mohr, Tübingen.

Albert, H. (**1991**) *Traktat über kritische Vernunft*. Mohr, Tübingen (5th Edition).

Albrecht, A. & Sorbo, L. (**2004**) *Can the Universe Afford Inflation?*, arXiv:hep-th/0405270v2.

Alexander, J. (**2007**) *The Structural Evolution of Morality*. Cambridge University Press, Cambridge, UK.

Anderson, M. (**2003**) *Embodied Cognition. A Field Guide*, Artificial Intelligence 149: 91-130.

Aristotle (**MET**) *Metaphysics*, URL="http://ebooks.adelaide.edu.au/a/aristotle/metaphysics/complete.html", Retrieved on 02.07.2009 (Translated by W. D. Ross).

Aristotle (**PHYS**) *Physics (Book II, Part 9)*, URL="http://classics.mit.edu/Aristotle/physics.2.ii.html", Retrieved on 20.08.2009.

Arthur, W. B. (**1994**) *Inductive Reasoning and Bounded Rationality. (The El Farol Problem)*, American Economic Review 84: 406-411.

Audi, R. (**2001**) *The Architecture of Reason. The Structure and Substance of Rationality*. Oxford University Press, Oxford, UK.

Autumn, K.; Sitti, M.; Liang, Y. A.; Peattie, A. M.; Hansen, W. R.; Sponberg, S.; Kenny, T. W.; Fearing, R.; Israelachvili, J. N. & Full, R. J. (**2002**) *Evidence for Van der Waals Adhesion in Gecko Setae*, Proceedings of the National Academy of Sciences of the United States of America 99: 12252-12256.

Axelrod, R. (**1984**) *The Evolution of Cooperation*. Basic Books, New York, NY.

Balaguer, M. (**2009**) *Realism and Anti-Realism in Mathematics*. In: Irvine, A. D. (Ed.), *Philosophy of Mathematics*, Elsevier/North-Holland, Amsterdam, NL.

Balaguer, M. (**1998**) *Platonism and Anti-Platonism in Mathematics*. Oxford University Press, New York, NY.

Ball, P. (**2008**) *Facing the Music*, Nature 453: 160-162.

Bambach, R. K.; Knoll, A. H. & Wang, S. C. (**2004**) *Origination, Extinction, and Mass Depletions of Marine Diversity*, Paleobiology 30: 522-542.

Barbour, J. (**1999**) *The End of Time. The Next Revolution in Our Understanding of the Universe*. Phoenix House, London, UK.

Bargh, J. & Chartrand, T. (**1999**) *The Unbearable Automaticity of Being*, American Psychologist 54: 479, 462.

Barnes, E. (**1991**) *The Causal History of Computational Activity: Maudlin and Olympia*, The Journal of Philosophy 88: 304-316.

Baron, J. (**2008**) *Thinking and Deciding*. Cambridge University Press, New York, NY (4th Edition).

Barrett, J. A. (**1999**) *The Quantum Mechanics of Minds and Worlds*. Oxford University Press, Oxford, UK (Reprint Edition).

Bartley, W. W. I. (**1984**) *The Retreat to Commitment*. Open Court, Chicago, IL (2nd Edition).

Bassani, G. (**1962**) *The Garden of the Finzi-Continis*. Knopf, New York, NY.

Bechara, A.; Damasio, H. & Damasio, A. R. (**2000**) *Emotion, Decision Making and the Orbitofrontal Cortex*, Cerebral Cortex 10: 295-307.

Bedau, M. A. (**1997**) *Weak Emergence*. In: Tomberlin, J. E. (Ed.), *Philosophical Perspectives 11: Mind, Causation and Word*, Blackwell Publishers, Malden, MA.

Berger, T. W. & Glanzman, D. (**2005**) *Toward Replacement Parts for the Brain: Implantable Biomimetic Electronics as Neural Prostheses*. MIT Press, Cambridge, MA.

Bernoulli, D. (**1738**) *Exposition of a New Theory on the Measurement of Risk*, Econometrica 22: 23-36.

Bernstein, M. (**2004**) *Neo-speciesism*, Journal of Social Philosophy 35: 380-390.

Binmore, K. G. (**2005**) *Natural Justice*. Oxford University Press, New York, NY.

Bird, A. (**2007**) *Nature's Metaphysics: Laws and Properties*. Clarendon Press, Oxford, UK.

Blackmore, S. (**1999**) *The Meme Machine*. Oxford University Press, Oxford, UK.

Block, N. (**1980**) *Troubles With Functionalism*. In: Block, N. (Ed.), *Introduction: What Is Functionalism?*, Harvard University Press, Cambridge, MA.

Bloom, P. & Weisberg, D. S. (**2007**) *Childhood Origins of Adult Resistance to Science*, Science 316: 996-997.

Boltzmann, L. (**1895**) *On Certain Questions of the Theory of Gases*, Nature 51: 413-415.

Borges, J. L. (**1942**) *Funes el memorioso*, La Nación.

Bostrom, N. (**2002**) *Existential Risks. Analyzing Human Extinction Scenarios and Related Hazards*, Journal of Evolution and Technology 9.

Bostrom, N. (**2003**) *Are We Living in a Computer Simulation?*, The Philosophical Quarterly 53: 243-255.

Bostrom, N. & Cirkovic, M. M. (**2008**) *Global Catastrophic Risks*. Oxford University Press, Oxford, UK.

Bounama, C.; von Bloh, W. & Franck, S. (**2004**) *Das Ende des Raumschiffs Erde*, Spektrum der Wissenschaft 10: 52-59.

Boyd, R. & Richerson, P. J. (**2005**) *The Origin and Evolution of Cultures*. Oxford University Press, New York, NY.

Boyer, P. (**2002**) *Religion Explained. The Evolutionary Origins of Religious Thought*. Basic Books, New York, NY (Reprint Edition).

Bradbury, J. (**2004**) *The Routledge Companion to Medieval Warfare*. Routledge, London, UK.

Branden, N. (**1989**) *Judgment Day: My Years With Ayn Rand*. Houghton Mifflin, Boston, MA.

Brandt, R. (**1979**) *A Theory of the Good and the Right*. Clarendon Press, Oxford, UK.

Breer, P. (**1989**) *The Spontaneous Self: Viable Alternatives to Free Will*. Institute For Naturalistic Philosophy, Cambridge, MA.

Bryson, J. (**2008**) *Embodiment versus Memetics*, Mind and Society 7: 77-94.

Buechner, J. (**2008**) *Gödel, Putnam, and Functionalism: A New Reading of Representation and Reality*. MIT Press, Cambridge, MA.

Buggle, F. (**1992**) *Denn Sie Wissen Nicht, Was Sie Glauben. Oder Warum Man Redlicherweise Nicht Mehr Christ Sein Kann. Eine Streitschrift*. Alibri, Aschaffenburg (2nd Edition).

Bunge, M. A. (**2006**) *Chasing Reality. Strife Over Realism*. University of Toronto Press, Toronto, CA.

Bunge, M. A. & Mahner, M. (**2004**) *Über die Natur der Dinge. Materialismus und Wissenschaft*. Hirzel, Stuttgart.

Burwood, S. (**2009**) *Are We Our Brains?*, Philosophical Investigations 32: 113-133.

Callender, C. (**Forthcoming**) *Time in Physics*, Draft (URL="http://philosophy.ucsd.edu/faculty/ccallender/index_files/Time%20in%20Physics.doc", retrieved on 21.01.2008) .

Camejo, S. A. (**2006**) *Skurrile Quantenwelt*. Fischer, Frankfurt am Main.

Campbell, D. (**1974**) *Evolutionary Epistemology*. In: Schilpp, P. A. (Ed.), *The Philosophy of Karl R. Popper*, Open Court, LaSalle, IL.

Carlson, E. & Olsson, E. J. (**2001**) *The Presumption of Nothingness*, Ratio 14: 203-221.

Carroll, L. (**1871**) *Through the Looking-Glass, and What Alice Found There*.

Carroll, S. (**2006**) *Boltzmann's Anthropic Brain*, URL="http://blogs.discovermagazine.com/cosmicvariance/2006/08/01/boltzmanns-anthropic-brain/", Retrieved on 23.01.2010.

Casebeer, W. D. (**2003**) *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition*. MIT Press, Cambridge, MA.

Chakravartty, A. (**2007**) *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge University Press, Cambridge, UK.

Chalmers, D. (**1996b**) *Does a Rock Implement Every Finite-State Automaton?*, Synthese 108: 309-333.

Chalmers, D. J. (**2003**) *Consciousness and its Place in Nature*. In: Stich, S. & Warfield, F. (Ed.), *Blackwell Guide to the Philosophy of Mind*, Blackwell, Malden, MA.

Chalmers, D. J. (**1995**) *Facing Up to the Problem of Consciousness*, Journal of Consciousness Studies 2: 200-219.

Chalmers, D. J. (**1996a**) *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press, New York, NY.

Chung-yuan, C. (**1970**) *Creativity and Taoism. A Study in Chines Philosophy, Art and Poetry.* Harper and Row, New York, NY.

Church, A. (**1936**) *An Unsolvable Problem in Elementary Number Theory'*, The American Journal of Mathematics 58: 345–363.

Clark, A. (**2001**) *Mindware. An Introduction to the Philosophy of Cognitive Science.: An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, Oxford, UK.

Clark, A. & Chalmers, D. J. (**1998**) *The Extended Mind*, Analysis 58: 10-23.

Clark, T. (**1999**) *Fear of Mechanism. A Compatibilist Critique of "The Volitional Brain"*, Journal of Consciousness Studies 6: 279-293.

Clarke, A. C. (**1953**) *Childhood's End*.

Clarke, A. C. (**1984**) *1984 Spring: A Choice of Futures*.

Cole, D. (**2000**) *Inverted Spectrum Arguments*, URL="http://www.d.umn.edu/~dcole/inverted_spectrum.htm", Retrieved on 06.06.2008.

Cole, D. (**2004**) *The Chinese Room Argument*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/chinese-room/".

Conway, J. & Kochen, S. (**2008**) *The Strong Free Will Theorem*, arXiv:0807.3286v1 [quant-ph].

Copeland, J. (**2004**) *Computation*. In: Floridi, L. (Ed.), *The Blackwell Guide to the Philosophy of Computing and Information*, Blackwell Publishing, Malden, MA.

Creel, H. G. (**1970**) *What is Taoism?*. University Of Chicago Press, Chicago, IL.

Croft, W. & Cruse, D. A. (**2004**) *Cognitive Linguistics*. Cambridge University Press, Cambridge, UK.

Curry, O. (**2006**) *Who's Afraid of the Naturalistic Fallacy?*, Evolutionary Psychology 4: 234-247.

Dainton, B. (**2001**) *Time and Space*. Acumen, Chesham, UK.

Damasio, A. (**1994**) *Descartes' Error. Emotion, Reason and the Human Brain*. Putnam and Sons, New York, NY (Vintage 2006 Edition).

Davids, T. W. R. (**1890**) *The Questions of King Milinda*, URL="http://www.holyebooks.org/budhism/milinda.html", Retrieved on 27.08.2009 (Translation).

Davidson, D. (**1987**) *Knowing One's Own Mind*, Proceedings and Addresses of the American Philosophical Association 60: 441-458.

Dawkins, R. (**1989**) *The Selfish Gene*. Oxford University Press, Oxford, UK (2nd Edition).

Dawkins, R. (**2006**) *The God Delusion*. Houghton Mifflin, Boston, MA.

DeGrazia, D. (**2005**) *Human Identity and Bioethics*. Cambridge University Press, Cambridge, MA.

Dehaene, S. (**2009**) *Origins of Mathematical Intuitions. The Case of Arithmetic*, Annals of the New York Academy of Sciences 1156: 232-259.

Dennett, D. C. (**1984**) *Elbow Room. The Varieties of Free Will Worth Wanting*. Clarendon Press, Oxford, UK.

Dennett, D. C. (**1995**) *Darwin's Dangerous Idea. Evolution and the Meanings of Life*. Simon and Schuster, New York, NY.

Dennett, D. C. (**2006**) *Breaking the Spell. Religion as a Natural Phenomenon*. Penguin Books, New York, NY.

Deschner, K. (**1986**) *Kriminalgeschichte des Christentums*. Rowohlt Verlag, Hamburg.

Deutsch, D. (**1998**) *David Deutsch's Many Worlds*, Frontiers Magazine.

Deutsch, K. W. (**1963**) *The Nerves of Government; Models of Political Communication and Control*. Free Press of Glencoe, London, UK.

Dipert, R. R. (**1997**) *The Mathematical Structure of the World. The World as Graph*, The Journal of Philosophy 94: 329-358.

Dorato, M. (**2006**) *Properties and Dispositions. Some Metaphysical Remarks on Quantum Ontology*, URL="http://philsci-archive.pitt.edu/archive/00002932/", Retrieved on 20.08.2009.

Doris, J. M. & Stich, S. P. (**2005**) *Empirical Perspectives on Ethics*. In: Jackson, F. & Smith, M. (Ed.), *The Oxford Handbook of Contemporary Philosophy*, Oxford University Press, Oxford, UK.

Dowe, P. (**2007**) *Causal Processes*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/causation-process/".

Doya, K.; Ishii, S.; Pouget, A. & Rao, R. P. W. (**2007**) *Bayesian Brain. Probabilistic Approaches to Neural Coding*. MIT Press, Cambridge, MA.

Drescher, G. L. (**2006**) *Good and Real. Demystifying Paradoxes from Physics to Ethics*. MIT Press, Cambridge, MA.

Dresner, E. (**2008**) *Turing-, Human- and Physical Computability. An Unasked Question*, Minds and Machines 18: 349-355.

Dretske, F. (**1996**) *Phenomenal Externalism or If Meanings Ain't in the Head, Where Are Qualia?*, Philosophical Issues 7: 143-158.

Dretske, F. I. (**1995**) *Naturalizing the Mind*. MIT Press, Cambridge, MA.

Duerr, D. (**2009**) *Bohmian Mechanics. The Physics and Mathematics of Quantum Theory*. Springer, New York, NY.

Dyson, F. (**1979**) *Time without End. Physics and Biology in an Open Universe*, Reviews of Modern Physics 51: 447-460.

Earman, J. (**2007**) *Aspects of Determinism in Modern Physics*. In: Butterfield, J. & Earman, J. (Ed.), *Handbook of the Philosophy of Science. Philosophy of Physics*, North Holland/Elsevier, Amsterdam, NL.

Earman, J. (**2004**) *Determinism. What We Have Learned and What We Still Don't Know*. In: Campbell, J. K.; Rourke, M. O. & Shier, D. (Ed.), *Determinism, Freedom, and Agency*, MIT Press, Cambridge, MA.

Edelman, G. M. & Gally, J. A. (**2001**) *Degeneracy and Complexity in Biological Systems*, Proceedings of the National Academy of Sciences of the United States of America 98: 13763-13768.

Edelman, G. M. & Tononi, G. (**2000**) *A Universe of Consciousness. How Matter Becomes Imagination*. Basic Books, New York, NY.

Egan, G. (**1992**) *Quarantine*.

Egan, G. (**1994**) *Permutation City.*

Egner, R. E. & Denonn, L. E. (**2009**) *The Basic Writings of Bertrand Russell*. Routledge, Oxford, UK.

Ekbia, H. R. (**2008**) *Artificial Dreams. The Quest for Non-Biological Intelligence*. Cambridge University Press, Cambridge, MA.

Epstein, J. M. (**2006**) *Generative Social Science. Studies in Agent-Based Computational Modeling*. Princeton University Press, Princeton, NJ.

Epstein, J. M. & Axtell, R. L. (**1996**) *Growing Artificial Societies. Social Science from the Bottom Up*. MIT Press, Cambridge, MA.

Esfeld, M. (**2009b**) *The Rehabilitation of a Metaphysics of Nature*. In: Heidelberger, M. & Schiemann, G. (Ed.), *The Significance of the Hypothetical in the Natural Sciences*, de Gruyter, New York, NY.

Esfeld, M. (**2009a**) *The Modal Nature of Structures in Ontic Structural Realism*, International Studies in the Philosophy of Science 23: 179–194.

Esfeld, M. & Lam, V. (**2008**) *Moderate Structural Realism about Space-time*, Synthese 160: 27-46.

Ettinger, R. C. W. (**1964**) *The Prospect of Immortality.* Doubleday, Garden City, N.Y.

Everett, H. I. (**1957**) *Relative State Formulation of Quantum Mechanics*, Reviews of Modern Physics 29: 454-462.

Fara, M. (**2006**) *Dispositions*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/dispositions/".

Fine, A. (**1984**) *The Natural Ontological Attitude*. In: Leplin, J. (Ed.), *Scientific Realism*, University of California Press, Berkeley, CA.

Fishman, Y. (**2009**) *Can Science Test Supernatural Worldviews?*, Science and Education 18: 813-837.

Flanagan, O. (**2002**) *The Problem of the Soul*. Basic Books, New York, NY.

Flanagan, O. (**2007**) *The Really Hard Problem. Meaning in a Material World*. MIT Press, Cambridge, MA.

Flew, A. (**1975**) *Thinking About Thinking – or Do I Sincerely Want to Be Right?*. Collins Fontana, Glasgow.

Floridi (**2004**) *On the Logical Unsolvability of the Gettier Problem*, Synthese 142: 61–79.

Floridi (**2009**) *Against Digital Ontology*, Synthese 168: 151-178.

Floridi, L. (**1999**) *Information Ethics. On the Philosophical Foundations of Computer Ethics*, Ethics and Information Technology 1: 37-56.

Floridi, L. (**2008a**) *The Method of Levels of Abstraction*, Minds and Machines 18: 303-329.

Floridi, L. (**2008b**) *A Defence of Informational Structural Realism*, Synthese 161: 219-253.

Fraassen, B. v. (**1980**) *The Scientific Image*. Oxford University Press, Oxford, UK.

Fraassen, B. v. (**1984**) *Belief and the Will*, Journal of Philosophy 81: 235-256.

Franzen, T. (**2004**) *Inexhaustibility. A Non-Exhaustive Treatment*. Association for Symbolic Logic, Urbana, IL.

Fredkin, E. (**1992**) *Finite Nature*, Proceedings of the XXVIIth Rencontre de Moriond.

Fredkin, E. (**2003**) *An Introduction to Digital Philosophy*, International Journal of Theoretical Physics 42: 189-247.

French, S. & Ladyman, J. (**1999**) *Reinflating the Semantic Approach*, International Studies in the Philosophy of Science 13: 103.

Friston, K. & Stephan, K. (**2007**) *Free-energy and the Brain*, Synthese 159: 417-458.

Futuyma, D. J. (**1998**) *Evolutionary Biology.* Sinauer Associates, Sunderland, MA (3rd Edition).

Gandhi, M. (**1927**) *Young India 1924-1926*.

Gelder, T. v. (**1998**) *The Dynamical Hypothesis in Cognitive Science*, Behavioral and Brain Sciences 21: 1-14.

Gelder, T. v. (**2005**) *Teaching Critical Thinking. Some Lessons from Cognitive Science.*, College Teaching 45: 1-6.

Gettier, E. L. (**1963**) *Is Justified True Belief Knowledge?*, Analysis 23: 121-123.

Gigerenzer, G. (**2003**) *Why does Framing influence Judgement?*, Journal of General Internal Medicine 18: 960-961.

Gillies, D. (**1998**) *The Duhem Thesis and the Quine Thesis*. In: Curd & Cover (Ed.), *Philosophy of Science. The Central Issues*, W.W. Norton and Company, New York, NY.

Goethe (**MR**) *Maximen und Reflexionen*. , Goethe-BA Bd. 18, S. 615.

Goethe (**WMW**) *Wilhelm Meisters Wanderjahre*. , Goethe-HA Bd. 8, S. 283.

Goldstein, J. (**2002**) *One Dharma. The Emerging Western Buddhism*. HarperSanFrancisco, San Francisco, CA.

Goodman, N. (**1983**) *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, MA (4th Edition).

Gottschall, J. (**2007**) *Fictional Selection*, New Scientist 3: 38-41.

Gould, S. J. (**1997**) *Nonoverlapping Magisteria*, Natural History 106: 16-22.

Gowder, P. (**2008**) *Beliefs Require Reasons, or: Is the Pope Catholic? Should he be?*, URL="http://www.overcomingbias.com/2008/11/beliefs-require.html", Retrieved on 27.06.2009.

Graham (**2007**) *How To Do Philosophy*, URL="http://www.paulgraham.com/philosophy.html", Retrieved on 23.03.2008.

Greene, J. D. (**2002**) *The Terrible, Horrible, No Good, Very Bad Truth About Morality And What To Do About It*, Thesis at Princeton University.

Greenspan, P. (**2004**) *Practical Reasoning and Emotion*. In: Mele, A. R. & Rawling, P. (Ed.), *The Oxford Handbook of Rationality*, Oxford University Press, Oxford, UK.

Grice, H. P. (**1975**) *Logic and Conversation*. In: Cole, P. & Morgan., J. (Ed.), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York, NY.

Grimes, J. A. (**1996**) *A Concise Dictionary of Indian Philosophy. Sanskrit Terms Defined in English*. State University of New York Press, Albany, NY.

Grimm, V. & Railsback, S. F. (**2005**) *Individual-based Modeling and Ecology*. Princeton University Press, Princeton, NJ.

Haidt, J. (**2006**) *The Happiness Hypothesis. Finding Modern Truth in Ancient Wisdom*. Basic Books, New York, NY.

Hajek, A. & Hartmann, S. (**Forthcoming**) *Bayesian Epistemology*. In: Dancy, J.; Sosa, E. & Steup, M. (Ed.), *Blackwell Companion to Epistemology*, Routledge, Oxford, UK.

Handfield, T. (**2009**) *Dispositions and Causes*. Clarendon Press, Oxford, UK.

Hanson, R. (**2007**) *Catastrophe, Social Collapse, and Human Extinction*. In: Bostrom, N. & Cirkovic, M. M. (Ed.), *Global Catastrophic Risks*, Oxford University Press, Oxford, UK.

Harnad, S. (**2003**) *Can a Machine Be Conscious? How?*, Journal of Consciousness Studies 10: 67-75.

Harris, J. (**2005**) *Scientific Research is a Moral Duty*, Journal of Medical Ethics 31: 242-248.

Harris, S. (**2004**) *The End Of Faith*. Free Association Press, New York, NY.

Hartung, J. (**1995**) *Love Thy Neighbor. The Evolution of In-Group Morality*, Skeptic 3: 86-99.

Hartung, J. (**1996**) *Prospects for Existence: Morality And Genetic Engineering*, Skeptic 4: 62-71.

Harvey, P. (**2000**) *An Introduction to Buddhist Ethics. Foundations, Values, and Issues*. Cambridge University Press, Cambridge, UK.

Hawkins, J. (**2004**) *On Intelligence*. Holt Paperbacks, New York, NY.

Hawthorne, J. (**2001**) *Causal Structuralism*, Philosophical Perspectives 15: 361–378.

Heath, J. & Anderson, J. (**Forthcoming**) *Procrastination and the Extended Will*. In: Andreou, C. & White, M. (Ed.), *The Thief of Time. Philosophical Essays on Procrastination*, Oxford University Press, Oxford, UK.

Heathcote, A. (**2003**) *Quantum Heterodoxy. Realism at the Plank Length*, Science and Education 12: 513-529.

Heil, J. (**2006**) *Commentary*. In: Esfeld, M. (Ed.), *John Heil. Symposium on His Ontological Point of View*, Ontos, Frankfurt am Main.

Heil, J. (**1998**) *Philosophy of Mind. A Contemporary Introduction*. Routledge, London, UK.

Heil, J. (**2003**) *From an Ontological Point of View*. Clarendon Press, Oxford, UK.

Heil, J. (**2005a**) *Kinds and Essences*, Ratio 18: 405–419.

Heil, J. (**2005b**) *Dispositions*, Synthese 144: 343-356.

Heinlein, R. A. (**1973**) *Time Enough for Love*.

Hempel, C. G. & Oppenheim, P. (**1948**) *Studies in the Logic of Explanation*, Philosophy of Science 15: 135-175.

Heylighen, F. (**2001**) *The Science of Self-organization and Adaptivity*. In: Kiel, L. D. (Ed.), *Knowledge Management, Organizational Intelligence and Learning, and Complexity*, EOLSS Publishers, Oxford, UK.

Heylighen, F.; Cilliers, P. & Gershenson, C. (**2007**) *Complexity and Philosophy*. In: Bogg, J. & Geyer, R. (Ed.), *Complexity, Science, and Society*, Radcliffe Publishing, Oxford, UK.

Holland, J. H. (**1995**) *Hidden Order. How Adaptation Builds Complexity*. Basic Books, New York, NY.

Holman, E. (**2008**) *Panpsychism, Physicalism, Neutral Monism and the Russellian Theory of Mind*, Journal of Consciousness Studies 15: 48-67.

Hooker, B. & Streumer, B. (**2004**) *Procedural and Substantive Practical Rationality*. In: Mele, A. R. & Rawling, P. (Ed.), *The Oxford Handbook of Rationality*, Oxford University Press, Oxford, UK.

Horkheimer, M. & Adorno, T. W. (**1947**) *Dialektik der Aufklärung. Philosophische Fragmente*. Fischer, Frankfurt (15th Edition).

Horwich, P. (**2005**) *Wittgensteinian Bayesianism*, From a Deflationary Point of View 1: 105-128.

Hurka, T. (**2006**) *Value Theory*. In: Copp, D. (Ed.), *The Oxford Handbook of Ethical Theory*, Oxford University Press, New York, NY.

Huxley, J. (**1942**) *Evolution. The Modern Synthesis*. Allen and Unwin, London, UK (3rd Edition).

Hyde, D. (**2005**) *Sorites Paradox*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://stanford.library.usyd.edu.au/archives/sum2009/entries/sorites-paradox/".

Irons, E. A. (**2008**) *Encyclopedia of Buddhism*. Facts on File, New York, NY.

Jackson, F. (**1982**) *Epiphenomenal Qualia*, Philosophical Quarterly 32: 127-136.

Jaynes, E. T. (**2003**) *Probability Theory. The Logic of Science*. Cambridge University Press, New York, NY (Reprint Edition).

Joyce, J. (**2008**) *Bayes' Theorem*, The Stanford Encyclopedia of Philosophy Fall 2008 Edition.

Joyce, J. M. (**2004**) *Bayesianism*. In: Mele, A. R. & Rawling, P. (Ed.), *The Oxford Handbook of Rationality*, Oxford University Press, Oxford, UK.

Kahneman, D.; Slovic, P. & Tversky, A. (**1982**) *Judgment under Uncertainty. Heuristics and Biases*. Cambridge University Press, New York, NY (Reprint Edition).

Kalat, J. W. (**2007**) *Biological Psychology*. Thomson Wadsworth, Belmont, CA (9th Edition).

Kant, I. (**1781**) *Kritik der Reinen Vernunft*.

Kant, I. (**1784**) *Beantwortung der Frage: Was Ist Aufklärung?*, Berlinische Monatsschrift Dezember-Heft: 481-494.

Kary, M. & Mahner, M. (**2002**) *How Would You Know if You Synthesized a Thinking Thing?*, Minds and Machines 12: 61-86.

Kauffman, S. (**2008**) *Reinventing the Sacred. A New View of Science, Reason, and Religion*. Basic Books, New York, NY.

Kemmerer, L. (**2006**) *In Search of Consistency. Ethics and Animals*. Brill, Leiden, NL.

Kim, J. (**2003**) *Philosophy of Mind and Psychology*. In: Ludwig, K. (Ed.), *Donald Davidson*, Cambridge University Press, Cambridge, UK.

Kirkham, R. L. (**1992**) *Theories of Truth. A Critical Introduction*. MIT Press, Cambridge, MA.

Kiverstein, J. (**2007**) *Could A Robot Have A Subjective Point Of View?*, Journal of Consciousness Studies 14: 127-139.

Knill, D. C. & Pouget, A. (**2004**) *The Bayesian Brain. The Role Of Uncertainty in Neural Coding and Computation*, Trends in Neurosciences 27: 712-719.

Koch, C. (**2004**) *The Quest for Consciousness. A Neurobiological Approach*. Roberts and Company Publishers, Denver, CO.

Koch, C. & Tononi, G. (**2008**) *Can Machines Be Conscious?*, IEEE Spectrum Special Report. The Singularity: 55-59.

Kolak, D. (**2004**) *I Am You. The Metaphysical Foundations for Global Ethics*. Springer, Dordrecht, NL.

Kolmogorov, A. N. (**1933**) *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin.

Krucoff, M. W.; Crater, S. W.; Gallup, D. & Others (**2005**) *Music, Imagery, Touch, and Prayer as Adjuncts to Interventional Cardiac Care. The Monitoring and Actualisation of Noetic Trainings (MANTRA) II Randomised Study*, The Lancet 366: 211-217.

Kuhlmann, M. (**2006**) *Quantum Field Theory*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/quantum-field-theory/".

Kuhn, T. S. (**1962**) *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL (3rd Edition).

Kurzweil, R. (**1999**) *The Age of Spiritual Machines. When Computers Exceed Human Intelligence*. Viking, New York, NY.

Ladyman, J. (**1998**) *What Is Structural Realism?*, Studies in History and Philosophy of Science 29A: 409-424.

Ladyman, J. (**2007**) *Structural Realism*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/structural-realism/".

Ladyman, J. (**Forthcoming**) *The Scientistic Stance: The Empirical and Materialist Stances Reconciled*, Synthese Online First(TM): 1-12.

Ladyman, J.; Ross, D.; Spurrett, D. & Collier, J. (**2007**) *Every Thing Must Go. Metaphysics Naturalized*. Oxford University Press, Oxford, UK.

Lakoff, G. & Nunez, R. (**2000**) *Where Mathematics Comes from. How the Embodied Mind Brings Mathematics Into Being*. Basic Books, New York, NY.

Lam, V. (**2006**) *Is a World Only Made up of Relations Possible? A Structural Realist Point of View.* In: Esfeld, M. (Ed.), *John Heil. Symposium on His Ontological Point of View*, Ontos, Frankfurt am Main.

Lamb, D. (**2001**) *The Search for Extraterrestrial Intelligence. A Philosophical Inquiry.* Routledge, London, UK.

Lambalgen, M. v. (**2003**) *Evolutionary Considerations On Logical Reasoning*, Proceedings of Twelfth International Conference on Logic, Methodology and Philosophy of Science .

LaoziMitchell (**1995**) *Tao Te Ching*, URL="http://academic.brooklyn.cuny.edu/core9/phalsall/texts/taote-v3.html", Retrieved on 27.07.2009 (Translated by Steve Mitchell).

LaoziMuller (**2004**) *Daode jing*, URL="http://www.acmuller.net/con-dao/daodejing.html#div-42", Retrieved on 27.08.2009 (Translated by Charles Muller).

Legg, S. (**2008**) *Machine Super Intelligence*, Thesis at University of Lugano, CH.

Leibniz, G. W. (**1698**) *Aufklärung der Schwierigkeiten, die H. Bayle in dem neuen „System der Vereinigung von Seele und Körper" gefunden hat (Hauptschriften zur Grundlegung der Philosophie)*.

Leibniz, G. W. (**1710**) *Essais de théodicée*.

Leiter, B. (**2007**) *Nietzsche's Theory of the Will*, Philosophers' Imprint 7.

Levin, J. (**2004**) *Functionalism*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/functionalism/".

Levy, D. A. (**2003**) *Neural Holism and Free Will*, Philosophical Psychology 16: 205-227.

Lewis, D. (**1980**) *A Subjectivist's Guide to Objective Chance*. In: Carnap, R. & Jeffrey, R. C. (Ed.), *Studies in Inductive Logic and Probability*, University of California Press, Berkeley, CA.

Lewis, D. (**1986**) *On the Plurality of Worlds*. Blackwell, Oxford, UK (Reissued (2001) Edition).

Lloyd, S. (**2006**) *Programming the Universe. A Quantum Computer Scientist Takes on the Cosmos*. Knopf, New York, NY.

Lockwood, M. (**1989**) *Mind, Brain, and the Quantum: The Compound 'I'*. Blackwell, Oxford, UK.

Lockwood, M. (**2005**) *The Labyrinth of Time: Introducing the Universe*. Oxford University Press, New York, NY.

Loll, R. (**2007**) *The Emergence of Spacetime, or, Quantum Gravity on Your Desktop*, arXiv:0711.0273v2 [gr-qc].

Lynch, M. P. (**2001**) *The Nature of Truth. Classic and Contemporary Perspectives*. MIT Press, Cambridge, MA.

Macdonald, G. & Papineau, D. (**2006**) *Teleosemantics. New Philosophical Essays*. Clarendon Press, Oxford, UK.

Maestripieri, D. (**2005**) *Early Experience Affects the Intergenerational Transmission of Infant Abuse in Rhesus Monkeys*, Proceedings of the National Academy of Sciences of the United States of America 102: 9726-9729.

Mahner, M. & Bunge, M. A. (**1997**) *Foundations of Biophilosophy*. Springer, New York, NY.

Mahner, M. & Bunge, M. A. (**2000**) *Philosophische Grundlagen der Biologie*. Springer, Berlin.

Mahner, M. & Bunge, M. A. (**2001**) *Function and Functionalism. A Synthetic Perspective*, Philosophy of Science 68: 75-94.

Mainzer, K. (**1997**) *Thinking in Complexity. The Complex Dynamics of Matter, Mind, and Mankind*. Springer, Berlin (3rd Edition).

Marchal, B. (**2004**) *The Origin of Physical Laws and Sensations*, URL="http://iridia.ulb.ac.be/~marchal/publications/SANE2004MARCHAL.htm", Retrieved on 27.06.2008.

Martin, C. (**1993**) *Power for Realists*. In: Bacon, J.; Campbell, K. & Reinhardt, L. (Ed.), *Ontology, Causality, and Mind. Essays in Honour of D. M. Armstrong*, Cambridge University Press, Cambridge, UK.

Martin, C. (**2008**) *The Mind in Nature*. Oxford University Press, Oxford, UK.

Martinich, A. & Sosa, D. (**2001**) *A Companion to Analytic Philosophy*. Blackwell, Malden, MA.

Maturana, H. R. & Varela, F. J. (**1987**) *The Tree of Knowledge. The Biological Roots of Human Understanding*. New Science Library (Shambhala), London, UK.

Maudlin, T. (**1989**) *Computation and Consciousness*, The Journal of Philosophy 86: 407-432.

Mayr, E. (**1993**) *What Was the Evolutionary Synthesis?*, Trends in Ecology and Evolution 8: 31–33.

Mayr, E. & Provine, W. B. (**1998**) *The Evolutionary Synthesis. Perspectives on the Unification of Biology*. Harvard University Press, Cambridge, MA.

McAllister, J. W. (**1996**) *Beauty and Revolution in Science*. Cornell University Press, Ithaca, NY.

McClelland, J. L. & Rumelhart, D. E. (**1988**) *Explorations in Parallel Distributed Processing. A Handbook of Models, Programs, and Exercises*. MIT Press, Cambridge, MA.

McFadden, J. (**2002**) *The Conscious Electromagnetic Field Theory. The Hard Problem Made Easy*, Journal of Consciousness Studies 9: 45-60.

McLaughlin, B. & Bennett, K. (**2005**) *Supervenience*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/supervenience/".

Mellor, D. H. (**1974**) *In Defense of Dispositions*, The Philosophical Review 83: 157–181.

Merchey, J. (**2004**) *Values of the Wise. Humanity's Highest Aspirations*. Infinity Publishing, Haverford, PA.

Merkle, R. C. (**1992**) *The Technical Feasibility of Cryonics*, Medical Hypotheses 39: 6-16.

Metzinger, T. (**2003**) *Being No One. The Self-Model Theory of Subjectivity*. MIT Press, Cambridge, MA.

Minsky, M. (**2006**) *The Emotion Machine. Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon and Schuster, New York, NY.

Misra, R. (**1998**) *The Integral Advaitism of Sri Aurobindo*. Motilal Banarsidass, Delhi.

Montero, B. (**2006**) *What Does the Conservation of Energy Have to Do with Physicalism?*, Dialectica 60: 383-396.

More, M. (**1997**) *Self-Ownership. A Core Transhuman Virtue*, URL="http://www.maxmore.com/selfown.htm", Retrieved on 09.05.2008.

More, M. (**2005**) *The Proactionary Principle*, URL="http://www.maxmore.com/proactionary.htm", Retrieved on 28.07.2009.

Muller, H. (**2008**) *Why Qualia Are Not Epiphenomenal*, Ratio 21: 85-90.

Mumford, S. (**2004**) *Laws in Nature*. Routledge, London, UK.

Musashi, M. (**BFR**) *A Book of Five Rings*, URL="http://www.miyamotomusashi.com/gorin.htm", Retrieved on 12.08.2009.

Nagel, T. (**1979**) *The Fragmentation of Value*. In: Nagel, T. (Ed.), *Moral Questions*, Cambridge University Press, New York, NY.

Nature (**2009**) *Across the Great Divide*, Nature Physics 5: 309.

Neander, K. (**2004**) *Teleological Theories of Mental Content*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/content-teleological/".

Neumann, J. v. (**1947**) *The Mathematician*. In: Adler, M. J. (Ed.), *Works of the Mind*, University of Chicago Press, Chicago, IL.

Newman, M. E. J.; Barabasi, A.-L. & Watts, D. J. (**2006**) *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, NJ.

Nida-Rümelin, M. (**2002**) *Qualia. The Knowledge Argument*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://www.science.uva.nl/~seop/archives/sum2009/entries/qualia-knowledge/".

Niemann, H.-J. (**1999**) *Reason. A Victim of Nazi Legacy*, The Philosophers' Magazine : 12-13.

Niemann, H.-J. (**2008**) *Die Strategie der Vernunft. Problemlösende Vernunft, Rationale Metaphysik und Kritisch-Rationale Ethik*. Mohr Siebeck, Tübingen (2nd Edition).

Nietzsche (**1888**) *Götzendämmerung*.

Nietzsche, F. (**1879**) *Menschliches Allzumenschliches*.

Nietzsche, F. (**Birth of Tragedy**) *Birth of Tragedy*. Richer Resources Publications, Arlington, VA.

Nietzsche, F. (**Will to Power**) *The Will to Power*. Vintage Books, New York, NY.

Nietzsche, F. W. (**Daybreak**) *Daybreak: Thoughts on the Prejudices of Morality*. Cambridge University Press, Cambridge, UK.

Nietzsche, F. W. (**GayScience**) *The Gay Science: With a Prelude in German Rhymes and an Appendix of Songs*. Cambridge University Press, Cambridge, UK.

Nikolic, H. (**2007**) *Quantum Mechanics: Myths and Facts*, Foundations of Physics 37: 1563-1611.

Noonan, H. W. (**2003**) *Personal Identity*. Routledge, London, UK (2nd Edition).

Norton, J. (**2007**) *Causation as Folk Science*. In: Price, H. & Corry, R. (Ed.), *Causation and the Constitution of Reality*, Oxford University Press, Oxford, UK.

Nowak, M. A. (**2006**) *Evolutionary Dynamics. Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

Nozick, R. (**1974**) *Anarchy, State, and Utopia*. Basic Books, New York, NY.

Nozick, R. (**1993**) *The Nature of Rationality*. Princeton University Press, Princeton, NJ (Paperback Edition).

Nozick, R. (**2001**) *Invariances. The Structure of the Objective World*. Belknap Press of Harvard University Press, Cambridge, MA.

Olson, E. T. (**1997**) *The Human Animal. Personal Identity Without Psychology*. Oxford University Press, Oxford, UK.

Papineau, D. (**2001**) *The Rise of Physicalism*. In: Loewer, B. (Ed.), *Physicalism and its Discontents*, Cambridge University Press, New York, NY.

Papineau, D. (**2003**) *Theories of Consciousness*. In: Smith, Q. & Jokic, A. (Ed.), *Consciousness. New Philosophical Perspectives*, Clarendon Press, Oxford, UK.

Papineau, D. (**1984**) *Representation and Explanation*, Philosophy of Science 51: 550-572.

Papineau, D. (**2002**) *Thinking About Consciousness*. Clarendon Press, Oxford, UK.

Papineau, D. (**2003**) *The Roots of Reason. Philosophical Essays on Rationality, Evolution and Probability*. Oxford University Press, Oxford, UK.

Papineau, D. (**Forthcoming**) *Kripke's Proof That We Are All Intuitive Dualists*, URL="http://www.kcl.ac.uk/ip/davidpapineau/Staff/Papineau/OnlinePapers/Kripke%27s %20Proof.htm", Retrieved on 18.07.2009.

Parfit, D. (**1984**) *Reasons and Persons*. Oxford University Press, New York, NY (Reprint Edition).

Pereboom, D. (**2001b**) *Living without Free Will. The Case for Hard Incompatibilism*. In: Kane, R. (Ed.), *The Oxford Handbook of Free Will*, Oxford University Press, Oxford, UK.

Pereboom, D. (**2001a**) *Living Without Free Will*. Cambridge University Press, Cambridge, UK (Reprint Edition).

Peres, A. (**2002**) *Quantum Theory. Concepts and Methods*. Kluwer Academic Publishers, Dordrecht, NL.

Petta, P. & Trappl, R. (**2001**) *Emotions and Agents*. In: Carbonell, J. G. & Siekmann, J. (Ed.), *Multi-agents Systems and Applications*, Springer, New York, NY.

Piccinini, G. (**2007**) *Computational Modelling vs. Computational Explanation. Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?*, The Australasian Journal of Philosophy 85: 93-115.

Piccinini, G. (**2009**) *Computationalism in the Philosophy of Mind*, Philosophy Compass 4: 515-532.

Pigliucci, M. (**2003**) *On the Relationship Between Science and Ethics*, Zygon 38: 871-894.

Plantinga, A. (**1981**) *Is Belief in God Properly Basic?*, Noûs 15: 41-51.

Pollock, J. & Ismael, J. (**2004**) *So You Think You Exist? In Defense of Nolipsism*. In: Crisp, T.; Davidson, M. & Laan, D. V. (Ed.), *Knowlege and Reality. Essays in Honor of Alvin Plantinga*, Springer, Berlin.

Pollock, J. L. (**2008**) *What Am I? Virtual Machines and the Mind/Body Problem*, Philosophy and Phenomenological Research 76: 237-309.

Popper, K. R. (**1972**) *Objective Knowledge. An Evolutionary Approach*. Clarendon Press, Oxford, UK.

Posner, R. A. (**2004**) *Catastrophe. Risk and Response*. Oxford University Pres, Oxford, UK.

Psillos, S. (**1999**) *Scientific Realism. How Science Tracks Truth*. Routledge, London, UK.

Putnam, H. (**1975**) *The Meaning of 'Meaning'*. In: Putnam, H. (Ed.), *Mind, Language and Reality*, Cambridge University Press, New York, NY.

Putnam, H. (**2005**) *A Philosopher Looks at Quantum Mechanics (Again)**, British Journal for the Philosophy of Science 56: 615-634.

Rae, A. (**2004**) *Quantum Physics, Illusion or Reality?*. Cambridge University Press, Cambridge, UK (2nd Edition).

Rand, A. (**1957**) *Atlas shrugged*.

Raup, D. M. & Sepkoski, J. J. (**1982**) *Mass Extinctions in the Marine Fossil Record*, Science 215: 1501-1503.

Reginster, B. (**2006**) *The Affirmation of Life. Nietzsche on Overcoming Nihilism*. Harvard University Press, Cambridge, MA.

Richerson, P. J. & Boyd, R. (**2005**) *Not By Genes Alone. How Culture Transformed Human Evolution*. University Of Chicago Press, Chicago, IL.

Rieffel, E. G. (**2007**) *Certainty and Uncertainty in Quantum Information Processing*, quant-ph/0702121.

Riegler, A. (**2001**) *Towards a Radical Constructivist Understanding of Science.*, Foundations of Science : 1-30.

Riemen, R. (**2008**) *Nobility of Spirit. A Forgotten Ideal*. Yale University Press, New Haven, CT.

Rockwell, W. T. (**2005**) *Neither Brain nor Ghost. A Nondualist Alternative to the Mind-Brain Identity Theory*. MIT Press, Cambridge, MA.

Ross, D. (**2000**) *Rainforest Realism. A Dennettian Theory of Existence*. In: Ross, D.; Thompson, D. & Brook, A. (Ed.), *Dennett's Philosophy. A Comprehensive Assessment*, MIT Press, Cambridge, MA.

Rovane, C. (**2004**) *Rationality and Persons*. In: Mele, A. R. & Rawling, P. (Ed.), *The Oxford Handbook of Rationality*, Oxford University Press, Oxford, UK.

Roy, D. (**2005**) *Grounding Words in Perception and Action. Computational Insights*, Trends in Cognitive Sciences 9: 389-396.

Russell, B. (**1927**) *The Analysis of Matter*. Harcourt, New York, NY.

Russell, S. J. & Norvig, P. (**2003**) *Artificial Intelligence. A Modern Approach*. Prentice Hall International, Upper Saddle River, NJ (2nd International Edition).

Sacks, O. W. (**1985**) *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. Summit Books, New York, NY.

Sale, K. (**1996**) *Rebels Against The Future. The Luddites And Their War On The Industrial Revolution. Lessons For The Computer Age*. Basic Books, New York, NY.

Sandberg, A. & Bostrom, N. (**2008**) *Whole Brain Emulation. A Roadmap.*, Future of Humanity Institute. Oxford University.

Schauer, F. (**1983**) *Free Speech. A Philosophical Enquiry*. Cambridge University Press, Cambridge, MA.

Schelling, T. C. (**1978**) *Micromotives and Macrobehavior*. Norton, New York, NY.

Schiff, J. L. (**2008**) *Cellular Automata. A Discrete View of the World*. Wiley-Interscience, Hoboken, NJ.

Schlagel, R. H. (**1999**) *Why not Artificial Consciousness or Thought?*, Minds and Machines 9: 3-28.

Schmidt, S. J. (**1987**) *Der Diskurs des Radikalen Konstruktivismus*. Suhrkamp, Frankfurt am Main.

Schmidt-Salomon, M. (**2006**) *Manifest des evolutionären Humanismus. Plädoyer für eine zeitgemäße Leitkultur*. Alibri, Aschaffenburg (2nd Edition).

Schnall, S.; Benton, J. & Harvey, S. (**2008**) *With a Clean Conscience. Cleanliness Reduces the Severity of Moral Judgments*, Psychological Science 19: 1219-1222.

Schoeman, F. D. (**1979**) *On Incapacitating the Dangerous*, American Philosophical Quarterly 16: 27-35.

Schopenhauer, A. (**1859**) *Die Welt als Wille und Vorstellung. Band I* (3rd Edition).

Schwitzgebel, E. & Rust, J. (**Forthcoming**) *The Moral Behavior of Ethicists. Peer Opinion*, Mind.

Searle, J. (**2002**) *I Married A Computer*. In: Richards, J. W. (Ed.), *Are We Spiritual Machines? Ray Kurzweil vs. the Critics of Strong A.I.*, Discovery Institute Press, Seattle, WA.

Searle, J. (**1980**) *Minds, Brains and Programs*, Behavioral and Brain Sciences, 3: 417-457.

Searle, J. (**1992**) *The Rediscovery of the Mind*. MIT Press, Cambridge, MA.

Searle, J. R. (**2007**) *Biological naturalism*. In: Velmans, M. & Schneider, S. (Ed.), *The Blackwell Companion to Consciousness*, Blackwell, Malden, MA.

Sellars, W. (**1962**) *Philosophy and the Scientific Image of Man*. In: Colodny, R. (Ed.), *In Frontiers Of Science And Philosophy*, University of Pittsburgh Press, Pittsburgh, PA.

Shagrir, O. (**2005**) *The Rise and Fall of Computational Functionalism*. In: Ben-Menahem, Y. (Ed.), *Hilary Putnam*, Cambridge University Press, Cambridge, MA.

Shapiro, L. (**2000**) *Multiple Realizations*, Journal of Philosophy 97: 635-654.

Shermer, M. (**1993**) *The Unlikeliest Cult In History*, Skeptic 2: 74-81.

Shields, C. J. (**2007**) *Aristotle*. Routledge, London, UK.

Shoemaker, D. (**2005**) *Personal Identity and Ethics*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/entries/identity-ethics/".

Shoemaker, S. (**1980**) *Causality and Properties*. In: Inwagen, P. v. (Ed.), *Time and Cause*, Reidel, Dordrecht, NL.

Singer, P. (**1993**) *Practical Ethics*. Cambridge University Press, Cambridge, MA (2nd Edition).

Sloman, A. (**2004**) *The Irrelevance of Turing Machines to AI*. In: Scheutz, M. (Ed.), *Computationalism. New Directions*, MIT Press, Cambridge, MA.

Sloman, A. (**2008**) *Progress Report on the Cognition and Affect project. Architectures, Architecture-Schemas, and The New Science of Mind*, University of Birmingham.

Sloman, A. & Chrisley, R. (**2003**) *Virtual Machines and Consciousness*, Journal of Consciousness Studies 10: 113-172.

Smerlak, M. & Rovelli, C. (**2007**) *Relational EPR*, Foundations of Physics 37: 427-445.

Smith, B. C. (**1996**) *On the Origin of Objects*. MIT Press, Cambridge, MA.

Smith, M. (**2005**) *Meta-Ethics*. In: Jackson, F. & Smith, M. (Ed.), *The Oxford Handbook of Contemporary Philosophy*, Oxford University Press, Oxford, UK.

Smolin, L. (**1997**) *The Life of the Cosmos*. Oxford University Press, New York, NY.

Smullyan, R. M. (**1977**) *The Tao Is Silent*. HarperSanFrancisco, San Francisco, CA.

Snow, C. P. (**1959**) *The Two Cultures*. Cambridge University Press, Cambridge, UK (Canto 1998 Edition).

Sousa, R. d. (**2007**) *Why Think? The Evolution of the Rational Mind*. Oxford University Press, New York, NY.

Sperber, D. (**2000**) *An Objection to the Memetic Approach to Culture*. In: Aunger, R. (Ed.), *Darwinizing Culture. The Status of Memetics as a Science*, Oxford University Press, Oxford, UK.

Spinoza, B. d. (**1677**) *Ethics*.

Stahl, B. (**2008**) *Discourses On Information Ethics. The Claim To Universality*, Ethics and Information Technology 10: 97-108.

Steinhardt, P. J. & Turok, N. (**2002**) *A Cyclic Model of the Universe*, Science 296: 1436-1439.

Steinhart, E. (**2004**) *Pantheism and Current Ontology*, Religious Studies 40: 63–80.

Stoljar, D. (**2001**) *Two Conceptions of the Physical*, Philosophy and Phenomenological Research 62: 253-281.

Strawson, G. (**2001**) *The Bounds of Freedom*. In: Kane, R. (Ed.), *The Oxford Handbook of Free Will*, Oxford University Press, Oxford, UK.

Strawson, G. (**2006**) *Realistic Monism. Why Physicalism entails Panpsychism*, Journal of Consciousness Studies 13: 3-31.

Strawson, G. & Others (**2006**) *Consciousness and Its Place in Nature. Does Physicalism Entail Panpsychism?*. Imprint Academic, Exeter, UK.

Sturgeon, N. L. (**2006**) *Ethical Naturalism*. In: Copp, D. (Ed.), *The Oxford Handbook of Ethical Theory*, Oxford University Press, New York, NY.

Sukopp, T. & Vollmer, G. (**2007**) *Naturalismus. Positionen, Perspektiven, Probleme*. Mohr Siebeck, Tübingen.

Suppes, P. (**1960**) *A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences*, Synthese 12: 287-301.

Svozil, K. (**1996**) *Undecidability Everywhere?*. In: Casti, J. L. & Karlquist, A. (Ed.), *Boundaries and Barriers. On the Limits to Scientific Knowledge*, Addison-Wesley, Reading, MA.

Svozil, K. (**2005**) *Computational Universes*, Chaos, Solitons and Fractals 25: 845-859.

Svozil, K. (**2007**) *Physical Unknowables*, arXiv:physics/0701163v2 [physics.gen-ph].

Swinburne, R. (**2005**) *Faith and Reason*. Clarendon Press, Oxford, UK (2nd Edition).

Tamarin, G. R. (**1966**) *The Influence of Ethnic and Religious Prejudice on Moral Judgment*, New Outlook 9: 49-58.

Tamarin, G. R. (**1973**) *The Israeli Dilemma. Essays on a Warfare State*. Rotterdam University Press, Rotterdam, NL.

Tegmark, M. (**1996**) *Does the Universe in Fact contain almost no Information?*, Foundations of Physics Letters 9: 25-42.

Tegmark, M. (**1997**) *On the Dimensionality of Spacetime*, Classical and Quantum Gravity 14: L69-L75.

Tegmark, M. (**2000**) *The Importance of Quantum Decoherence in Brain Processes*, Physical Review E 61: 4194-4206.

Tegmark, M. (**2007**) *The Mathematical Universe*, arXiv:0704.0646v2.

Thagard, P. (**2002**) *The Passionate Scientist. Emotion in Scientific Cognition*. In: Carruthers, P.; Stich, S. & Siegal, M. (Ed.), *The Cognitive Basis of Science*, Cambridge University Press, New York, NY.

Thomson, G. (**2008**) *Counting Subjects*, Synthese 162: 373-384.

Tononi, G. (**2004**) *An Information Integration Theory of Consciousness*, BMC Neuroscience 5.

Tononi, G. (**2008**) *A Bit of Theory. Consciousness as Integrated Information*, URL="http://spectrum.ieee.org/computing/hardware/a-bit-of-theory-consciousness-as-integrated-information", Retrieved on 02.06.2008 (IEEE Spectrum, Special Report: The Singularity).

Trakakis, N. (**2007**) *Whither Morality in a Hard Determinist World?*, Sorites 19: 14-40.

Turchin, P. (**2003**) *Historical Dynamics. Why States Rise and Fall*. Princeton University Press, Princeton, NJ.

Turchin, P. (**2008**) *Arise 'Cliodynamics'*, Nature 454: 34-35.

Turing, A. M. (**1936**) *On Computable Numbers, with an Application to the Entscheidungsproblem*, Procedings of the London Mathematical Society 42: 230-265.

Turkle, S. (**2005**) *The Second Self. Computers and the Human Spirit*. MIT Press, Cambridge, MA (Twentieth Anniversary Edition).

Tversky, A. & Kahneman, D. (**1983**) *Extensional versus Intuitive Reasoning. The Conjunction Fallacy in Probability Judgment*, Psychological Review 90: 293-315.

Tye, M. (**2009**) *Consciousness Revisited. Materialism Without Phenomenal Concepts*. MIT Press, Cambridge, MA.

Vaidman, L. (**2002**) *Many-Worlds Interpretation of Quantum Mechanics*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/archives/sum2009/entries/qm-manyworlds/".

Vaidman, L. (**2008**) *Counterfactuals in Quantum Mechanics*, arXiv:0709.0340v1 [quant-ph].

Vidal, C. (**2007**) *An Enduring Philosophical Agenda. Worldview Construction as a Philosophical Method*, URL="http://philsci-archive.pitt.edu/archive/00003335/", Retrieved on 28.01.2008.

Vohs, K. & Schooler, J. (**2008**) *The Value of Believing in Free Will. Encouraging a Belief in Determinism Increases Cheating*, Psychological Science 19: 49–54.

Vollmer, G. (**1975**) *Evolutionäre Erkenntnistheorie*. Hirzel, Stuttgart (8th Edition).

Vollmer, G. (**2003**) *Wieso können wir die Welt erkennen? Neue Beiträge zur Wissenschaftstheorie*. Hirzel, Stuttgart.

Vriend, N. J. (**2000**) *An Illustration of the Essential Difference between Individual and Social Learning, and its Consequences for Computational Analyses*, Journal of Economic Dynamics and Control 24: 1-19.

Waldrop, M. M. (**1992**) *Complexity: The Emerging Science at the Edge of Order and Chaos*. Simon and Schuster, New York, NY.

Walter, A. (**2006**) *The Anti-Naturalistic Fallacy. Evolutionary Moral Psychology and the Insistence of Brute Facts*, Evolutionary Psychology 4: 33-48.

Walter, H. (**2001**) *Neurophilosophy of Free Will*. In: Kane, R. (Ed.), *The Oxford Handbook of Free Will*, Oxford University Press, Oxford, UK.

Watson, J. D. (**1969**) *The Double Helix*. New American Library, New York, NY.

Watts, A. (**1957**) *The Way of Zen*. Pantheon Books, New York, NY (Vintage Books 1989 Edition).

Wegner, D. M. (**2002**) *The Illusion of Conscious Will*. MIT Press, Cambridge, MA.

Weinberg, S. (**1977**) *The First Three Minutes. A Modern View of the Origin of the Universe*. Basic Books, New York, NY.

Whitfield, J. (**2007**) *Survival of the Likeliest*, PLoS Biology 5: e142.

Wigner, E. P. (**1960**) *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, Communications in Pure and Applied Mathematics 13: 1-14.

Wilkes, K. V. (**1988**) *Real People. Personal Identity Without Thought Experiments*. Clarendon Press, Oxford, UK.

Williams, B. (**1996**) *Toleration. An Impossible Virtue?*. In: Heyd, D. (Ed.), *Toleration: An Elusive Virtue*, Princeton University Press, Princeton, NJ.

Wilson, G. V. (**2006**) *Wo Klemmt Es Wirklich Bei Wissenschaftlichen Berechnungen?*, Spektrum der Wissenschaft 11: 118-120.

Wimmer, F. M. (**2007**) *Cultural Centrisms and Intercultural Polylogues in Philosophy*, International Review of Information 7.

Wimsatt, W. C. (**1994**) *The Ontology of Complex Systems. Levels of Organization, Perspectives, and Causal Thickets*, Canadian Journal of Philosophy 20: 207-274.

Witherall (**2006**) *The Zero Ontology*, URL="http://www.hedweb.com/witherall/zero.htm", Retrieved on 05.03.2009.

Wolpert, L. & Richards, A. (**1997**) *Passionate Minds. The Inner World of Scientists*. Oxford University Press, Oxford, UK.

Worrall, J. (**1989**) *Structural Realism. The Best of Both Worlds?*, Dialectica 43: 99-124.

Worrall, J. (**2004**) *Why Science Discredits Religion*. In: Peterson, M. & Vanarragon, R. (Ed.), *Contemporary Debates in Philosophy of Religion*, Blackwell Publishing, Malden, MA.

Wright, R. (**1995**) *The Moral Animal. Why We Are, the Way We Are. The New Science of Evolutionary Psychology*. Pantheon Books, New York, NY.

Yalowitz, S. (**2004**) *Anomalous Monism*, Stanford Encyclopedia of Philosophy Summer 2009 Edition URL="http://plato.stanford.edu/entries/anomalous-monism/#tafam".

Yamada, K. (**2004**) *Mumonkan. Die torlose Schranke. Zen-Meister Mumons Koan-Sammlung*. Kösel, München.

Yudkowsky, E. (**2007c**) *Zen and the Art of Rationality*, URL="http://lesswrong.com/lw/m7/zen_and_the_art_of_rationality/", Retrieved on 20.06.2009.

Yudkowsky, E. (**2007d**) *Cached Thoughts*, URL="http://lesswrong.com/lw/k5/cached_thoughts/", Retrieved on 18.06.2009.

Yudkowsky, E. (**2007e**) *Universal Law*, URL="http://lesswrong.com/lw/hr/universal_law/", Retrieved on 2009.08.14.

Yudkowsky, E. (**2007f**) *Semantic Stopsigns*, URL="http://lesswrong.com/lw/it/semantic_stopsigns/", Retrieved on 07.07.2009.

Yudkowsky, E. (**2007g**) *An Alien God*, URL="http://lesswrong.com/lw/kr/an_alien_god/", Retrieved on 2009.08.27.

Yudkowsky, E. (**2007h**) *Thou Art Godshatter*, URL="http://www.overcomingbias.com/2007/11/thou-art-godsha.html", Retrieved on 2009.08.27.

Yudkowsky, E. (**2007i**) *Not for the Sake of Happiness (Alone)*, URL="http://lesswrong.com/lw/lb/not_for_the_sake_of_happiness_alone/", Retrieved on 27.07.2009.

Yudkowsky, E. (**2007a**) *Religion's Claim to be Non-Disprovable*, URL="http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/", Retrieved on 20.06.2009.

Yudkowsky, E. (**2007b**) *Expecting Short Inferential Distances*, URL="http://lesswrong.com/lw/kg/expecting_short_inferential_distances/", Retrieved on 2009.08.10.

Yudkowsky, E. (**2008a**) *Newcomb's Problem and Regret of Rationality*, URL="http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/", Retrieved on 18.06.2009.

Yudkowsky, E. (**2008b**) *How An Algorithm Feels From Inside*, URL="http://lesswrong.com/lw/no/how_an_algorithm_feels_from_inside/", Retrieved on .

Yudkowsky, E. (**2008c**) *Extensions and Intensions*, URL="http://lesswrong.com/lw/nh/extensions_and_intensions/", Retrieved on 20.06.2009.

Yudkowsky, E. (**2008d**) *Probability is Subjectively Objective*, URL="http://lesswrong.com/lw/s6/probability_is_subjectively_objective/", Retrieved on 18.06.2009.

Yudkowsky, E. (**2008e**) *The Cluster Structure of Thingspace*, URL="http://lesswrong.com/lw/nl/the_cluster_structure_of_thingspace/", Retrieved on 20.06.2009.

Yudkowsky, E. (**2008f**) *Superexponential Conceptspace, and Simple Words*, URL="http://lesswrong.com/lw/o3/superexponential_conceptspace_and_simple_words/", Retrieved on 20.06.2009.

Yudkowsky, E. (**2008g**) *Living in Many Worlds*, URL="http://lesswrong.com/lw/qz/living_in_many_worlds/", Retrieved on 23.07.2009.

Yudkowsky, E. (**2008h**) *Possibility and Could-ness*, URL="http://lesswrong.com/lw/rb/possibility_and_couldness/", Retrieved on 01.07.2009.

Yudkowsky, E. (**2008i**) *Prolegomena to a Theory of Fun*, URL="http://lesswrong.com/lw/wv/prolegomena_to_a_theory_of_fun/", Retrieved on 27.07.2009.

Yudkowsky, E. (**2008j**) *Complex Novelty*, URL="http://lesswrong.com/lw/wx/complex_novelty/", Retrieved on 27.07.2009.

Yudkowsky, E. (**2009**) *31 Laws of Fun*, URL="http://lesswrong.com/lw/y0/31_laws_of_fun/", Retrieved on 27.07.2009.

Zimmer, H. R. & Campbell, J. (**1969**) *Philosophies of India*. Princeton University Press, Princeton, NJ.

Zumdahl, S. (**2007**) *Chemistry. Media Enhanced Edition*. Houghton Mifflin, Boston, MA (7th Edition).

Zurek, W. H. (**2007**) *Relative States and the Environment. Einselection, Envariance, Quantum Darwinism, and the Existential Interpretation*, arXiv:0707.2832v1 [quant-ph].