universität wien

Masterarbeit

Titel der Masterarbeit

Inference to the Best Explanation as a Method for Solving the Hard Problem of Consciousness

Verfasser

Lukas Schwengerer

angestrebter akademischer Grad

Master of Arts (MA)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A066 941
Studienrichtung lt. Studienblatt:	Masterstudium Philosophie
Betreuer:	Mag. Dr. habil. Sven Bernecker, PhD

universität wien

Master Thesis

Title

Inference to the Best Explanation as a Method for Solving the Hard Problem of Consciousness

Author

Lukas Schwengerer

submitted in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Vienna, 2013

Program ID number:	A066 941
Program:	Masterstudium Philosophie
Advisor:	Mag. Dr. habil. Sven Bernecker, PhD

Contents

Introduction 6

1 Setting Up the Project 8

1.1 A Guiding Picture 8

1.2 The Mysterian Challenge 9

1.2.1 Consciousness 9

1.2.2 The Hard Problem of Consciousness 12

1.2.3 Explanation I – Deductive-Nomological 13

1.2.4 Explanation II – Causal 17

1.2.5 Stating the Mysterian Challenge 24

1.2.6 Mysterian Denial and Epistemic Limits 25

1.3 An Answer in Sight 34

2 The Method 37

2.1 The Pool of Explanation Candidates 37

2.1.1 Example I – Galen Strawson's Fundamental Duality Monism 39

2.1.2 Example II – 'Nonreductive' Physicalism(s) 41

2.1.3 Example III – An Interactive Dualism à la Descartes 45

2.1.4 Example IV – A Type Identity Theory 46

2.2 The Filter: Consistency and Coherence 47

2.2.1 Starting Steps to Coherence 47

2.2.2 A Computable Account of Coherence 51

2.2.3 Constraints 56

2.2.4 Accepted Science-Based Propositions 59

2.2.5 Accepted Intuition-Based Propositions I 61

2.2.6 Accepted Intuition-Based Propositions II – Philosophical Agreement 63

2.2.7 Accepted Intuition-Based Propositions III – Experimental Philosophy 67

2.3 The Filter: Exemplifying the Usage 69
2.3.1 Interactive Dualism (Descartes) 70
2.3.2 Type Identity Theory 72
2.3.3 Supervenience Physicalism 73
2.3.4 Supervenience Physicalism with Ability Hypothesis 75
2.3.5 Supervenience Physicalism with Ability Hypothesis and Conceivability (Intuition) Production 77
2.3.6 Fundamental Duality Monism 80
2.4 What we achieved: Improving the Filter, 'Supertheories' and Inference to the Loveliest Explanation 82
3 Epistemic Status 86
3.1 Justification as Having Reasons 86
3.2 Justification as Reliability 86
3.3 Removed Defeaters 89
3.4 Truth 89
4 Conclusion 90
5 Bibliography 92
6 Appendix 101
Abstract 101
Zusammenfassung 102
Curriculum Vitae 103
Lebenslauf 105

Acknowledgements

I would like to express my deep gratitude to my advisor Dr. Sven Bernecker, who continued to provide valuable input after moving back to the University of California, Irvine. My special thanks are extended to the members of the Vienna Forum for Analytic Philosophy and the 'Sperl' reading circle for various discussions and ideas. Finally, I wish to thank my parents for support and encouragement throughout my study, and my brother for a variety of lively and fun debates.

Introduction

We are forever explaining happenings in the world. We explain why it can be freezing cold in March, or why a stick seems to be bent when it is held into water. Most of our everyday explanations come surprisingly easy to us. There is never any doubt about us being able to know why a billiard ball moves when hit by another. Of course, the precise calculations necessary to predict a billiard balls direction are not obvious. Nevertheless, we feel that we can explain its movement just fine. Sometimes, however, we stumble upon a phenomenon that seems to be slipping completely out of our hands. We are perplexed how the phenomenon in question is even possible, but are certain that it is reality. Consciousness is one of those. There is barely anything that we are more certain about, than the existence of a pain in the moment we experience it. Unfortunately, there is also barely anything we find ourselves in such great difficulties explaining. The debate in philosophy over consciousness in various aspects is wide and diverse, but not a single theory has been able to convince the scientific community of being the correct, or at least adequate explanation for the problem of phenomenal experience, e.g. in feeling pain. This very fact motivated skeptics, like Colin McGinn, to claim that we might be even in principle unable to provide such an explanation. This leads to the basic question we will be dealing in this inquiry: What should a method of choice for finding an explanation for the existence of phenomenal consciousness look like? The presentation of our answer will be provided as follows:

Chapter 1 sets up the project by giving an analogy that leads to the basic idea of the method. Then we define the problem of consciousness in question and the notion of explanation. We will also spell out the 'mysterian challenge' that the inference to the best explanation should be able to answer.

Chapter 2 provides an account of such a method by showing the tools used to rule out inadequate explanation candidates from a pool of theories for phenomenal consciousness. It exemplifies the method by testing several explanation candidates: Interactive Dualism; Type Identity Theory, Supervenience Physicalism; Supervenience Physicalism with Ability

Hypothesis; Supervenience Physicalism with Ability Hypothesis and clauses for cases of mistaken conceivability; and Fundamental Duality Monism.

Chapter 3 embarks on a short epistemological evaluation of the method's success based on our definition of explanation in Chapter 1.

In Chapter 4 we conclude our results and reflect upon future work.

1 Setting Up the Project

1.1 A Guiding Picture

Suppose there is a species of intelligent slugs.[1] The slugs live on a mosaic completely constructed by two sorts of tiles: triangles and pieces of pie. The mosaic itself can come in and has parts of different shapes - circles, rectangles, and so on. The slugs' perceptual access to the mosaic is limited to two shape-detecting systems. One detects triangles and the other circles. The slugs build their natural sciences on the basis of the triangle detection system. These sciences work surprisingly well in the mosaic world. Some members of the slug society are philosophers – *slugists* – who question what there really is in the mosaic world. Some of these slugists claim that the circles are not fundamental, but rather based on the triangles. Others object that the triangle-slugists leave something out. There is no place for circles in a triangle world. Circles have to be something totally independent and fundamental. A third group thinks that there is some truth in the circle being built by other tiles and suggest that there are parts in the world that are close siblings to normal triangles, but are round on one side. They postulate pieces of pie in the mosaic world. Within the slugists a huge debate arises about which solution to the circle problem is correct and whether the slugs can ever know which one is correct. It turns out that slug mathematicians prove that triangles can never build up a circle. At the same time the slug scientists believe that their science works very well because of their triangles having three corners. Everything in the slug science can be predicted based on three corners. Now the slugists infer that the best explanation for circles is the pieces of pie explanation, because triangles are ruled out and the circle fundamentalists do not fit well with the slug sciences. The slugists get things right. They know what fundamental parts there are in their mosaic world.

What is the point of this picture? The slugs cannot grasp all fundamental parts of the mosaic by perception but nevertheless they solve the circle problem. They hypothesize what might exist on a basic level that brings about the circle. They invoke new concepts and test their hypotheses by confronting it with their well working sciences. One hypothesis stands out because it fits best with the scientific knowledge and it happens to be true.

[1] The slugs are loosely based on Jackson (1982) and the tiles on the modified version by Stoljar (2006). The story differs from their stories though.

It might be the same case with consciousness. As the slugs could not, we also cannot grasp directly what brings about consciousness. Nevertheless we can hypothesize what might be the cause of experience and we can compare different hypothesis. We rate how they fit with our sciences and intuitions and maybe we get things right as well, solving the problem of consciousness. However, is this even possible?

1.2 The Mysterian Challenge

What we will call the 'm*ysterian challenge'*[2] is more or less the denial of the slug's story for consciousness. The so-called mysterians deny the possibility of an explanation of consciousness completely. There is no direct way to reach an explanation by introspection or just looking at the physics of the brain and, secondly, there is also no possibility to infer one. Consciousness will be mysterious forever and we better not waste our time in futile explanation attempts. The aim of this project is to accept the first part, the denial of directly reaching an explanation of consciousness, but to refute the second part. There might be a possible workaround to explain consciousness indirectly, which we need to explore to full extent. However, before we embark on the challenge in a more philosophical vocabulary, let us shortly consider some initial considerations and definitions.

1.2.1 Consciousness

When talking about consciousness in the guiding picture section, we have been rather vague. The notion, however, needs to be treated cautiously and in a more precise way. There are several ways of defining different forms of consciousness. Christopher Hill (2009), for example, distinguishes between agent-, propositional-, introspective-, experential-, access-, relational-, and phenomenal consciousness, with some of them being reducible to others.[3] *Agent consciousness* is what we have in mind when using sentences like 'She is regaining/losing consciousness'. *Propositional consciousness* is based on our use of the phrase 'conscious that…'. *Introspective consciousness* is what we usually relate to some tendencies being conscious, while others are unconscious. *Experiential consciousness* is a

[2] The name is loosely based on Owen Flanagan's baptism of the *new mysterians*. See Flanagan (1992).
[3] The Access- and Phenomenal Consciousness distinction here is taken from Ned Block (1997).

label for experiences of cognitive (e.g. perception) and conative (e.g. volitions) mental states and propositional attitudes. *Access consciousness* denotes states that are "[…] poised for direct control of thought and action" (Block, 1997, p. 382). Those are "[…] poised for free use in reasoning and for direct 'rational' control of action and speech" (Block, 1997, p. 382). *Relational consciousness* is connected to the expression 'conscious of' - when we express our perceiving of an object. *Phenomenal consciousness* is what is usually understood as the possession of qualia. Qualia are subjective, directly experienced qualities that are given to the experiencer. Experiencing a quale has a certain way it is like to[4] experience it. For example: Being in pain has a subjective quality for the person experiencing it. The notion of qualia will be further refined later.

Note that Hill has one of the richer taxonomies of consciousness. Others like Galen Strawson (2010) tend to speak only about 'experience', without any difference in experience and the phenomenal. The decision whether those two terms are interchangeable is mostly based on the view on cognitive phenomenology. Proponents of cognitive phenomenology argue that various cognitive processes come with a certain way it is like to experience them.[5] Phenomenology, so they argue, is not limited to perceptual states. Understanding a sentence has it (Strawson, 2010, pp. 5-13; Goff, 2012), as well as having a belief (Goldman, 1993), a desire or thoughts in general too. The amount of mental states that are accompanied with a certain way they feel like varies on different accounts. The larger the set of mental states with phenomenology, the more the distinction between the experienced and the phenomenal vanishes. We will leave the question of cognitive phenomenology open. It is, however, important to keep in mind that the contested extension of phenomenal consciousness is a problem that needs to be solved. It will come up on various occasions throughout this inquiry.

For this project we consider only phenomenal consciousness and therefore use 'consciousness' interchangeable with 'phenomenal consciousness'. We do the same with 'being conscious' and similar terms. A general difficulty in defining phenomenal consciousness and qualia is the, at least seemingly, inevitable circularity. As described by

[4] Taking the phrase from Nagel (1974).
[5] Cf. Bayne & Montague (2011)

Mark Rowlands (2003), three different ways to convince one of the existence of qualia are used in general. All of them involve qualia already as a part of the argument. The first option is to give examples and hope that the reader, who experienced something similar to the examples already, remembers how it felt like. The second option is to provide rough synonyms for qualia: 'experiential', 'qualitiative character', 'feel', 'subjective', to name some of the most common. The third option is to invite the reader to create a phenomenal, mental state. One could, for example, encourage her to bite into a lemon. There will certainly be something it is like to experience that. We invoked a little bit of the first two options (and got a glimpse of the third a moment ago). The question whether there are qualia will not be pursuit by us. We want to explain why there is phenomenal consciousness, presupposing that there are qualia. We do not argue for the existence of qualia. Some arguments and thought experiments, which come up later in this work, can potentially be used to do so, but we use them for a different purpose. In general, it is rather unlikely to convince an eliminativist of the existence of qualia and a waste of space to recapitulate old arguments for an already pro-qualia audience. We do not provide any original arguments which might do better. Denying the existence of something as basic as qualia seems very hard to grasp. As Galen Strawson talking about realistic physicalism acknowledged:

> "Well, one thing is absolutely clear. You're certainly not a realistic physicalist, you're not a real physicalist, if you deny the existence of the phenomenon whose existence is more certain than the existence of anything else: experience, 'consciousness', conscious experience, 'phenomenology', experiential 'what-it's-likeness', feeling, sensation, explicit conscious thought as we have it and know it at almost every waking moment." (Strawson, 2006b, p. 3)

This, however, has no impact on our pursuit for an answer to the mysterian challenge without denying the mysterians reasoning completely. The mysterians do accept the presupposition of qualia and eliminative materialists do not feel threatened by the mysterian challenge anyway.

We continue with a characterisation of the main problem of consciousness by David Chalmers to further elucidate on phenomenal consciousness. This leads into a more concrete formulation of our research goal.

1.2.2 The Hard Problem of Consciousness

We rely on a basic distinction made popular by David Chalmers (1995). He provides us with two different kinds of problems of consciousness. The first one – the 'easy problems' – are problems of information processing. Those include, but are not limited to, "the ability to discriminate, categorize, and react to environmental stimuli; the integration of information by a cognitive system; the reportability of mental states; the ability of a system to access its own internal states; the focus of attention; the deliberate control of behaviour; the difference between wakefulness and sleep" (Chalmers, 1995, p. 200). While these are certainly not fully solved and answered, there is, in principle, no problem about how we should be able to solve them. The hard problem of consciousness is different. It is not about the complex information processing, but rather about the subjective aspect it is like to be in, when the information processing happens.

> "This subjective aspect is experience. When we see, for example, we experience visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations, from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion, and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience. It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does." (Chalmers, 1995, p. 201)

The first point to note is that Chalmers uses the term 'experience' for all phenomenal consciousness, whether perceptual states, emotions or mental images. There is something it

is like to be in any of these states. The question that arises is: Why and how? It is indeed perplexing that something like phenomenal consciousness exists. It certainly can be perplexing in various ways. We may, for example, wonder what the biological or evolutionary function of consciousness is. Our curiosity is different and starts with how little of an idea of its fit into the usual scientific worldview we have. We do have good reasons to believe that it is somehow connected to the brain, which does the processing. Nevertheless there seems to be an explanatory gap between the brain processes and phenomenal consciousness. (Levine, 1983) Chalmers therefore asks rightfully: "Why is all this processing accompanied by an experienced inner life?" (Chalmers, 1996, p. xii) We want to have an explanation for the existence of phenomenal consciousness, i.e.: what brings about qualia? We need an explanation for these qualia that accompany so many of our cognitive processes.

The hard problem poses the primary question for this thesis. Like the slugists wants to find out why there are circles, we now want to find out where phenomenal consciousness comes from. In other words: *We want to explain the existence of phenomenal consciousness. We want to explain qualia.*

1.2.3 Explanation I – Deductive-Nomological

As claimed above we want an explanation of phenomenal consciousness. But what exactly is an 'explanation'? In general, an explanation should be an answer to a question why an event took place or will take place. If we want to explain why the billiard ball moved, we point out the other ball hitting it with a certain speed. If we want to explain why a person eats all the cookies of a plate in front of him, we can provide reasons for her to do so. She may be hungry. Examples like these can be found innumerably in our everyday life and we are inclined to identify them as explanations. We need, however, a general account of the notion to get a grasp on a possible answer to the hard problem of consciousness. Such an account will not lead to a solution itself, but it restrains the range of possibly valid answers. Accounts of explanation vary and we will limit ourselves to a discussion of two particular models with a decision for one of them. We will not discuss other accounts like the reason model, the familiarity model, the unification model or the necessity model. For a short overview of those see Lipton (2004, pp. 21-29).

In the philosophy of science of the present and last century the notion of explanation is dominated by versions of the deductive-nomological account[6]. This model was made popular by Carl G. Hempel and Paul Oppenheim beginning in the 1940s[7]. According to this account, the basic structure of any explanation is constituted by an explanandum and an explanans. The explanandum is a description of what is to be explained, whereas the explanans denotes a class of sentences that adduce an account for the phenomenon. (Hempel & Oppenheim, 1948, pp. 136-137) The relation between explanandum and explanans is analogue to the relation of premises and conclusion in a deductive argument (hence the *deductive*-nomological account). The explanandum is further distinguished into two subclasses: Antecedent conditions (C_1, C_2, ...C_n); and sentences representing general laws (L_1, L_2, …L_n). Every single explanation has both antecedent conditions and general laws as part of the explanandum. There are no exceptions. Let us go over a sample of a deductive-nomological explanation provided by Hempel:

> "The metal screwtop on a glass jar is tightly stuck; after being placed in warm water for a short while, it can be readily removed. The familiar explanation of this phenomenon is, briefly, to the effect that the metal has a higher coefficient of thermal expansion than glass, so that a given rise in temperature will produce a larger expansion of the lid than of the neck of the glass jar; and that, in addition, though the metal is a good conductor of heat, the temperature of the lid will temporarily be higher than that of the glass—a fact which further increases the difference between the two perimeters. Thus, the loosening of the lid is here explained by showing that it came about, by virtue of certain antecedent circumstances, in accordance with certain physical laws. The explanation may be construed as an argument in which the occurrence of the event in question is inferred from information expressed by statements of two kinds: (a) general laws, such as those concerning the thermal conductivity of metal and the coefficients of expansion for metal and for glass, as well as the law that heat will be transferred from one body to another of lower temperature with which it is in contact; (b) statements describing particular circumstances, such as that the jar is made of glass, the lid of metal; that initially, at room temperature, the lid fitted very tightly

[6] Also known as the covering-law model of explanation.
[7] Cf. (Hempel, 1942; Hempel & Oppenheim, 1948)

> on the top of the jar; and that then the top with the lid on it was immersed in hot water. To show that the loosening of the lid occurred "by virtue of" the circumstances in question, and "in accordance with" those laws, is then to show that the statement describing the result can be validly inferred from the specified set of premises." (Hempel, 2001a (1962), pp. 87-88)

We can write any such explanation like a deductive argument:

$(C_1, C_2, ...C_n)$	(The jar is made of glass.; The lid is made of metal; The lid is tightly fit at room temperature.;…)
$(L_1, L_2, …L_n)$	(<Laws of thermal conductivity of metal and glass>; <Laws of heat transfer>;…)
Explanandum	The metal screwtop on a glass jar was tightly stuck, but after being placed in warm water for a while, it can be removed.

Now we obviously do not explain with complete arguments in everyday life and even in science we rarely do. Does that imply a problem of our everyday use of 'explanation' or is the deductive-nomological model simply too strong? One possible answer is to relativize our everyday explanations. They may be either incomplete, or be no explanations at all. In these cases they may indicate some connection, but do not really explain. Something is a real explanation only if there is nothing missing to the deduction. In other cases the whole argumentative structure might be tacit. (Hempel & Oppenheim, 1948, p. 139) The billiard ball moving can be sufficiently explained by being hit by the other ball, if it is crystal clear to everyone, that the momentum-transfer is a result of certain obvious physical laws. The single sentence explanation 'It was hit by the other ball' serves as an enthymeme in such cases. It is, however, a very unpleasant consequence of the model, that even in science we either have incomplete or enthymematic explanations in most cases. There is a strong pull in the intuition that the ordinary use of 'explanation' is correct and the deductive-nomological account is simply too demanding.

Another problem comes up with the demand of general laws as part of the explanans. It is anything but clear what differentiates real laws of nature from simple generalisation (Woodward, 2003, pp. 182-184). Hempel characterizes general laws already in 1942:

> "By a general law, we shall here understand a statement of universal conditional form which is capable of being confirmed or disconfirmed by suitable empirical findings. […] In every case where an event of a specified kind C occurs at a certain place and time, an event of a specified kind E will occur at a place and time which is related in a specified manner to the place and time of the occurrence of the first event." (Hempel, 1942, p. 35)

This, however, does not provide a sufficient way to distinguish laws from generalisations. We might try to define a difference by using modality. Laws are necessary, whereas ordinary generalisations only hold contingently. An unpleasant bullet to bite with this option is the commitment to necessary laws of nature with its prima facie implausibility. We better have good reasons to deny possible worlds with other physical laws, before we choose to vouch for this option. Problematic in a similar way is the definition by Hempel regarding other types of laws, namely probabilistic laws. How should we characterize cases in which not every event of type C is followed by an event of type E, but only in 76% of them? Hempel (2001a (1962); 2001b (1963); 2001c (1977)) himself and others (Railton, 1978) (Fetzer, 1974) tried to reconcile probabilistic explanation with his covering-law model, but we will not go into this debate.

The deductive-nomological account works with general laws of causation[8], but it is not limited to those (even though the examples usually are). This openness for different forms of laws used in an explanation is contested by a causal view of explanation. David Lewis (1986) argues that all cases of supposed non-causal explanation in the natural sciences are either causal, just stated in a less obvious way, or no explanations at all. Lewis' account differs not only in restricting explanation to cases of causality, he also does not commit to an argument-like schema.

[8] The „C" and „E" in the last quote suggest the terms 'cause' and 'effect', as Hempel (1942, p. 35) acknowledges.

1.2.4 Explanation II – Causal

Lewis main thesis about explanation is this: "*to explain an event is to provide some information about its causal history*" (Lewis, 1986, p. 217).[9] The core intuition guiding Lewis concerns a conversation situation in which someone, who possesses some information about the causal history of an event, discloses this information to someone else. In this process no complete argument-like structure is needed. It is not required that she has complete information about the causal history.[10] It is sufficient for her to have only some parts of information. Explanations can vary in scale. A physicist might know more about what causes a window to brake, when a ball hits it, than the soccer player knows. Nevertheless, both are able to give valid explanations, even with the physicist's being *better* or *richer* in a sense. This certainly does not imply that deductive-nomological explantions are no explanations at all, they just come by very rarely.

We can easily redefine Lewis' definition for an explanation of an event to general explanations and explanations of regularities: If we examine events of a single kind and their causal history involves events of a single kind, we can deduce a general explanation: *Information about what is common to all those causal histories.* (Lewis, 1986, p. 225) Granted, there is a difficulty involved in using kinds to provide general explanations. This does not seem insurmountable though. Paradigmatic cases of kinds in our everyday life seem to pose no problems to us. In addition, we did not commit ourselves to any view on natural kinds. We are free to define certain kinds directly for our explanation's purposes.

If the explanandum is a causal regularity, we can explain it by providing information about the specific mechanism that links cause and effect. (Lipton, 2004, p. 30) If we explain why extended sunbathing gets one sunburned, we cannot simply name the sun as the cause. We rather have to give information about the causal mechanism that leads from certain wavelengths of the light to a specific effect on the skin.

[9] The term 'information' is used in a loose fashion.
[10] This would be unattainable anyway.

There are three basic objections to the causal account. The first one is already mentioned before: There are explanations that do not include causation at all. Mathematical or geometrical explanations are independent of any causal relations. Some of them seem even to be suited to explain physical events. (Lipton, 2004, p. 31) Lewis would probably deny the last sentence and argue that there are no real explanations given in those cases. Nevertheless, he would have to accept explanation in the realm of math being non-causal. We better accept that the causal account is not the sole way of understanding 'explanation'. As long as the exceptions are only in the abstract realm of math and logic, however, we do not need to be concerned. Our task revolves around an explanation of phenomenal consciousness, which surely is part of the concrete world.

The second objection points to an unanalysed term 'causal'. It is certainly true that we do not have a fully adequate analysis of causation. Nevertheless, we do know a fair share of it by ordinary life, sciences and philosophy. We are also not required to provide a complete analysis to be able to use a concept to explain other things. (Lipton, 2004, p. 31) In fact, the vagueness of causality is a good thing for our project. If we want to find an indirect method to explain phenomenal consciousness, we better not restrict the notion of causality too much beforehand. Otherwise we would prematurely and unreasonably remove explanation candidates. For example: If we define causation as only possible between physical events, any dualist explanation would be ruled out. The accepted vagueness of causation prevents us from doing so.

The third objection is the claim that the causal model is too weak and therefore does not fit with our explanatory practices. We already noted that causal explanations show up in a variety of scales. Some information about a causal history is very detailed and extended; others are just tiny bits of intelligence. They can provide very useful explanations, or they might be of no use at all. If we want to explain the blue colour of the sky, it does not help us much if we shout "the big bang is part of its causal history". A causal explanation account therefore needs to distinguish between different explanations based on our interest in it. Some explanations are useful, others are not. Our explanatory interest determines which one we choose from all possible explanations. This explanatory interest can be found in the way we ask questions. If we ask anyone to explain why there is a bottle on the table and she answers with citing the big bang as a cause, we try to refine our question. We may ask why

there is a bottle on the table right now, when a minute ago there was not. This familiar method gives us a glimpse of how the third objection can be answered.

With the objections sufficiently answered, we can conclude that the causal explanation account is the one to prefer. The model relates well to our everyday notion of explanation, while being able to provide elegant ideas how particular events, general kind of events and regularities can be similarly explained. We will therefore use the causal model for the rest of this thesis. We can put forth a first part of our definition of explanation:

Def: Any real explanation has to be useful. We will henceforth use 'explanation' only for those explanation candidates that fit our explanation interest. Fitting our explanation interest is a necessary, but not solely sufficient condition.

Def: An explanation candidate[11] will be any given (supposed) information of the causal history of the explanandum.

In the definition of an explanation candidate we already included a 'supposed' in brackets. The reason for this clause is the relation of explanation to truth. The content of an explanation candidate might be false, whereas a real explanation has to be true.

Def: An explanation has to be true.

With explanations having truth conditions we have to ask if they also have general knowledge conditions. Does an explanation need to be justified and believed? Lipton thinks it does, when he writes: "Understanding is not some sort of super-knowledge, but simply more knowledge: knowledge of causes." (Lipton, 2004, p. 30) Understanding is here to be read simply as possessing an explanation. But is there a possibility of an explanation neither being the content of a belief, nor justified? There are reasons to think there is not: Imagine

[11] Sometimes this is labelled a 'potential explanation'. Cf. (Lipton, 2004, p. 58)

Gregg and John discussing why redshift occurs. Gregg is a physics expert and points to the movement of the objects emitting the light and the Doppler-Effect. John knows about Gregg's expertise and has no reason to doubt Gregg. We are inclined to grant, that Gregg explained the redshift to John. Now imagine another situation with Tim and John discussing the same topic. Tim has absolutely no clue about physics, but nevertheless claims to know why the redshift happens to exist. By sheer luck he correctly postulates the light emitting objects as moving away from the observer and cites the Doppler-Effect. If Tim and John were to discuss any other problem, for example blueshift, Tim would also answer in the exact same way and thereby claiming something false. Tim is therefore highly unreliable. John knows that pattern about Tim and does not acquire the belief corresponding to Tim's testimony. John thinks Tim rambles random things, which better not be really listened to. Did Tim explain redshift to John? Certainly not. With the testimony given the epistemic situation of John did not change at all. Before he talked to Tim, John did not have any information about the cause. Afterwards he still does not, because whatever Tim told him was not identified as information regarding redshift. John will only identify testimonial information as such, if it provides evidence for a belief explaining redshift and therefore leads him to acquire or strengthen the belief in question. The testimonial assertion is information regarding the explanandum only if it is adequately processed. It is only information if it makes a difference.[12]

But what about a case like the Gregg and John one without belief acquisition? Gregg again asserts the correct answer to the question why redshift occurs and he is again reliable. John, however, does not think Gregg gives him any information about the causal mechanism behind redshift. Did Gregg explain redshift? There is a sense in which we are inclined to answer negatively. John's information base did not change at all because he did not identify Gregg's testimony as information. On the other hand: from a third person point of view we do feel a pull for the intuition that Gregg *did* explain it because he provided something *we* identify as information. Only John failed to do so. That is: We do have the belief of the testimonial assertion being information regarding the explanandum. Nonetheless: Providing information and explaining are different. Even from a third person standpoint, Gregg only provided information, but he did not explain to John why redshift occurs.

[12] Cf. Bateson (2000 (1970))

There might also be cases in which the assertion is taken to be potential information, but defeaters exist, which deny this status. Imagine again the Tim and John example, but this time John concedes that Tim talks specifically about Redshift and does not only ramble about random stuff. John believes that Tim is reliable. However, it is widely known in John's social environment that Tim is still unreliable. John ought to know about that unreliability, which provides a 'common knowledge'[13] defeater. In this case Tim provides evidence in some sense, but John should better not let that evidence influence his beliefs. If we ask John, if Tim explained redshift, John will probably answer in the affirmative. Nevertheless we should not concede that Tim explained anything here. Even though Tim's testimony is taken to be information by John, the existence of a normative defeater prohibits the information status on our third person point of view. The truth of Tim's testimony was purely luck-based, but an explanation does not fit well with epistemic luck. An explanation ought to be something provided by someone knowing, or at least justifiably believing (parts of) the causal history or mechanism, not by a lucky guess.[14] If we were told that a testimonial assertion, taken to be an explanation, was just a lucky guess, we stop labelling it explanation, even though it might still be true. An explanation therefore does not only need to be taken as information, it has to be undefeated.

Let us assume a similar case with just the common knowledge defeater missing. John thinks Tim provides information and there are no undefeated defeaters to be found. Does Tim explain, even though he is unreliable? An internalist's answer will be affirmative. An externalist's might be negative. The letter will argue that it is not enough for John to believe in Tim's reliability, he really has to be reliable to provide information. Regardless of the internalism and externalism debate on justification, the relation of an explanation to knowledge can be treated equally. Explanation does require knowledge.

Any assertion only explains if

1) it is believed,
2) it is justified,

[13] A kind of normative defeater.

[14] This does not necessarily imply that the testifier really knows. We are only committed to the claim that whoever acquires knowledge through testimony has to treat the testifier as trustworthy and has no defeaters. Whether the testifier really has to know is still under debate. For a view of testimony as a generative source see (Lackey, 2008).

3) it is true,
4) it is undefeated,
5) it is not based on epistemic luck.

Def: An explanation has to be known.

Therefore explanation and understanding is "[…] simply more knowledge: knowledge of causes." (Lipton, 2004, p. 30)

One might object that we discussed situations of person A explaining something to person B, but what about 'having an explanation'? Having an explanation can easily be modelled based on our discussion. We have an explanation of redshift if we possess knowledge that is information (on first and third person point of view) on redshifts causal history.

We get our account of 'explanation':

An explanation candidate will be any given (supposed) information of the causal history of the explanandum.

Def: An explanation is an explanation candidate that:

1) fits the explanation interest,
2) is believed,
3) is justified,
4) is true,
5) is undefeated,
6) is not based on epistemic luck.

There are two different classes of explanations in general: Basic (or plain) explanations and contrastive explanations. So far we only talked about basic explanations like "why does x occur?", but there can be question similar to "why x rather than y?" as well. Those questions need to be answered by a contrastive explanation. In some cases contrastive explanations cannot be given at all. Let us assume a Copenhagen interpretation of the double-slit experiment and ask why a single electron is detected exactly at position x at time t. This is

not precisely determined, but rather only given a probability. Why was it spotted at x rather than somewhere else in the detector? Contrastive explanations for chance events cannot be given. (Lewis, 1986, p. 230) Of course some events may look like chance events but we can give contrastive explanations perfectly fine and some hidden variable theories think this may be even the case for the double-slit experiment in the future.[15]

Contrastive explanation cannot always be reduced to non-contrastive ones. "Why P rather than Q" is distinct from "why P" and "why not Q". If we want to answer the contrastive question, information about the causal history of P will not satisfy us, except when it accidentally happens to be part of the causal history of not-Q too. P and Q may both be true. Assume, for example, a thermometer with mercury. If we want to explain why the mercury rose in the glass, rather than breaking the glass, our causal story of the mercury rising will not be enough, because it could be compatible with the glass breaking at some point. (Lipton, 2004, p. 37) Contrastive questions are regularly used in everyday life and we usually fall back to contrastive ways of stating a question to refine our explanation interest. We may ask why Anna watched an episode of 'Breaking Bad' and get the answer that she was bored. We may then refine our question and ask why Anna watched 'Breaking Bad' rather than 'Lost', if the boredom was obvious and we never wanted information about that, but on the specific choice of watching an episode of exactly *that* series.

The question asked with the hard problem of consciousness, the question why phenomenal consciousness exists, has a related contrastive form too. We could ask why there is phenomenal consciousness rather than not. Those two questions seem close enough to be both answered with a causal history of phenomenal consciousness.

There is still a last distinction for explanations available. In detail, we can classify explanations regarding their precision or vagueness. An explanatory claim can involve strictly defined terms and cite several micro steps involving causation. It may as well be a claim about a causal process only on a macro level, without saying anything more detailed. In some cases even an analogy or metaphor may pass as an explanation. An example: If we want to explain redshift, we may only say that it is similar to the change in sound of an ambulance driving by. We could, however, also say that interference processes cause the

[15] This view of impossible contrastive explanations of chance events has been contested. See for example Hitchcock (1999). We will not go further into this.

effect, to give a more detailed and precise explanation. If one happens to be a physicist, she will probably give an even more accurate explanation.

How precise a potential explanation has to be, to suffice as a real explanation, depends on the context and especially our explanation interest. In some cases we need an explanation that opens up exact calculations, whereas in others we do just fine with a metaphor.[16]

With these considerations on consciousness and explanation in mind, we can state the mysterian challenge.

1.2.5 Stating the Mysterian Challenge

For solving the hard problem of consciousness we need to explain the existence of phenomenal consciousness. We need an as detailed as possible, causal history of qualia. With no overwhelmingly convincing explanation candidate in place, the possibility of reaching such an explanation is questioned. The mysterians, today mostly associated with Colin McGinn (1999; 1993b), argue that we are unable to reach any such explanation. This marks a *mystery* and not just a solvable problem in the labelling of Noam Chomsky (1976). Most of the so called *new* mysterians tend toward a non-provable naturalism, whereas other mysterian movements are mostly dualist in nature. Both groups, however, accept our inability to solve the hard problem.[17] More agnostic views claim that we cannot grasp how such an explanation by current natural sciences may look like. (Nagel, 1986; 1974) (Strawson, 2010; 2006a; 2008 (2003)) While Strawson claims that this requires us to change our natural sciences towards panpsychist physics, the destructive part of his project, the limit of current sciences, can be viewed individually. It poses the same challenge as the fully committed mysterian position does. All those positions blame our poor epistemic abilities for our inability to solve the hard problem of consciousness. The challenge therefore is as follows:

> Mysterian Claim: *We are unable to explain the existence of phenomenal consciousness because of our epistemic limitations.*

[16] The accessible precision is often limited and we have to accept vague explanations.

[17] Steven Horst (2007, p. 117) goes even further and considers every position in the philosophy of mind that accepts the explanatory gap as some sort mysterianism.

If this claim is correct, all our attempts to solve the hard problem are futile. In this case we better stop wasting our time and go on with more productive projects. McGinn even provides a list of better things to do than being puzzled about the hard problem (1999, pp. 69-76). The usual answer to the mysterian claim denies the epistemic limits. We will refrain from fully doing so, but rather argue that most, or maybe even all, of the proposed limits do not stop us from explaining phenomenal consciousness. They only force a different, indirect method. But why should we even feel obliged to answer the mysterian challenge? Innumerable relevant epistemic limits were proposed and argued for, in the history of philosophy. We will present some of them in a schematic fashion to motivate our project.

1.2.6 Mysterian Denial and Epistemic Limits

In the beginning we need to acknowledge that our discussion of epistemic limits and mysterianism will certainly fail to be complete. Nevertheless, we inquire the most common paths for arguing in favour of mysterianism. Initially we have to distinguish epistemic limits regarding the hard problem of consciousness from other types of limits of knowledge. The idea that we might not be able to know everything is obviously an old one. Discussions about epistemic scepticism were already present in ancient Greek philosophy[18] and reintroduced in modern philosophy prominently by Descartes' evil demon and dream scenarios in his *Meditations on First Philosophy*. An epistemic skeptic doubts the possibility of knowledge in general, i.e. we might be unable to generate knowledge on epistemic interesting propositions. We may be unable to know anything about the external world, about the past, etc., even though common sense might lead us into thinking otherwise. This can be interpreted to imply the denial of knowledge, or just withhold judgment whether we have knowledge or not.[19] A denial of knowledge is often a result based on some mistrust to our sense perception and therefore ignorance of the real external world. Descartes' skeptic scenarios hinge on arguments like that (Stroud, 1984). Epistemic skepticism implies being skeptic towards explaining the existence of phenomenal consciousness. This, however, is not result of ignorance specific to the hard problem of consciousness. The skeptic is also committed to deny any indirect way of grasping an explanation and therefore our whole

[18] For example in Pyrrhonism. Cf. (Vogt, 2011)
[19] That is typically taken to be a difference of 'academic' and 'pyrrhonian' sceptics. Cf. (Klein, 2011).

project in this thesis. Hence we will not further talk about skepticism and focus on more constrained epistemic limits.

We will also not discuss epistemic limits that are purely *contingent* in contrast to essential limits. Contingent limits can in principle be overcome. Humans cannot know how a certain small rock on a planet on the other side of the Milky Way looks like. However, there is no reason why we should forever be unable to do that. It is certainly difficult to fly there, but it is obviously not impossible.

The third kind of limits we will not discuss are logical limits. As Gödel's (1931) incompleteness theorems show, there are limits to proving axiomatic systems. Another limit closely related to logic is based on an anonymous reviewer and put forward by Fitch (1963), who pointed to a theorem of necessary unknowable propositions. The core of Fitch's argument is this: for every unknown truth x there is an unknown truth y about x being an unknown truth. To every contingent ignorance corresponds a necessary ignorance. (Williamson, 2000, p. 270) The corresponding prove of Fitch is still part of a wide debate in which we will not go into.

After listing which limits we are not discussing, we need to provide examples of epistemic limits that do challenge the explanation we seek for, without denying knowledge in general. Such limits come in different forms and we will provide two distinct types of them for an overview. They can target scientific knowledge or focus on our cognitive abilities (or the brain). We will provide examples for the first option based on a view on physics by Bertrand Russell and Arthur Eddington and illustrate the second way with Noam Chomsky and Colin McGinn.

A general way of attacking natural sciences, and especially physics, in the debate on the hard problem of consciousness, is to point towards physics limit to grasp intrinsic properties. When Strawson (2006b) argues against the possibility of a physical explanation he claims:

> "Another thing we know about it, let us grant, is everything (true) that physics tells us. But what is this second kind of knowledge like? Well, there is a fundamental sense in which it is 'abstract', 'purely formal', merely a matter of 'structure', in

> Russell's words. This is a well established but often overlooked point. 'Physics is mathematical', Russell says […]" (Strawson, 2006b, pp. 9-10)

William Seager adds that "[…] another way to put this, in line with the way Russell views things, is that all that science provides, or can provide, is structural or purely mathematical information about the world." (Seager, 2006, p. 135) In the Russellian view this is certainly true and can be explained by at least two, somewhat connected, approaches. The first one is based on his differentiation between knowledge by acquaintance and knowledge by description.[20] Acquaintance is a direct relation to sense data, universals and perhaps oneself (Russell, 1910, pp. 109-111). Anything we are not acquainted with can only be known by description. Descriptions are complexes of directly acquainted parts fit together based on rules. The clue in denying physics the possibility to explain phenomenal consciousness is implied in this distinction. Physics only works by description based on previously acquainted material and therefore provides knowledge only by description. This descriptive nature implies a restriction to relational properties. Unfortunately, one might argue that qualia come with intrinsic properties which cannot be explained by relations.

The second way to argue for a limit of physical knowledge is based on Russell's thoughts put forward in the *Analysis of Matter* (2007 (1927)). We find him starting with two general assumptions:

1. All knowledge relevant to physics is either analytic, like pure mathematics and logic, or (partially) derived from perceptions, which Russell calls 'empirical' (Russell, 2007 (1927), pp. 175-176)
2. The basic principle of epistemology must be to put the more certain before the less certain (Russell, 2007 (1927), p. 179)

Based on those general assumptions Russell finds the most primitive starting point for empirical physics in the percepts. Those are more primitive for this purpose than the process of perception, as the knowledge of our percepts is immediate whereas the knowledge of our perception requires additional reflection. Through this method of disentangling the primitive

[20] Cf. Russell (1910; 1912)

from the interfered elements, Russell begins his pursuit of the foundation of physics. The most primitive basis needs to be discovered, and doing so seems to be an easy task for Russell:

> "[…] I find that the task is not really very difficult, except in certain niceties. The primitive part seems something like this: There are coloured shapes which move, there are noises, smells, bodily sensations, the experiences which we describe as those of touch, and so on. There are relations among these items: time-relations (earlier and later) among all of them, and space-relations (up-an-down, right-and-left, and the relations by which localization in the body is effected) among many of them. There are recollections of some of these things […]. There are also expectations […]. Of thoughts other than memories and expectations, it is not necessary to take account when our sole purpose is to reach the primitive basis of our knowledge of matter." (Russell, 2007 (1927), pp. 180-181)

The interesting part is that the basis of physics is not at all as it was expected to be. From an epistemological point of view, there are no physical objects serving as part of this foundation. Russell describes this in a famous paragraph:

> "In the above account, I have omitted many things which I formerly 'knew,' and which, apparently, most other people 'know.' I have omitted 'objects.'" (Russell, 2007 (1927), p. 181)

The foundation differs from common sense solely because of the fact that it is an expression of the essential principle of epistemology, the search for the most primitive basis. Everyday objects like tables, books, trees, etc. are less primitive than those percepts listed by Russell. The theoretical entities, usually assumed to be the material of those tables or books (atoms, electrons, protons, etc.), are likewise less primitive than the percepts through which they occur. Objects in general are inferred from the epistemologically primitive. (Russell, 2007 (1927), p. 181) At this point Russell finds himself caught up to explain how we infer from these primitives some alleged substantial identity with varying properties, which we call 'objects'. Following Russell, we need to drop the feature of 'substantial identity' from our notion of objects. What we usually label as object is rather a series of occurrences with certain relations that link them to each other. We do not need to go into Russell's

explanation of the way we build such a relational object. It is sufficient to note that any physical object is merely a group of percepts linked to each other.

What we need to grasp a complete account of the Russellian epistemic limit is an interrogation of another important aspect of scientific inference: 'structure'. Structure is a notion describing relations or systems of relations. Two structures are similar if you can map them one to one via a function. P and Q are similar if a function F exists that maps all terms related P to correlate terms related Q. Russell's examples are helpful for clarification:

> "A map, for example, if accurate, is similar to the region which it maps. A book spelt phonetically is similar to the sounds produced when it is read aloud. A gramophone record is similar to the music which it produces. And so on." (Russell, 2007 (1927), p. 249)

Consequently two relations that are similar have the same structure. The important point for our task of reconstructing a possible limit of physics is that the structures of relations determine their logical properties. Logical properties, such as transitiveness, symmetry, having n terms in its field and others are part of the structure of the relation, a connection which implies one of the pivotal points for physics:

> "When two relations have the same structure (or relation-number), all their logical properties are identical." (Russell, 2007 (1927), p. 251)

With that in mind, we can take a look at the relation between the space of physics and the space of perception. There is a difference between perceived space relations among visual percepts and the space relations that physics assumes among physical objects. They are not identical, but are related. They are related in a certain kind of correspondence in their relations: their structures are the same. Russell explains his view:

> "My point is that the relations which physics assumes in assigning angular coordinates are not identical with those which we perceive in the visual field, but merely correspond with them in a manner which preserves their logical (mathematical) properties." (Russell, 2007 (1927), p. 253)

The correspondence preserves the logical properties because of the sameness in structure. A same structure implies identical logical properties. The same considerations hold to colours, tastes, sounds, etc. The foundation in the primitives, the percepts, provides the basis for their relations.

> "This applies to all varieties of percepts, and accounts for the fact that our knowledge of physics is mathematical: it is mathematical because no non-mathematical properties of the physical world can be inferred from perception." (Russell, 2007 (1927), p. 253)

With these considerations in mind we see that there are no non-mathematical properties to be inferred because we cannot grasp anything more certain to us than percepts and their relations. Physics is something that begins after the occurrence of those percepts and corresponds to them. The recognition of similar structure is all that it can provide. This position implies an identity of all logical properties and asserts that physics cannot go beyond those properties to reach any intrinsic basis. If we assume that phenomenal consciousness has intrinsic properties, we find physics unable to explain phenomenal consciousness.

Sir Arthur Eddington is usually the authority listed next to Russell in arguments for purely structural and therefore limited physics. Eddington's views on the nature of exact science and its limits are expressed in his *The Nature of the Physical World* (Eddington, 1929). When Eddington writes about 'exact science', he has physics in mind. He introduces the way physics handles problems with a formidable example that we will recount here. Any usual examination paper in physics starts with an observation in everyday terms such as "an elephant slides down a grassy hillside…." (Eddington, 1929, p. 251). The next step is to replace the 'elephant' with its mass; but to which property does 'mass' refer? Because we cannot answer this question, we should rather ask a different one: "how has the elephant actually entered in so definite a way into our experience?" (Eddington, 1929, p. 251) What *is* two tons? Eddington gives a simple and plausible answer: "two tons *is* the reading of the pointer when the elephant was placed on a weighing-machine." (Eddington, 1929, p. 251) Following him, we can replace all the inexact everyday terms such as 'elephant', 'hillside',

and ‘grassy’, with their exact mathematical interpretations for the physical description of the elephant’s slide. Now we speak of certain masses, slope of the hill, friction of the grass and so on. Those terms are defined purely through the reading of pointers by devices in measurements. “And so we see that poetry fades out of the problem, and by the time the serious application of exact science begins we are left with only pointer readings.” (Eddington, 1929, p. 252) This is a mere observation by Eddington: the properties and values that physics works with are based on constant measurements and cannot reach beyond their status as pointer readings. Eddington’s observation implies a huge limitation in physical knowledge. If properties are solely responses to various ways of confronting objects with measurement devices, then these measures can only capture the interaction, which is to say, the relation between the indicator and the object being analysed. “After all, knowledge of this kind is fairly comprehensive. A knowledge of the response of all kinds of objects – weighing-machines and other indicators – would determine completely its relation to its environment, leaving only its inner ungetatable nature undetermined.” (Eddington, 1929, p. 257) Knowledge of physics is limited to relational information. We cannot go beyond this limit through the means of an exact science. We are unable to grasp the inner nature of any object through the means of measuring devices. “But now we realize that science has nothing to say as to the intrinsic nature of the atom.” (Eddington, 1929, p. 259)

Similar limits were also argued and discussed by, but are not limited to, David M. Armstrong (1961; 1968), Simon Blackburn (1990) and are already present in David Hume’s *Treatise* (1.4.4). A recent version based on the cellular automaton *Game of Life* has been proposed by Rosenberg (2004). We could also argue by observing how the physical particles are defined. If we ask what an electron is all we will be answered is that it is a ‘particle’ with a certain mass (9.10938188 * 10^{-31} kilogram), electric charge -1, spin ½, etc. - all relational attributes. (Seager, 2006, pp. 134-135) They all provide a relevant epistemic limit for an explanation of phenomenal consciousness’ existence, if we grant that qualia are or have intrinsic properties. Exactly this is frequently accepted. Ned Block for example holds that “Qualia realism […] is the view that there are intrinsic mental features of our experience” (Block, 1990, p. 53). Nagel claims as well that whatever it is like to be in a state, is intrinsic. (Nagel, 1974, p. 445) And Rosenthal also concedes that “the idea that being conscious is an intrinsic property has strong intuitive appeal” (1997, p. 737).

Following Noam Chomsky, Colin McGinn argues for an epistemic limit distinct from scientific knowledge. However, according to Akeel Bilgrami and Carol Rovane (2005), McGinn's way of spelling out the idea does not fit the intention of Chomsky completely. Chomsky thinks that we should distinguish between problems and mysteries. Problems are solvable as long as we formulate them clearly, whereas mysteries cannot be answered due to our limited cognitive capacities. (Chomsky, 1976) McGinn believes that the mind-body problem, which we take to be synonymous to the hard problem of consciousness, is one of those mysteries. Chomsky himself presents it differently. We are rather unable to provide a clear question regarding the mind body problem. He states:

> "Nagel states that 'the mind–body problem was posed in its modern form only in the seventeenth century, with the emergence of the scientific conception of the physical world on which we are now all brought up' (1993: 97) (the Newtonian conception). But that has the story reversed. The mind–body problem made sense in terms of the mechanical philosophy that Newton undermined, and has not been coherently posed since. If so, discussion cannot proceed in Nagel's terms without some new account of the nature of body (material, physical, etc.) and mind." (Chomsky, 2000, p. 86)

and

> "The presupposed concepts 'physical' or 'material' have no clear sense; nor will 'mental', unless some sense can be given to the notion of potential consciousness and, even then, it is unclear what the interest of this particular category would be, as distinct from many others. It is not the business of the sciences to express the content of ordinary discourse about anything, physical or mental. There seems to be no coherent doctrine of materialism and metaphysical naturalism, no issue of eliminativism, no mind–body problem." (Chomsky, 2000, p. 91)

Most philosophers would disagree that there is no mind-body problem. We may concede that we do not have a concise formulation of the problem in a way everyone agrees with, but the existence of a hard problem of consciousness is widely accepted. This is also supported by the huge amount of published papers struggling with the problem and responding to each other.

Chomsky's distinction between mysteries and problems can become valuable for the philosophy of mind even though he does not believe that the hard problem is a mystery. In contrast to his assessment, McGinn claims it is one and points to our cognitive closure. The basic idea behind his view is an evolutionary one that is easily acceptable. Our cognitive capacities are a result of an evolution with selection processes that made us excel in certain areas and deficient in others. We are pretty good at recognizing faces, acquiring a language, see and identify objects of a certain size in a small to medium range, etc. On the same token, we are unable to fly, cannot hear or smell as good as, say, dogs, are subject to illusions like the Müller-Lyer arrows, etc.. McGinn finds that "evolution has equipped us with a large number of these modules, each with its own job to do. Thus we have certain biases built into our mental architecture, certain things we are adapted for as against others." (1999, p. 41) This is certainly true, but is solving the hard problem of consciousness one of those actions we are unable to do? One argument we can provide is the sheer dissatisfaction of the answer-attempts in history. The hard problem[21] is discussed for a very long time without solution. We may take this history at face value and take it as an inductive reason that there will never be a solution. An argument, which McGinn himself uses, is as follows: consciousness seems to be a property that is less intellectually demanding than others. Conscious states seem biologically primitive, comparatively speaking, and it is rather natural to ascribe conscious states to various, less complex animals. If that is correct: why should something not especially complex be hard to grasp, based on its complexity? It is plausible that the mind-body problem is not too difficult to overcome due to its size, but rather because it is of a type that we cannot solve. (McGinn, 1993a (1989), p. 19) He combines this with attempts to show that neither introspection, nor a neuroscientist position can bridge the gap between phenomenology and brain. His reasoning incorporates the problem of the brain's spatiality and consciousness lack thereof. (McGinn, 1993a (1989), pp. 11-12) The philosophical puzzlement of the mind-body problem is a result of our belief that the problem must be scientific, but that any science we come up with cannot solve it. It is a scientific problem without the means to solve it because of our cognitive closure. (McGinn, 1993a (1989), p. 18)

[21] Not always with this specific name.

Those examples of possible limits relevant to explaining the existence of qualia are not exhaustive. Nevertheless they provide a good idea why we might be unable to solve the hard problem. Each of those limits can be objected. We might deny the Russellian idea of the acquaintance/description distinction. We could also doubt that McGinn gave us enough of a reason to believe in relevant cognitive closure, or claim that Eddington's account of physics is simply false. Nonetheless those approaches put reasonable doubt on our ability to directly access a solution to the hard problem. It is anything but certain that we live in a non-mysterian world. Regardless of the question if these are the strongest ways to put mysterian intuitions forward, we need to be able to cope with the possibility of only one of them, or any sufficiently similar thesis, being true. With this in mind, the question arises: What is the correct response to a somewhat successful mysterian challenge, if we want to stick to the aim of explaining the existence of phenomenal consciousness?

1.3 An Answer in Sight

The answer we put forward is an approach of inference to the best explanation that evaluates explanation candidates and lets us conclude to the best of them. We may be limited in our access of relevant facts for the hard problem, but we certainly can work with the evidence we have. Maybe we cannot defeat all forms of mysterianism, but we should succeed in countering most of them. Our proposal is this: we should form and collect as many plausible hypotheses as possible. In the next step, we filter the inadequate ones by consistency with hard sciences and coherence with our intuitions related to the hard problem. After this filter mechanism, we can either hope for the very unlikely case of finding a filter that leaves us with only one explanation candidate, build a disjunctive super-theory, or chose the loveliest out of the explanation candidates left. If everything works out well, this results in an explanation, not only a mere explanation candidate. Our aim is to spell out this basic idea. We will discuss the pool of explanation candidates; the filter mechanism in both consistence with scientific knowledge and coherence with our intuitions; what kind of intuitions we can use; our options after the filtering; and lastly the epistemic status of an outstanding explanation candidate.

This answer is able to face the mysterian challenge as follows:

1. To reject Russellian and Eddingtonian speculations, we have two different responses. First, we may demand a more refined notion of intrinsicality and an argument why qualia are or have intrinsic properties. As long as these criteria are not met, the limit of physical knowledge cannot be treated seriously. Our inference to the best explanation, however, enables us to go for the second and preferred response. We can answer the Russellian-Mysterian, even if intrinsicality and intrinsic qualia are accepted or sufficiently argued for. Our following proposal for an inference to the best explanation is not restricted to knowledge by means of physical theories or experiments. Even if physics cannot grasp intrinsic properties, we can handle hypothetical, intrinsic, physical properties. Although physical theory cannot deal with intrinsic properties, we are not committed to deny intrinsic, physical properties as such. We now leave the ground of physics and free ourselves from its limitations. We lose the tight connection to its justificatory processes, but will replace it with our method of inference.
2. McGinns attack is a bit trickier. If McGinn is right, we should stop wasting time with the hard problem. He suggests studying the physical side or phenomenology singularly instead. We should not try to combine them, but rather research the bounds of our cognitive access in each dominion separately. (McGinn, 1999, pp. 68-76) This is a very unsatisfactory conclusion and we ought to provide a better alternative in the second chapter. This is possible, because McGinn's denial of a possibility of an inference to the best explanation aims only at those inferences based on perception. McGinn thinks that there is homogeneity in concept introduction that prevents us from forming concepts that differ from perception in an extent that would free it from perception's limits.[22] However, we deny this assumption. There is no reason why there should be any such homogeneity. We know that there are concepts in physical theories that are at best loosely connected to perception. For example, wave-particle-dualism is hardly perception based. Moreover, mathematical concepts are usually[23] not treated as being based on perceptions. Although McGinn seems to accept this for mathematical concepts

[22] Cf. McGinn (1993a (1989))
[23] Some sort of Platonism may argue for that.

(1993a (1989), p. 16), he rejects it for concepts that can explain phenomenal consciousness. However, there seems to be no good reason for doing so. We can form concepts without depending on perception. Having the (possibility of) concepts is only one part of solving the mind-body problem though, and we need to question whether we can reach justification. We cannot justify by perception or introspection, because those are in McGinn's sense restricted. Nevertheless, we can use perception and introspection to collect data that is used to generate justification by an inductive inference to the best explanation. In a sense we split up McGinn's attack into a conceptual mysterianism and a justificational mysterianism, both related to perception. We deny the conceptual side while accepting the limits of justification by perception. Therefore we reach concepts suited for phenomenal consciousness, but are in need of an independent justificational practice: the proposed inference to the best explanation.

2 The Method

2.1 The Pool of Explanation Candidates

Setting up the pool of explanation candidates is difficult. There are potentially infinite of them and we certainly cannot assemble them all. A more promising approach is to concentrate on those that are or have been in fact proposed in the mind-body debate. Unfortunately, a huge number of them exist and listing them all is neither possible nor necessary for this thesis. Some of the most prominent are: Reductive Physicalisms (Type-Identity Theory); Non-reductive Physicalism (Supervenience; Realization; Emergence); Representationalism (first and higher order); Functionalism; Substance Dualism (Interactionist; Epiphenomenalist; Parallelist); Property Dualisms (Interactionist; Epiphenomenalist; Parallelist); Panpsychism; Neutral Monism; – most with or without smallism[24]. Restricting ourselves to actually proposed theories does imply that the justification we gain is limited. We can provide only prima facie justification relative to the set of all possible explanations. Nonetheless, our approach justifies, at least in the sense that competing explanation candidates are filtered, ruled out and the remaining theories therefore get an epistemic boost. We will discuss this in chapter 3.

Our part here is modest. To begin with, we inquire on a general level what attributes an explanation candidate should have and in what form it should be stated. Note that explanation candidates are built based on speculation. Speculation might seem pejorative at first sight, but it is not. Hypotheses motivated by a specific explanation interest are the perfect starting point for any inference to the best explanation. We grant the mysterians that we cannot find an answer by direct perception. We accept that we cannot solve the hard problem by merely looking at the brain or introspecting. With these limits in place, the best starting line is moderate speculation. 'Moderate' because we can derive a first approximation of what requirements those hypotheses have to fulfil, by considering what our goal is and how we want to accomplish it. As we aim at an explanation for the existence of phenomenal consciousness, it is obvious that theories, which do not involve this very

[24] Smallism is the doctrine that the smallest fundamental particles on the micro-level determine the larger phenomena on the macro-level. The term is coined by Coleman (2006).

explanandum, will not be selected into our set of candidates. They do not even remotely fit our explanation interest. This explanation interest can be broadly understood so that any theory that makes claims about the causal history of phenomenal consciousness is suitable.[25] This includes vague theories about the existence of qualia. Those are very unlikely to be filtered out with consistency or coherence arguments and will pose problems later on. However, at the current state of our inquiry we can include them. We only need to be aware that for most theories more vague forms exist, which might be included in our pool as well. For example: a theory may be detailed about a 3-step causal mechanism, whereas a more general formulation provides a vague description about the same mechanism. If anything counts in favour of the first theory, it does so for the broader version as well. However, the explanatory power of the more tangible theory will be higher and in general favourable. We will come back to this point later.

We plan to filter with coherence and consistency considerations and therefore need to formulate the explanation candidates in a way they can actually be evaluated based on those. The best way to fulfil this requirement is to present the theories as a set of sentences. Sentences involve the propositions needed for coherence considerations and can be treated as potential beliefs. However, philosophical theories are rarely presented as a concise set of sentences. Often they are presented as a mesh-up of arguments, anecdotes, analogies, examples, laws, regularities, postulated entities etc., and usually spread out over several different papers. We present an idea how an explanation candidate can be put forward in the rigor fashion needed for our purpose and provide examples after the general remarks.

A further problem that we are confronted with in picking the theories for our initial pool is that different variations of theories are frequently labelled as being a single one. We find not only the broadest taxonomies like ‘dualism’ or ‘materialism’ denoting innumerable theories, but rather even smaller labels like ‘interactionism’ or ‘parallelism’ falling into the same trap. It would therefore be misleading to attempt a characterisation of a single theory like ‘dualism’, ‘physicalism’, etc. and we have to be careful with such labels.

[25] Stating a concrete explanation interest in a concise fashion is tough. The ordinary practice is just to ask the question and reframe it if it is misunderstood or not sufficiently answered.

2.1.1 Example I – Galen Strawson's Fundamental Duality Monism

A relatively easily accessible example how our explanation candidates should look like, can be found in what Strawson himself sometimes calls *real physicalism* (Strawson, 2006b), *real materialism* (Strawson, 2008 (2003)) or *real naturalism* and is frequently and correctly named panpsychism. The more unique and therefore useful label he provides is 'Fundamental Duality Monism' (Strawson, 2006a). This is our first example because Strawson himself provides a list of possible theorems of a theory of mind and chooses which ones of them ought to be in his theory. (Strawson, 2006a, pp. 222-223) It is a list of metaphysical and epistemological theses based on which Strawson proposes different theories. One is as follows[26]:

(FDM 1) There is a fundamental sense in which there is only one kind of stuff in reality.

(FDM 2) There is experiential[27] reality, experiential being.

(FDM 3) There cannot be experience without a subject of experience.

(FDM 4) There cannot be a subject of experience without experience.

(FDM 5) There is physical reality, physical being.

= (FDM 5*) There is non-experiential being.

(FDM 6) All reality (or being) is either experiential or physical.

= (FDM 6*) All reality is either experiential or non-experiential.

(FDM 7) There is only one kind of stuff and it has both experiential reality and physical reality and no other kind of reality.

(FDM 8) There is only physical reality.

(FDM 9) All experiential reality is physical reality.

(FDM 10) All physical being involves experiential being (given that Physicalism is true).

= (FDM 10*) (All concrete) being involves experiential being.

(FDM 11) At least some ultimates are experience-involving.

(FDM 12) All physical reality is (at bottom) the same kind of stuff.

= (FDM 12*) All reality is (at bottom) the same kind of stuff.

[26] The numbers differ from Strawson's table. Experential is synonymous to phenomenal. '*' marks restated formulation with the term 'physical' being replaced.

[27] 'Experiential' can be read as 'phenomenal' here.

(FDM 13) Experiential reality cannot be wholly non-experiential reality.

(FDM 14) Non-experiential reality cannot be wholly experiential reality.

(FDM 15) Experiential reality cannot possibly emerge from wholly and utterly non-experiential reality.

(FDM 16) Non-experiential reality cannot possibly emerge from wholly and utterly experiential reality.

(FDM 17) All facts are (fully) determined by facts about ultimates.

(FDM 18) There is only one fundamental kind of stuff. It (all of it) has both experiential reality and non-experiential reality. There is no other kind of reality. (Strawson, 2006a, pp. 221-236)

(FDM 18) is the core thesis of fundamental duality monism. How does the whole theory try to explain phenomenal consciousness? It accepts it as a fundamental part of the basic reality. The one fundamental kind of stuff has experiential properties and every occurrence of qualia is caused by properties that are experiential themselves. The causal history of an experiential event always involves an experiential property itself. This might look circular at first glance, but it is an adequate explanation candidate. An explanation does not require to be completely explained itself. It only demands information about the causal history, which is satisfied with the thesis that every experiential event is caused by other experiential events. Following this concept, the fundamental duality monist seems to be in the same boat as loads of dualistic views, e.g. forms of parallelism. This is mistaken because experiential events are, in this view, not independent from non-experiential events. There is only one kind of fundamental stuff and fundamental physical entities (ultimates) do have experiential reality as well. This is not a necessity, but rather a contingent feature of our world. To complete the explanation candidate, we may provide a thesis of the causation-relation. The following fits the role: (FDM 19) Experiential (phenomenal) events are caused by events involving the experiential reality of ultimates.

2.1.2 Example II – 'Nonreductive' Physicalism(s)

Providing a single theory of nonreductive physicalism is certainly impossible. There are too many different views denoted with this single label. Some of them require the existence of only physical entities, which has its downfall with social reality or mathematical objects. Most of them focus on physical properties or being somehow based on them. Usually, physicalism is taken to be a thesis about the empirical world. However, sometimes it is proposed as a view that a true complete physical scientific description states all the empirical world contains.[28] Physicalisms based on this idea are *Theory View Physicalisms*. (Stoljar, 2010) The main problem for them is Hempel's dilemma. On one hand, we cannot define physicalism based on what current physics tells us. We can see the force of a pessimistic meta-induction, claiming that certainly something in today's physics is false, because there were always scientific mistakes in the past. At the same time, future physics might be too vague to provide a coherent position. (Ney, 2008, pp. 1036-1037) Crane and Mellor (1990) even claim that this stops physicalism before it can even start.[29] Nonetheless, there are approaches to physicalism that do not rely on physics. Daniel Stoljar (2006) suggests an understanding of the physical based on paradigm objects: "Take objects like this toaster, and now take whatever principles and theories are required to explain their nature and operation. […] these are the principles and theories that […] provide the supervenience base for whatever it is that explains the whole world." (Stoljar, 2006, p. 30) *Object-based physicalism* is also vague in nature and therefore difficult as an ontological basis. We will leave the question of the best definition for the term 'physical' aside and continue with a specific form of physicalism.

One way to spell out nonreductive physicalism can be found in the *Necessity View of Physicalism*. (Stoljar, 2010) The gist of this position is the claim that physicalism is true if and only if every instantiated property is necessitated by some physical property. We thereby accept non-fundamental, non-physical properties. Supervenience-relations are a possible option for this idea. We can define a (global) supervenience physicalism as follows:

[28] 'Physical sciences' refers not only to physics, but also the likes of chemistry, molecular biology or neurophysiology, but certainly not economics or sociology. (Crane & Mellor, 1990, pp. 185-186)
[29] Others like Pettit (1993) are less pessimistic.

> Physicalism is true at world w if and only if for any possible world w* if w* is a physical duplicate of w, then w* is a duplicate of w simpliciter. (Stoljar, 2010, p. 116)

We can provide a basic supervenience physicalism as follows:

> (SP 1) There is a fundamental sense in which there is only one kind of stuff in reality.
> (SP 2) There is physical reality.
> (SP 3) There is phenomenal reality.
> (SP 4) Phenomenal reality is not wholly physical reality.
> (SP 5) Physical reality is not wholly phenomenal reality.
> (SP 6) In the actual world w it is true that: for any possible world w* if w* is a physical duplicate of w, then w* is a duplicate of w simpliciter.
> (SP 7) Any instantiated property is necessitated by some instantiated physical properties.
> (SP 8) Any phenomenal property is supervenient on physical properties.
> (SP 9) On the fundamental level there are only physical properties.
> (SP 10) All facts are (fully) determined by facts about ultimates.

Now we used the term 'physical property' which we define either by paradigm objects, physical theory, epistemic accessibility, scientific method, or by contrast to something non-physical. We can invoke a cluster concept of the theory view using most of them:

> F is a physical property if and only if F:
> (a) is expressed by a predicate of a physical theory that is true at some possible world,
> (b) is not one of the distinctive properties of souls, ectoplasm, etc,
> (c) is objective, and
> (d) is the sort of property about which one could find out using the methods of the sciences. (Stoljar, 2010, p. 84)

This supervenience physicalism explains the existence of qualia by pointing to the relation of the supervenient property to the supervenience basis and the causal history of the supervenience basis. Hence, we may add:

(SP 11) Experiential (phenomenal) events are caused by and supervene on physical events.

An alternative physicalist route is to replace the supervenience-relation with a realization-relation. Note that realization is defined as follows:

> Token x realizes token y iff (i) y is a token of some functional type, F, such that, necessarily, F is tokened iff there is a token of some or other type that meets condition, C; (ii) x is a token of some type that in fact meets C; and (iii) the token of F whose existence is logically guaranteed by the holding of condition (ii) is numerically identical with y. (Melnyk, 2003, p. 21)

Realization invokes a second order property move for physicalism. A realized property is a second-order property that can be realized by multiple first-order properties. The cloths property of being absorbent can be realized by various physical configurations, while being of the same functional type. A realization physicalism can be defined as being true at world w iff every property instantiated at w is either identical to some physical property instantiated at w or is realized by some physical property instantiated at w. (Stoljar, 2010, p. 123)[30] A realization physicalism would be a theory like this:

(RP 1) There is a fundamental sense in which there is only one kind of stuff in reality.
(RP 2) There is physical reality.
(RP 3) There is phenomenal reality.
(RP 4) Phenomenal reality is not wholly physical reality.
(RP 5) Physical reality is not wholly phenomenal reality.
(RP 6) Every property instantiated at w is either identical to some physical property instantiated at w or is realized by some physical property instantiated at w.
(RP 7) On the fundamental level there are only physical properties.

[30] Similar definition by Melnyk: Every property instance (object, event) is *either* an instance of a physical property (an object of a physical kind, an event of a physical kind) *or* an instance of some functional property (an object of a functional object kind, an event of a functional event kind) that is *realized by* an instance of a physical property (an object of a physical kind, an event of a physical kind). (Melnyk, 2003, p. 22). Melnyk adds that this definition might be too weak and we better talk about physical realization and not mere realization. (2003, pp. 23-26)

(RP 8) Realization definition (see above).

(RP 9) Any possible world that has exactly the same distribution of physical tokens and laws of physics as the actual world contains all the tokens of nonphysical (functional) types that the actual world contains. (Melnyk, 2003, p. 51)[31]

(RP 10) All facts are (fully) determined by facts about ultimates.

A physical property is defined by cluster as above. Any event incorporating phenomenal consciousness is explained by the realization relation and the causal history of the physical event that realizes the experiential event.

(RP 11) Experiential (phenomenal) events are caused and realized by physical events.

Those two versions are formulation of a basic kind that can be enhanced. For example: We could add a hypothesis about coming to know physical reality or phenomenal reality. David Lewis (1983) proposed this idea. We can formulate this as follows:

(Ab 1) Theories about physical reality cannot wholly teach us about phenomenal reality.

(Ab 2) In experiencing phenomenal reality we acquire an ability to recognize, imagine and predict one's behaviour by means of imaginative experiments. This ability cannot be exhaustively taught by a physical theory.

A close relative is to be found in replacement of 'ability' with the notion of know-how.

Another enhancement can be found in Conee's (1994) acquaintance hypothesis.

(Ac 1) Theories about physical reality cannot acquaint us to all properties.

(Ac 2) In experiencing phenomenal reality we become acquainted with a property.

It could also be enhanced by a theory of conceivability (intuition) production. E.g. with the principles provided by Hill (1997).

(Ip 1) "If there are no substantive a priori ties between two concepts, then it is possible to conjoin either of the concepts with the negation of the other without

[31] That does not imply that nonphysical tokens in other possible worlds are realized equally to their realization in the actual world. It is not supervenience!

producing an inconsistency. That is to say, it is possible to use the concepts to conceive coherently of situations (i.e., to construct internally coherent descriptions that purport to represent situations) in which there are particulars that fall under one of the concepts but do not fall under the other concept. If it is within a subject's power to conceive coherently of its being the case that p, then, unless the subject has a good a posteriori reason to think that the proposition that p could turn out to be inconsistent with a set of propositions that have been shown independently to be necessary, the subject will come to believe that it is possible - or at any rate, come to believe that it is prima facie possible - for it to be the case that p." (Hill, 1997, p. 75)

(Ip 2) "If it is within a subject's power to conceive coherently of its being the case that p, then, unless the subject has a good a posteriori reason to think that the proposition that p could turn out to be inconsistent with a set of propositions that have been shown independently to be necessary, the subject will come to believe that it is possible - or at any rate, come to believe that it is prima facie possible - for it to be the case that p." (Hill, 1997, p. 75)

(Ip 3) There is no substantive a priori tie between physical reality and phenomenal reality.

(Ip 4) We lack a posteriori reasons that those concepts are necessarily coextensive.

2.1.3 Example III – An Interactive Dualism à la Descartes

Descartes take on the mind-body problem is interpreted in many different ways. Every now and then a new 'real Descartes' is proposed[32]. For this thesis, we use the simple and well-known substance dualist interpretation of him to create a basic theory of interactionism based on the *Meditations* and *The Passions of the Soul*. The claims for such a view are:

(ID 1) There is physical reality.

(ID 2) There is mental reality.

(ID 3) Mental reality is not physical reality.

(ID 4) Physical reality is not mental reality.

[32] For example: Galen Strawson builds his views on a different reading of Descartes. Cf. Strawson (2006a)

(ID 5) Mental reality is fundamental.

(ID 6) Physical reality is fundamental.

(ID 7) Phenomenal reality is mental reality.

(ID 8) Physical events can cause and be caused by physical events.

(ID 9) Mental events can cause and be caused by mental events.

(ID 10) Physical events can cause and be caused by mental events.

(ID 11) Mental events cause physical events by interaction at the Pineal Gland.

(ID 12) Physical events cause mental events via the senses and the Pineal Gland.

Phenomenal consciousness is explained by being caused either by physical or mental events. We can provide a vaguer version by dropping (ID 11) and (ID 12). It is possible to give other detailed accounts of the mental-physical interaction by replacing (ID 11) and (ID 12) with other interaction hypothesis.

2.1.4 Example IV – A Type Identity Theory

The type identity theory has been proposed in the 1950s. The core idea is to understand phenomenal mental states as type-identical with physical brain processes. It claims that "Sensations are nothing over and above brain processes" (Smart, 1959, p. 145). My feeling pain *is* some sort of c-fibre firing in my brain. Nothing is added to the brain process to bring about qualia, because a quale is just a certain neural process. Physical theories and studies of the human brain just need to be continued and, hopefully, sometimes in the future we will be able to give type identity statements for every mental state. Such a theory can be sketched as follows:

(IT 1) There is a fundamental sense in which there is only one kind of stuff in reality.

(IT 2) There is physical reality.

(IT 3) There is phenomenal reality.

(IT 4) Phenomenal reality is wholly physical reality.

(IT 5) Phenomenal event types are identical with physical event types.

(IT 6) On the fundamental level there are only physical properties.

(IT 7) All facts are (fully) determined by facts about ultimates.

Phenomenal consciousness is explained by the type identity theory just like any other physical event is explained, because it is just a physical event. (IT 8) Phenomenal consciousness is physical and is caused by physical events.

2.2 The Filter: Consistency and Coherence

Having set up the pool of explanation candidates, the next step is to provide a filter mechanism to rule out inadequate explanation candidates. We use two different mechanisms: one based on scientific knowledge and one connected to our intuitions.

If we want to use any data, scientific knowledge or intuitions, as a filter, we can do so in two different ways. We either demand our explanation candidates to be consistent or to cohere with them. The choice depends on how high we put the bar for an explanation candidate to stay relevant. Depending on the used notion of coherence, it is easier for a theory to be consistent with sciences than to cohere with them. Coherence implies consistency, but the opposite does not hold. To reduce our pool of potential explanations as much as possible, we should use coherence. But what precise coherence account should we use?

2.2.1 Starting Steps to Coherence

The first option is the notion of coherence from A. C. Ewing (1961 (1934)). His account of coherence requires that a set of propositions or facts does not only be consistent, but rather that the elements entail each other. Any element x of the set Y has to be entailed by the remaining elements n, m, … ∈ Y. E.g.: A set {A, B, C} coheres iff {A, B} $\rightarrow$ C and {A, C} $\rightarrow$ B and {B, C} $\rightarrow$ A. In addition, it requires that there is no subset of propositions Z within Y that is logically independent of all propositions in the remainder of the set. (Ewing, 1961 (1934), pp. 230-231) This account of coherence is certainly too strong. We can easily grant that a set fulfilling these requirements is coherent, but there are intuitively coherent sets that do not satisfy the conditions. Take a set of three propositions “Gregg is jailed for another 3 years”, “I saw Gregg yesterday hiding from the police” and “Gregg is especially talented in escaping from prison”. None of these is entailed by the other. Nevertheless, they intuitively fit together. Similarly we cannot expect to be able to logically entail any explanation candidate

from our intuitions and accepted scientific theories/data. We need a more lenient account of coherence.

C. I. Lewis (1946) employs a coherence account that might be better suited. What C. I. Lewis calls 'congruence' is determined by mutual support of propositions in the relevant set. The core idea, which serves as a basis for his model, is a situation at court. Suppose several relatively unreliable witnesses independently tell the same circumstantial story. Any such testimony taken singularly provides very little justification, but taken together the reports establish a high probability on what they agree upon. It is very unlikely that they just happen to tell the same story, and it is more probable that some truth is the reason for their agreement. A set of testimonial reports is coherent even though there are not logical entailment relations between the single reports. They just fit together well and this fit provides justification and an increased probability of being true. C. I. Lewis defines congruence (coherence) as follows:

> "A set of statements, or a set of supposed facts asserted, will be said to be congruent if and only if they are so related that the antecedent probability of any one of them will be increased if the remainder of the set can be assumed as given premises." (Lewis, 1946, p. 338)

He names only statements and asserted supposed facts, but it similar with propositions in general. Congruence by itself, however, can never provide any probability for a proposition's truth, but only boost pre-existing likelihood. This requires that some element of the coherent set brings initial probability (justification). The amount of initial probability needed to result into a probability capable of being relied on varies because it depends on the amount of congruence. C. I. Lewis thinks that any initial probability can be transformed into a likelihood of truth which is sufficient for rational and practical reliance. However, the question arises if this notion of coherence fares better for our purpose.

At first sight, it does. There are situations in which congruence works well. C. I. Lewis gives a good example: Say four persons, N, E, S and W, receive one card from an unexamined standard deck of cards. Each one looks at the card: N is dealt, first, the extra card belonging to no suit. E is dealt the king of hearts, S the queen of hearts and W the Jack of hearts. This is exactly the way never before shuffled decks are sorted. Each person has a low justification

for the belief that the deck was never shuffled, but taken all of them together, they have good evidence and are therefore well justified. (Lewis, 1946, pp. 344-345) However, we need to be cautious when sticking to examples like this. Simple scenarios lead us into thinking that we may in principle be able to assign the initial probabilities and then calculate the resulting likelihood. In more complex and abstract problems it becomes apparent that this is not the case. How should we reasonably attribute a starting probability to any of our explanation candidates? C. I. Lewis thinks this is no problem, but it certainly is, because we cannot know which amount of congruence is needed for providing sufficient justification. Olsson shows this problem with a memory example:

> "Suppose Peterson finds himself having this and that memory, wondering whether he can rely on them at large, rationally and practically. According to Lewis, Peterson may assume that whatever the degree of individual credibility may be, so long as it is positive there is always a degree of congruence sufficient to boost probabilities up to a level where the contents of the memories can be accepted as true. But learning this is of no help to Peterson if, as Lewis would insist, he has no clue what the actual degree of individual credibility is. Purged of this knowledge, he will not be able to determine whether the actual degree of congruence of his memories is high enough to warrant acceptance of their contents." (Olsson, 2005, p. 57)

The problem might not hit a coherence account directly[33], but it certainly effects any filtering by coherence because we need to assign an amount of justification based on coherence and C. I. Lewis' account makes this impossible. In the best scenario, we should be able to keep a score on how coherent an explanation candidate is and dismiss any below a certain score. His account of congruence does not provide us with enough tools to accomplish this. We do not know the initial probabilities and even if we fix that somehow, we would still need a grasp on how the initial probabilities have to be combined and processed. It is easy to see that the deck being never shuffled becomes likelier, but it is difficult to convert this change in likelihood into a computable formula. We have no idea how to assess the increase in probability. Nevertheless, C. I. Lewis is a good starting point. If we can get rid of assigning

[33] They can argue that it is only an epistemic problem.

initial probabilities and find a way to make the process computable, we would be able to rate the coherence of explanation candidates with a given set of accepted propositions.

A step into this direction may be BonJour's complex coherence account. Although BonJour does not give a precise definition of coherence because "[…] the task of giving an adequate explication of the concept of coherence is not uniquely or even primarily the job of coherence theories" and "[…] the absence of an adequate explication of coherence does not count against coherence theories any more than against their rivals" (BonJour, 1985, p. 94), he does give a schema of five fundamental criteria:

> (1) A system of beliefs is coherent only if it is logically consistent.
>
> (2) A system of beliefs is coherent in proportion to its degree of probabilistic consistency.
>
> (3) The coherence of a system of beliefs is increased by the presence of inferential connections between its component beliefs and increased in proportion to the number and strength of such connections.
>
> (4) The coherence of a system of beliefs is diminished to the extent to which it is divided into subsystems of beliefs which are relatively unconnected to each other by inferential connections.
>
> (5) The coherence of a system of beliefs is decreased in proportion to the presence of unexplained anomalies in the believed content of the system.
>
> (BonJour, 1985, pp. 95-99)

The criteria implement many of the ideas already presented in Ewing and C. I. Lewis. They require consistency and some sort of connection. Any logically independent subset Z in the complete set Y diminishes the coherence a lot, similar to the restriction Ewing had in place. Also, like C. I. Lewis, BonJour does not require that every element is entailed by the remainders. The new addition of BonJour is the role of explanation to coherence. He takes explanations to be covering-law ones, like Hempel, and therefore subsumes their positive part for coherence into inferential connections. Unexplained propositions, however, get their own criteria and decrease coherence. (BonJour, 1985, p. 99) This catalogue of criteria accomplishes a lot of what we want in a coherence account. It has the requirement of consistency and it uses positive and negative relations between the elements to assign

coherence to a set. With the introduction of explanation as a relevant factor for coherence, we have a natural access to determine the relevant relations for our explanation candidates. Nothing relies on the covering-law model in BonJour's coherentism. We may give potential explaining an increase of coherence as a unique criterion instead of subsuming it into the presence of inferential connections to differentiate them from logical connections.

One of the biggest steps into the right direction in BonJour's account is the drop-out of probability. C. I. Lewis used likelihood to define congruence, but that was not much help for us, because we want to use coherence as a way towards finding out which explanation candidate is the most probable. We do not know any of their probabilities beforehand. However, we know logical relations and potential explanatory connections. We want to use those to determine coherence and to be able to work with coherence to determine likelihood afterwards.[34] There is still an open question how to compute coherence with a view similar to BonJour's. Fortunately there is an even more refined account in this fashion that is made for computation: the coherence view of Paul Thagard (2000; 1989).

2.2.2 A Computable Account of Coherence

The basic move to develop a criterion-like coherence account into a computable version is to transform it towards a constraint satisfaction notion of coherence. Following Thagard and Verbeurgt (1998), we can understand coherence in terms of maximal satisfaction of multiple constraints. Basic elements representing propositions, concepts, beliefs, actions and so on, can be related to another based on fitting together (cohere) or resist fitting together (incohere). We can work with different coherence relations: explanation, deduction, association, semantic relatedness, similarity, purpose and so on. Those relations are positive constraints between elements. Negative constraints are inconsistency, competition or negative association. Elements are divided into accepted and rejected ones based on those constraints. (Thagard & Verbeurgt, 1998) (Thagard, 2000) Constraints work only binary, but this is no problem, as nonbinary constraints can be transformed into binary ones.[35] This

[34] We need to discuss later if this last step is even possible.
[35] For example via hidden variables or dual-graph-translations. Cf. Bacchus & van Beek (1998)

multitude of positive and negative constraints keeps the advantages of BonJour's criterions in place, while it is better suited for computation. Thagard's aim is slightly different from ours. We want a measurement to score the coherence of explanation candidates, whereas Thagard concentrates on coherence problems as ordering a set of elements together with a set of constraints into accepted and rejected elements.[36] Those are formally defined as follows:

> "[…] Let E be a finite set of elements $\{e_i\}$ and C be a set of constraints on E understood as a set $\{(e_i, e_j)\}$ of pairs of elements of E. C divides into C+, the positive constraints on E, and C-, the negative constraints on E. With each constraint is associated a number w, which is the weight (strength) of the constraint. The problem is to partition E into two sets, A and R, in a way that maximizes compliance with the following two coherence conditions:
>
> 1. if (e_i, e_j) is in C+, then e_i is in A if and only if e_j is in A.
> 2. if (e_i, e_j) is in C-, then e_i is in A if and only if e_j is in R.
>
> Let W be the weight of the partition, that is, the sum of the weights of the satisfied constraints. The coherence problem is then to partition E into A and R in a way that maximizes W." (Thagard & Verbeurgt, 1998, p. 3)

The computation can be done with different methods: An exhaustive search algorithm that considers all possible solutions; an incremental algorithm that considers elements in arbitrary order; a connectionist algorithm using an artificial neural network; a greedy algorithm that uses local optima to approximate global optima; a semidefinite programming algorithm (Thagard & Verbeurgt, 1998, p. 7).

However, this problem, which is a variant of a constraint satisfaction problem, is known to be NP complete. This means that the effort to solve increases exponential in relation to the problem size and it is unlikely (unless P = NP) that there will ever be an effective algorithm to solve it optimal. These kind of problems are also called intractable and brute force algorithms, which enumerate all possible solutions, are in general not usable because of their excessive runtime. In addition, also incremental algorithms are unreliable, as they highly depend on the starting point of the search.

[36] This is not entirely true. Part of our coherence measurement is that we want the set of rejected elements to be empty.

Greedy, semidefinite and connectionist algorithms provide good approximations in reasonable time, according to Thagard and Verbeurgt (1998). The psychologically most plausible of them is the connectionist algorithm. We do not need to be concerned with psychological plausibility, nevertheless we will settle for the connectionist approach out of pragmatic reasons. Thagard (2000) provides a good overview on coherence modelled via neural networks and has already experimented with inference to the best explanation. A connectionist algorithm for coherence was proposed in the program ECHO (Thagard, 1989).[37] We build up on his research for our purpose. An example: Suppose a bottle of water is on my table (E1). It is table water (E2) and I think I did not bring it (E3). Now assume I have two different hypotheses how the water came to be in this position. Either my brother brought it with him (B1) and he loves table water (B2); or I did and forgot afterwards (I1), but I hate table water (I2). We use this as input for ECHO with the positive constraint of explanation and the negative of contradiction (data assigns the initial activation to evidence). When the network settles, we can find positive activation at accepted elements and negative activation at rejected elements.

Evidence:
E1 - A bottle of water is on my table
E2 - It is table water
E3 - I think I did not bring it
Hypotheses:

Brother:
B1 - My brother brought the bottle with him
B2 - My brother loves table water

Me:
I1 - I brought the bottle and forgot afterwards
I2 - I hate table water

'Brother' explains:
explain ((B1) E1)
explain ((B1 B2) E2)
explain ((B1) E3)

'Me' explains:
explain ((I1) E1)
explain ((I1 I2) E3)

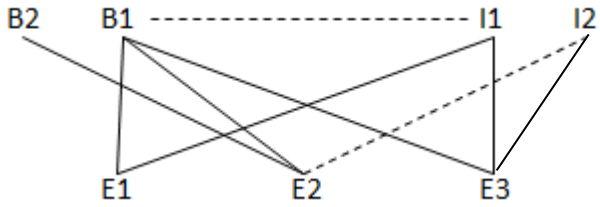

Figure 1 – A simple connectionist network. Dotted lines are inhibitory links (negative constraints), solid lines excitatory links (positive constraints).

[37] Source codes for ECHO and similar algorithms for practical reasoning can be found online at http://cogsci.uwaterloo.ca/ (Feb. 4, 2013). Technical details of ECHO were first published in Thagard (1989)

Data:
data (E1 E2 E3)

Contradictions:
contradict (B1 I1)
contradict (I2 E2)

The result is a rejection of I1 and I2. The most coherent set of elements in these circumstances is {E1, E2, E3, B1, B2}.

There are a couple of problems with the account in this example. First and foremost, it does something different from what we aim for. ECHO in this example decides between two competing hypothesis, but we want to filter out incoherent explanation candidates based on a given set of accepted propositions. Can ECHO do something like this as well? It can. If we drop B1, B2, and all relations involving them, we can test the coherence of the network with the elements {E1, E2, E3, I1, I2}. This results in I2 being rejected. We can interpret this as the complete tested set being incoherent (just like intuitively expected). Another way to assign coherence to a total set of element via ECHO is to measure the goodness-of-fit (harmony) of the neural network, which is defined as function $f = \sum_j w_{ij} a_i(t) a_j(t)$ (Thagard, 2000, p. 38). Harmony measures how well the activations match the weights.[38] However, this has to be modified to cancel the effects of large networks. For this purpose Thagard provides a mean-goodness function reflecting the total number of elements and incorporating the distance between elements (e.g. the mean-goodness of {E1, E2, E3, I1, I2, B1, B2} is 0.011155972). We can derive a basic scheme to filter based on coherence by goodness-of-fit > x, no rejections and no inconsistency. We want no inconsistency in a set as a minimal requirement for all our explanation candidates. Moreover, we want no rejections because our explanation candidates come in sets of propositions and stop functioning if one of them is rejected. We also do not want to reject any already accepted proposition/evidence (due to sciences or intuition) solely based on our explanation candidates. And at last, we want the total mean-goodness-of-fit to be greater or equal some value x, which gives us a minimal coherence condition.

[38] Hence elements with negative activation can still provide high harmony.

We still need to discuss which constraints we should actually employ, which connectionist algorithm works with them and which accepted propositions/evidence we should use to assign coherence. Before that, we have to discuss a general worry that a connectionist algorithm cannot reliably provide the globally maximal coherence, but only local maxima. Elijah Millgram (2000) brought up this problem. As already mentioned, coherence problems are NP-complete. Thus, solving them in reasonable time is not feasible due to their exponential complexity in relation to the problem size. The usual strategy when being confronted with a NP-complete problem is to work with approximations. This is the way the constraint satisfaction account of coherence works in computation. ECHO does not find the absolute best solution, but rather waits until the activations settle to a local optimum in coherence. Millgram's criticism focuses on this difference. He argues that approximating maximum coherence does not necessarily go together with approximation of the best solution, because the second most coherent theory may be drastically different from the most coherent theory. Not in terms of coherence, but in terms of their content. Copernicus theory of the movement of the planets might be close in coherence to Ptolemy's, but the theories are not even remotely similar. (Millgram, 2000, pp. 86-88) The suggested conclusion is that coherence approximations do not work for philosophical (scientific) needs. We do not aim for maximum coherence just for coherence's sake; we rather use it as a tool to grasp the likelihood of a theory being true. If two theories can be similar in coherence but dramatically different in content, the tool is likely to lose its usability.

Responding this issue, there is a two-fold answer available. Thagard argues first that the alternative too coherence based theory choice, namely Bayesian reasoning, is flawed in the same way. Bayesian inference is not computationally tractable as well. (Thagard, 2012, p. 45) This reply does not provide any good reason to stick to coherence as a tool. The second response fares better because it tries to show a connection between coherence and truth. It uses the following argument:

> (Premise 1) Scientists have achieved theories that are at least approximately true.
> (Premise 2) Scientists use explanatory coherence.
> (Conclusion) So explanatory coherence leads to theories that are at least approximately true.
> (Thagard, 2012, p. 46)

If we accept that scientific theories are approximately true and that scientists use explanatory coherence for their theory choice, then we can infer that explanatory coherence is a reliable guide to approximate truth. The argument is valid, but its soundness may be contested. In favour of P2, we can point to scientific practice in general. Thagard himself notes that one can support P2 by giving computational models of historical scientific theory choices.[39] This reasoning is not too compelling, because the computational models are always created when the historical decision has been already done and accepted. Hence, those models are possibly biased. P1 can be defended by pragmatic considerations. Scientific theories provide accurate predictions and are the basis of our reliable technology. These are good reasons for the argument being sound. Nevertheless, they can at best show that theory choice by coherence works in the long run. It is still impossible to conclude whether a single instance of it does lead to truth. Fortunately, this does not need to worry us too much. If we use coherence as a filter, we are not committed to picking a single theory out of all possible, we rather rule out incoherent and therefore likely false theories. We do not choose one of the explanation candidates solely on coherence, therefore Millgram's problem of choosing a false, but close to maximal coherent theory does not affect us. We only need a general reliable connection from coherence to truth.

2.2.3 Constraints

The argument by Thagard, mentioned in the previous paragraph, did more than just establishing the required connection of explanatory coherence and truth. It naturally suggests using explanation as a positive constraint for coherence. We need to be cautious, because Thagard's uses an unanalysed notion of explanation. We will use potential explanation as a positive constraint.[40] Thagard describes explanatory coherence with the following principles:

> **Principle E1: Symmetry** Explanatory coherence is a symmetric relation, unlike, say, conditional probability. That is, two propositions p and q cohere with each other equally.

[39] For simple examples see Thagard (1989).
[40] For simplicity we will write only 'explain' and 'explanation' in the algorithm. See the ECHO example above.

Principle E2: Explanation (a) A hypothesis coheres with what it explains, which can either be evidence or another hypothesis. (b) Hypotheses that together explain some other proposition cohere with each other. (c) The more hypotheses it takes to explain something, the lower the degree of coherence.

Principle E3: Analogy Similar hypotheses that explain similar pieces of evidence cohere.

Principle E4: Data Priority Propositions that describe the results of observations have a degree of acceptability on their own.

Principle E5: Contradiction Contradictory propositions are incoherent with each other.

Principle E6: Competition If p and q both explain a proposition, and if p and q are not explanatorily connected, then p and q are incoherent with each other (p and q are explanatorily connected if one explains the other or if together they explain something).

Principle E7: Acceptance The acceptability of a proposition in a system of propositions depends on its coherence with them.

(Thagard, 2000, p. 43)

E1 might be surprising on first sight. If we understand potential explanation (being an explanation candidate) as providing (supposed) information on causal history, symmetry is not implied. The symmetry constraint here does not need to be understood as two propositions being mutual explanations. Symmetry refers to the propositions hanging together by virtue of the potential explanation. This connection of fit is symmetric, even though any potential explanation itself is not a symmetric relation. Hence "'explanatory coherence' is a symmetric relation" (Thagard, 2000, p. 43).

E2 is straightforward. A hypothesis coheres with what it potentially explains.

E3 is unfortunately vague and Thagard does not say much about it. He claims that it is in fact used in scientist's theory choice (for example by Darwin). (Thagard, 1992, p. 68) The claim is reasonable, but it does not establish a coherence relation. We can drop E3 until we find a good reason to accept it.

E4 is especially important and creates a *foundherentism*, a mix of coherentism and foundationalism. This principle provides the basis for the implementation of already accepted propositions. Those can be accepted by foundational epistemic sources, e.g. perception, or introspection. Thagard gives those propositions that fall into E4 some degree of acceptability, but nevertheless they can be rejected. The amount of credence we should give to the accepted propositions depends on the acceptability criteria. If we only count propositions we are certain to be true as data (accepted propositions), then we can give them a very high initial acceptability. We can put the bar even higher if we require that no hypothesis contradicts the data. If anything goes as data, the initial acceptability should be very low. Our aim in this thesis is to move towards an inference to the best explanation based on scientifically accepted propositions and intuitions. For this purpose we set the bar as high as possible. We drop any hypothesis that contradicts science based propositions or rejects data in general.

E5 is again straightforward. Contradictory sets are incoherent and therefore provide a negative constraint. Contradiction as a negative constraint does not refer only to logical contradictions, but also to those which Thagard calls *pragmatic*. Pragmatic contradictory are explanatory unconnected propositions with a very unlikely conjunction. (Thagard, 1989, p. 438)

E6 is needed for choosing between different hypotheses with a given set of data. While it is in principle possible to build a single algorithm that lets all explanation candidates compete with each other, such a move would be unreasonable. We do not want to decide everything by coherence competition, but rather use coherence as a filter before looking at other explanatory virtues.

E7 is obvious. We accept what coheres well and reject propositions that do not.

All those principles are implemented in ECHO, but E3 and E6 are not necessary for the computation of a coherence problem.

Thagard (2000) proposes at least four other kinds of coherence with their own principles: analogical, deductive, visual and conceptual coherence. Our problem resembles theory

choice problems and therefore is best done with explanatory coherence. It is also favourable for our goal of ruling out inapt explanation candidates based on their potential explanation's (lack of) fitting to our accepted propositions. Furthermore it is the only kind of coherence that has a sufficient relation to truth, even though only established by an indirect argument.[41]

2.2.4 Accepted Science-Based Propositions

What propositions should be used as data? We want to use scientific data and intuitions, beginning our discussion with the first. Scientific data and intuitions are not independent. Science involves interpretation and theory-ladenness. Intuitions are related to beliefs formed on scientific evidence. Sciences will in most cases[42] not provide knock-down arguments for any theory of consciousness, but they can point against one. This fits the nature of negative constraints, as they do not only involve logical contradictions, but also explanatory unconnected propositions with a very unlikely conjunction. This improbability is not a fact, but rather based on a judgement that is itself often connected to intuitions. The difference, however, is that the science-based propositions we want to use are not determined solely by intuition, but rather have a foothold in empirical observations. Having said that, we can continue with discussing how we can select the science-based propositions for our coherence test.

In a perfect world we would start by employing every proposition that is part of or implied by a sufficiently tested scientific theory. However, that is certainly not feasible. We neither have the sources, nor abilities, nor time to gather all of those theories. One option is to use only a limited set of science-based propositions, but picking out a certain set may bias the coherence measurement in favour of some explanation candidates.

A more promising approach is to use any negative constraining science-based proposition and collect anything else into a single element. Let us call this element 'remainder element'. An element that is a negative constraint for explanation candidate A will be also used as data for explanation candidate B, even though B is positively connected to the same element. This has to be done to establish sameness of data elements for all coherence measurements.

[41] See above.

[42] It can get very close though. E. g.: A theory claims that all consciousness is caused by a certain brain region x. Then a person has x removed due to surgery and there are no noticeable changes. While it is possible to interpret the situation in favor of the theory, it does not seem to be a rational.

Without this condition, the mean goodness-of-fit would not be comparable. The switch from countless propositions of sciences to the contradicting plus remainder element does not come for free. We lose detailed explanation-coherence with sciences and use them mostly as condition for consistency (logical and pragmatic).

We can also use the 'remainder plus some relevant propositions'-attempt in a positive way. In this approach we pick data based on providing positive constraints and then collect the rest in a remainder element. Alternatively we could also use both negative and positive constraining elements and a remainder element. Even though this option may be the most promising one, we will stick to using only the negative constraining elements for pragmatic reasons. Providing all elements with positive constraints is nearly as difficult as using all science based propositions and presents analogue problems. We lack sources, ability and time to provide them. Using more elements also comes at the price of longer computation. Hence, we have another pragmatic reason for only negative constraining scientific elements.

For the examples of our coherence measurement, we will use the explanation candidates provided in 2.1. Negative constraints to those are, for example, the following:

(C1) Various beings, that we attribute a single type of phenomenal mental state to, differ in biological/neural/physical properties.

(C2) The pineal gland is understood as part of the endocrine system mainly functioning as a source for the hormone melatonin and regulating the biological rhythmic of sleeping/awakeness cycles. For the first paper published on the glands melatonin production cf. Lerner et al. (1960). For the relation to day/night cycles cf. Reiter (1980).

(C3) There are no observed instances of mental properties without correlating physical properties.

(C4) There is no evidence for mental properties of single physical ultimates. Consciousness seems to be a feature of complex and structured nervous systems.

2.2.5 Accepted Intuition-Based Propositions I

Why intuitions? Philosophers use intuitions as a tool for theory choice frequently.[43] Most famous is probably the employment of intuitions by Edmund Gettier (1963) to dismantle the justified true belief account of knowledge. However, do we have good reasons to assign such a pivotal role to intuitions? There are a couple of ways to defend an epistemic weight of intuitions. One we discuss here treats intuitions as evidence.[44] We do not discuss intuitions as beliefs or dispositions to believe. Furthermore, we need to distinct two aspects of intuitions: intuit*ings* – the fact that a person has an intuition; and intuit*eds* – the content of the intuition (Pust, 2012). The mere intuitings are usually not interesting. Only if we take the propositional content as part of the evidence it can justify any belief.

The justificatory relevance can be reasoned by a sort of phenomenal generativism. If we accept what Michael Huemer calls 'the principle of phenomenal conservatism' (PC), we are to some degree justified to believe that p, just because it seems to us that p (Huemer, 2007, p. 30). The idea behind this is to model all kinds of epistemic boost on perception. When we visually perceive any object x, we have at least some justification that there is x (if there are no defeaters). Appearing to us that p justifies our believing that p.[45] Appearance or seeming provides a propositional attitude that is part of perception, introspection, apparent memory, and, most importantly for us, intuitions. When we have an intuition that p, it seems to us that p. This fares well if we are internalists on justification and handle intuiting that p as a reason for believing that p. If we demand an externalist version of guidance towards truth, e.g. reliability, this is not satisfying. We need an argument in favour of the reliability of our seeming that p. This has, as far as we know, not been presented. Nevertheless, the way is already pointed out. Colleen M. Kelley and D. Stephen Lindsay (1993) tested the felt ease of retrieval for answers to general knowledge questions. Feeling easier retrieval correlated with more confidence in the answer and in most cases, higher reliability. However, various factors can lead to errors and misplaced confidence. Those experiments are neither exhaustive nor

[43] We did so too when committing to the intuition that there is an accessible answer to the hard problem of consciousness.

[44] There are numerous attacks on the epistemic relevance of intuitions. A common argument against any such relevance is the empiricist explanationist skepticism about intuitions (Pust, 2000), put forward against moral intuitions (Harman, 1977), metaphysical intuitions (Goldman, 1987; 1992) and semantic or logical intuitions (Stich, 1975). Other attacks on intuition as a source of knowledge aim at the reliability of intuitions itself, or our justification of any reliability claim toward intuition. (Pust, 2012)

[45] The justification for PC is again intuition. PC is this sense circular. Circularity, however, is not a problem itself.

conclusive, but they show what needs to be done. The phenomenal seeming that p has to be connected (in the right circumstances) to p being true.

Another (not exclusive) take on using intuitions as evidence claims that they are result of explicating our doxastic and representational system. Let us assume that I possess the concepts ‘red’, ‘blue’ and ‘apple’. ‘Red’ is connected to ‘apple’, but ‘blue’ is not. Now I am confronted with Anna who reports that she saw an apple like plastic sculpture. If I guess the colour of the plastic sculpture I will certainly rather say it is red, than that it is blue. The idea portrait by this example is that concepts are mental representations in one’s head which are related to each other. An intuition is something that either is entailed or just fits well with the existing concepts. It “[…] bears evidentially on […] personal concepts”, as Goldman (2007, p. 13) claims. If the content of a concept C implies that C applies to x, then the bearer of C is disposed to intuit that C applies to x (when triggered by any input). If the possessed concepts are adequate (in regard to the world), the intuitions related also tend to be true. One might, however, have false information, inadequate concepts or simply be careless and thus be an unreliable intuitionist. False theories about concepts can interfere as well. (Goldman, 2007, p. 15) We can therefore, at least in principle, explain why an intuition is wrong. This enables us to use potential explanation constraints in our coherence measurement between theories and intuitions, even though the theory claims the intuition is false.

In general, intuitings will be reliable, given adequate and correct concepts. We can also see how intuitions are brought about. We can notice a causal relationship between personal concepts and intuitions based on those concepts. (Goldman, 2007, p. 16)

We now only need to secure satisfying concepts and are ready to use intuitions as data in our coherence measurement. Hence, we need to move from personal concepts to shared concepts. We need agreement on concepts and thereby on intuitions. This increases the likelihood of our concepts being adequate by having them tested by each individual. If the application of a concept is inadequate, the (reasonable) person will modify that very concept and therefore produce disagreement. Due to this process it is probable that a concept is

adequate when there is little to no disagreement on it. Ernest Sosa (2007) argues that we should privilege some intuitioners because they acquire a certain competence and become experts. It is, however, difficult to find and argue for reasonable criteria of expertise in intuitioning. And we thereby better not lose the starting point of philosophical inquiry – our folk concepts.[46] Hence, we should deny such a requirement of expertise. We should also not demand perfect agreement, because it is seldom the case.

2.2.6 Accepted Intuition-Based Propositions II – Philosophical Agreement

Having argued for intuitions as part of the accepted propositions, we need to discuss which intuitions in particular can be used. We certainly want intuitions that are predominantly agreed upon. In general, we have at least two different ways to gather those: either by looking at the philosophical discourse, or by experimental approaches. Similar to the impossibility to include all science based propositions, we cannot use all intuitions. Consequently we limit ourselves to those relevant to the hard problem of consciousness that are already proposed in the discourse. We include any intuition (which is predominantly agreed upon) that is reasonably put forward as impacting the debate (and does not closely resemble another intuition). Certainly agreement towards intuitions based on common thought experiments (‘Mary’ (Jackson, 1982), ‘Zombies’ (Chalmers, 1996), ‘inverted qualia/earth’ (Block, 1990)) can be used. Other thought experiments can be used under certain conditions. If we accept a broader notion of qualia or phenomenal consciousness[47], we can use a wider array of such thought experiments (e.g. ‘Mary on plus and quus’ (Goff, 2012)).

Let us investigate Frank Jacksons ‘Mary’:

> “Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on. She discovers, for

[46] Goldman agrees to some extent. (Goldman, 2007, p. 19)
[47] E.g. accept cognitive phenomenology.

example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'. (It can hardly be denied that it is in principle possible to obtain all this physical information from black and white television, otherwise the Open University would of necessity need to use colour television.) What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she learn anything or not?“ (Jackson, 1982)

The case of Mary is well discussed in philosophical literature. Jackson initially proposed it as part of an anti-physicalist argument (the knowledge argument). He argued that Mary intuitively learns something new. Something not incorporated by physiscs. The argument is arguably unsound and maybe not even valid. The most common responses are as follows[48]:

1. One may deny that Mary learns something new. E.g. Churchland (1985)
2. One may argue that what is learned is no propositional knowledge, but ability. E.g. Nemirow (1980) and Lewis (1983)
3. One may argue that what is learned is no propositional knowledge, but rather new acquaintance. E.g. Conee (1994)
4. One may argue that Mary learns something she already knew in a different type. E.g. Horgan (1984b) and Tye (1995)
5. One may argue that Mary learns only in a fine-grained sense. E.g. Loar (1997)

The important point for us here is that the critics mostly accept the core intuition that Mary learns something. As soon as this intuition is accepted, any theory of mind has to deal with it. We will use the accepted intuition:

(I1) We have the intuition that: In a situation similar to what Frank Jackson tells about Mary, Mary learns something new.

[48] A good overview is provided by Van Gulick (2004)

We can interrogate the possibility of philosophical zombies similarly. Those are used in the conceivability argument, recently most prominently defended by David Chalmers (1996). A philosophical zombie is a complete physical duplicate of a human being, but it lacks any phenomenal consciousness. Now Chalmers argues with the possibility of a zombie world. His complete anti-materialist argument is as follows (Chalmers, 1996, p. 123):

1. In our world there are conscious experiences.
2. There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold.
3. Therefore, facts about consciousness are further facts about our world, over and above the physical facts.
4. So materialism is false.

The second premise is pivotal and established via an inference from conceivability to possibility.[49] We should, however, discuss how conceivability comes into play. The ability to represent something as possible is closely related to intuition. Usually, if we think about the conceivability of something we employ no logical deduction, but rather intuit if it seems possible or not. Establishing a premise by conceivability is therefore to some extent intuition in disguise. It is not completely intuition based, because the transition from conceivability to logical possibility has a supporting role that can be contested. Even if an inference of this kind is accepted in general, there are restrictions that can lead into failure. One may have defeaters for the possibility of something supposedly conceivable. A defeater of this kind shows how a person S happens to think x is conceivable, even though this is actually false. (Yablo, 1993, p. 37) The usual criticism on the conceivability argument focuses on this connection from conceivability to possibility. Opponents deny either that this inference is reliable in general, or that it is not applicable to this case (e.g. Hill (1997), Hill and Mclaughlin (1999), Loar (1999)). However, they rarely deny the conceivability of zombies (e.g. Dennett (1995) and Shoemaker (1999)). Hence we will use the predominant intuition:

> (I2) We have the intuition that: physical duplicates are not necessarily phenomenal duplicates.

[49] In favor of which has been argued by Chalmers (2002)and Yablo (1993).

There is a connected argument for ghosts, which is arguably stronger. Philip Goff (2010) proposes that a being that is wholly determined by state of consciousness – a ghost – is intuitively possible and even easier to accept. The reason for this is the natural resistance to doubt, which consciousness involves as argued by Descartes. If scepticism towards an outer world being true is imaginable, then such philosophical ghosts are imaginable. Unfortunately, there is little discussion of philosophical ghosts, but relying on the zombie debate we should also accept the ghost intuition:

> (I3) We have the intuition that: phenomenal duplicates, which are not physical duplicates, are possible.

The fourth thought experiment we will look at is the scenario of inverted qualia.[50] Imagine a person S_1 that has the phenomenal experience 'red' when light with a frequency of w_1 hits the eye. S_1 uses the word 'red' when exposed to such an experience. S_1 has also the phenomenal experience 'blue' when light with the frequency w_2 hits the retina and uses the term 'blue' accordingly. Now suppose another person S_2. S_2 has inverted qualia: when exposed to w_1 light he experiences 'blue' and with w_2 'red'. Nevertheless, he uses the term 'red' at w_1 and 'blue' at w_2, just like S_1. Both persons are behaviourally perfectly alike. Is this scenario possible?

Inverted qualia scenarios come in different variety and usually aim at dismantling behaviourism, functionalism and various other kinds of reductive and non-reductive physicalism.[51] They work just like the argument from possible, philosophical zombies in pumping an intuition that a scenario not smoothly explained by physicalism is possible. As with the above thought experiments, there are only few, who deny the intuition in question (e.g. Harman (1990)). Most try to reconcile inverted qualia with some kind of physicalism (e.g. Tye (1994), Horgan (1984a), Campbell (2004)). We will use the core intuition:

> (I4) We have the intuition that: behavioural and functional duplicates, which are not phenomenal duplicates, are possible.

[50] Cf. Block (1990)
[51] Cf. Byrne (2010)

2.2.7 Accepted Intuition-Based Propositions III – Experimental Philosophy

Experimental philosophy is a very young field, but may be suitable for our purpose. The main goal of this approach is to investigate the intuitions of non-philosophers with regard to philosophical problems and questions. It is, in some sense, a way to grasp folk concepts. As such it is a good candidate as a tool to select intuitions that are predominantly agreed upon. Unfortunately, there is very little research done with methods of experimental philosophy in the field of phenomenal consciousness. Research done by Joshua Knobe and Jesse Prinz (2008) investigated people's mental state ascriptions. They aimed towards the question whether ordinary people – those who never studied philosophy – have a concept of phenomenal consciousness and use it in state attributions. The first experiment tested in what circumstance people ascribe phenomenal states. In particular, they tested whether people would attribute the following states to a company: intends…, decides…, tries…, wants…, believes…, hopes…, loves…, hates…, feels depressed, experiences joy, feels happy, feels pain, feels angry, and feels scared. Persons usually did not attribute any 'feel x' or 'experience x' properties, but they were less reluctant to ascribe an intentional state like 'wants x'. (Knobe & Prinz, 2008, pp. 73-74) In further studies this distinction between phenomenal states and merely intentional states was confirmed and further related to moral judgements. Knobe and Prinz concluded that there is a folk concept of phenomenal consciousness, which is not used purely for behavioural prediction and analysis, but rather connected to moral judgments. (Knobe & Prinz, 2008, pp. 74-82) If they are correct, then we can derive an agreed upon intuition that phenomenal consciousness is at least distinct from behavioural dispositions.

However, there are also studies coming to the converse conclusion. Justin Sytsma and Edouard Machery (2010) argue that in the folk conception no distinction between phenomenal and any other kind of consciousness is observable. The criticism targeted at Knobe and Prinz points towards a problem in their studies. Knobe and Prinz tested mental state ascription on Corporations, which are behavioural and functional different from single individuals. (Sytsma & Machery, 2010, p. 304) Indeed it would be a better test if there were no behavioural differences as a potentially deciding factor. Such a study has been carried out by Sytsma and Machery[52]:

[52] A similar study is discussed by Huebner (2010).

> "In the first experiment, participants were given a description of a relatively simple, non-humanoid robot performing tasks expected to elicit ascriptions of either a perceptual experience (seeing red) or a bodily sensation (feeling pain) in humans. Given the role of visual experiences and bodily sensations such as pain in spelling out the philosophical concept of phenomenal consciousness, we hypothesized that philosophers' attribution of the perceptual experience would not dissociate from their attribution of the bodily sensation. If the folk conceive of subjective experience in the same way as philosophers, they should give a similar pattern of answers. A total of 671 participants completed surveys through the Mental State Intuitions Study website. Participants were divided into two groups based on their responses to biographical questions concerning their philosophical training. They randomly received one of four scenarios describing an agent manipulating a red object. The scenarios were varied along two dimensions. In half of the scenarios, the agent was a simple robot (Jimmy), while in the other half, the agent was a normal human male (Timmy). Moreover, in half of the scenarios, the manipulation was successful and participants were asked whether the agent (either the robot or the human) ''saw red'' on a 7-point scale anchored at 1 with ''clearly no,'' at 4 with ''not sure,'' and at 7 with ''clearly yes.'' In the other half, the agent was electrically shocked and participants were asked whether the agent ''felt pain'' on the same 7-point scale. In each pair of vignettes the agents are described in behaviourally identical ways." (Sytsma & Machery, 2010, p. 305)

The results for 'pain' were as expected: both philosophers and non-philosophers ascribed it to the human but not to the robot. 'Red', however, was attributed to the robot by non-philosophers way more frequent than by philosophers. (Sytsma & Machery, 2010, p. 307) One may object that 'seeing red' is ambiguous insofar as there is a phenomenal and a purely informational sense of 'red'. Sytsma and Machery argue that this is not the cause of the divergence, because if it would be so, the amount of non-philosophers claiming that the robot sees red ought to be lower than it is. (Sytsma & Machery, 2010, pp. 309-311) This is hardly convincing, because for this line of argument to work, non-philosophers with two concepts of 'red' need to choose equally often either of the two. If non-philosophers use the informational sense with higher priority, but nevertheless have the phenomenal notion in

store, the study becomes inconclusive. Nevertheless it puts doubt on Knobe's and Prinz's conclusion.

A study by Huebner et al. (2010) shows, that there might be even more problems with the phenomenal mental state ascription test. Knobe and Prinz studied attributions to collectives – e.g. corporations – in an environment that has a particularly individualistic culture. Huebner et al. completed a study similar to Knobe and Prinz, but with two groups of test persons. One group of students based in the University of North Carolina and the other from Chinese University of Hong Kong. Interestingly, the Chinese students ascribed phenomenal states to the group more frequently. (Huebner, et al., 2010, p. 235) However, the sample size of students in those tests is unfortunately very low; therefore it cannot completely invalidate the former results.

However, we can conclude that there is currently no convincing study that can be used for providing accepted intuitions. There is no a priori reason why such a study is impossible. It just has not been done yet and it is not our primary aim to provide one. Hence we will have to restrict ourselves to intuitions that are agreed upon in the philosophical discourse.

2.3 The Filter: Exemplifying the Usage

We will show how our filter mechanism works with explanation candidates put forward in 2.1, the data ("contradictions" (C1, C2, C3, C4) and intuitions (I1, I2, I3, I4)) and the remainder element (E). We rule out any explanation candidate that contradicts or rejects our data. We also dismiss any explanation candidate with a mean goodness-of-fit below 0.03[53] and those which include a rejected hypothesis. For this purpose we use ECHO with a calibration of: decay: .05; excitation: .04; inhibition: -0.06; data-excitation: .05 and can be interpreted as the example in 2.2.2. Each measurement uses the same data and, additionally, propositions and explanation-relations depending on the specific theory. The ECHO inputs are the following:

[53] This is relatively low, but should be sufficient to get rid of badly coherent explanation candidates. It will, for example, usually rule out explanation candidates with unconnected propositions.

Data used for all coherence measurements:

Intuitions:

I1 - We have the intuition that: In a situation similar to what Frank Jackson tells about Mary, Mary learns something new.
I2 - We have the intuition that: physical duplicates are not necessarily phenomenal duplicates.
I3 - We have the intuition that: phenomenal duplicates, which are not physical duplicates, are possible.
I4 - We have the intuition that: behavioural and functional duplicates, which are not phenomenal duplicates, are possible.

Contradictions:

C1 - Various beings, that we attribute a single type of phenomenal mental state, differ in biological/neural/physical properties.
C2 - The pineal gland is understood as part of the endocrine system mainly functioning as a source for the hormone melatonin and regulating the biological rhythmic of sleeping/awakeness cycles.
C3 - There are no observed instances of mental properties without correlating physical properties.
C4 - There is no evidence for mental properties of single physical ultimates. Consciousness seems to be a feature of complex and structured nervous systems.

Remainder:

E - Remainder element

2.3.1 Interactive Dualism (Descartes)

ID1 - There is physical reality.
ID2 - There is mental reality.
ID3 - Mental reality is not physical reality.
ID4 - Physical reality is not mental reality.
ID5 - Mental reality is fundamental.
ID6 - Physical reality is fundamental.
ID7 - Phenomenal reality is mental reality.
ID8 - Physical events can cause and be caused by physical events.
ID9 - Mental events can cause and be caused by mental events.
ID10 - Physical events can cause and be caused by mental events.
ID11 - Mental events cause physical events by interaction at the Pineal Gland.
ID12 - Physical events cause mental events via the senses and the Pineal Gland.

Interactive Dualism, positive constraints:

explain ((ID1 ID2 ID5 ID6) ID3)	;explaining mental =/= physical
explain ((ID1 ID2 ID5 ID6) ID4)	;explaining physical =/= mental
explain ((ID2) ID7)	;explaining phenomenal as mental
explain ((ID2) ID9)	;explaining inter-mental causation
explain ((ID1) ID8)	;explaining inter-physical causation
explain ((ID10 ID11) ID12)	;explaining mental-physical causation
explain ((ID1 ID2 ID3 ID4 ID7 ID10) I1)	;explaining Mary's learning vaguely
explain ((ID1 ID2 ID3 ID3 ID7 ID12) I1)	;explaining Mary's learning more concrete
explain ((ID1 ID2 ID3 ID4) I2)	;explaining physical, but not phenomenal duplicates
explain ((ID1 ID2 ID3 ID4) I3)	;explaining phenomenal, but not physical duplicates
explain ((ID1 ID2 ID3 ID4) I4)	;explaining behavioural and functional, but not phenomenal duplicates
explain ((ID1 ID2 ID3 ID4) C1)	;explaining differences in physics without differences in types of mental states.
explain ((ID3 ID4) C4)	;explaining no mental properties in physical ultimates
explain ((ID1) E)	;explaining remainder by physical reality

Negative constraints:

contradict ((ID11) C2)	;pineal gland function differs
contradict ((ID12) C2)	;pineal gland function differs
contradict ((ID9) C3)	;pure mental causation in question

The network settled by cycle 153 with the following activation results:

ID1	0.8331512		E	0.6187295
ID2	0.8270133			
ID3	0.8179777		I1	0.6588252
ID4	0.81929713		I2	0.6177471
ID5	0.5416653		I3	0.6177471
ID6	0.5416653		I4	0.6177471
ID7	0.618074			
ID8	0.3959448		C1	0.6177471
ID9	0.088558786		C2	0.687167
ID10	0.15597795		C3	0.46770072
ID11	-0.49209663	rejected	C4	0.61710566
ID12	-0.5655446	rejected		
Mean goodness-of-fit:		0.035431854		

Table 1 - Echo results for Interactive Dualism.

Interactive Dualism scores low because the constraint satisfaction rejects ID11 and ID12. It also is inconsistent with the scientific data C1 and C2.

Interactive Dualism (Descartes)	Consistency (with sciences)	x
	No Rejections	x
	Mean goodness-of-fit > 0.03	✓

2.3.2 Type Identity Theory

IT1 - There is a fundamental sense in which there is only one kind of stuff in reality.
IT2 - There is physical reality.
IT3 - There is phenomenal reality.
IT4 - Phenomenal reality is wholly physical reality.
IT5 - Phenomenal event types are identical with physical event types.
IT6 - On the fundamental level there are only physical properties.
IT7 - All facts are (fully) determined by facts about ultimates.
IT8 - Phenomenal consciousness is physical and is caused by physical events.

Type Identity, positive constraints:

explain ((IT1 IT2 IT3) IT4)	;explaining phenomenal = physical
explain ((IT1 IT2 IT3 IT4) IT5)	;explaining physical $=_{TYPE}$ mental
explain ((IT1 IT2 IT4) IT6)	;explaining fundamental properties
explain ((IT1 IT2 IT6) IT7)	;explaining smallism
explain ((IT1 IT2 IT4 IT5 IT6 IT7)(IT8)	;explaining phenomenal consciousness explicitly
explain ((IT2) E)	;explaining the remainder
explain ((IT2) C2)	;explaining pineal gland by physical reality
explain ((IT4 IT5) C3)	;explaining mental, physical correlation
explain ((IT6) C4)	;explaining no mental properties at ultimates

Negative constraints:

contradict ((IT5) C1)	;type-identity is incompatible with C1
contradict ((IT5) I1)	;type-identity is incompatible with I1
contradict ((IT5) I2)	;type-identity is incompatible with I2
contradict ((IT5) I3)	;type-identity is incompatible with I3
contradict ((IT5) I4)	;type-identity is incompatible with I4

The network settled by cycle 135 with the following activation results:

IT1	0.65104395		E	0.6068066
IT2	0.7296721			
IT3	0.48169562		I1	0.6427025
IT4	0.5793558		I2	0.6427025
IT5	-0.70866174	rejected	I3	0.6427025
IT6	0.6288432		I4	0.6427025
IT7	0.44484654			
IT8	0.23285642		C1	0.6427025
			C2	0.6068066
			C3	0.48169562
			C4	0.594444
Mean goodness-of-fit:		0.024975905		

Table2 - Echo results for Type Identity Theory.

Type identity theory fares in no way better than the implausible form of interactive dualism. It even has a lower mean constraint satisfaction (goodness-of-fit).

Type Identity Theory	Consistency (with sciences)	x
	No Rejections	x
	Mean goodness-of-fit > 0.03	x

2.3.3 Supervenience Physicalism

SP1 - There is a fundamental sense in which there is only one kind of stuff in reality.
SP2 - There is physical reality.
SP3 - There is phenomenal reality.
SP4 - Phenomenal reality is not wholly physical reality.
SP5 - Physical reality is not wholly phenomenal reality.
SP6 - In the actual world w it is true that: for any possible world w* if w* is a physical duplicate of w, then w* is a duplicate of w simpliciter.
SP7 - Any instantiated property is necessitated by some instantiated physical properties.
SP8 - Any phenomenal property is supervenient on physical properties.
SP9 - On the fundamental level there are only physical properties.
SP10 - All facts are (fully) determined by facts about ultimates.
SP11 - Experiential (phenomenal) events are caused by and supervene on physical events.

Supervenience, positive constraints:

```
explain ((SP1 SP2 SP3) SP4)          ;explaining phenomenal is not wholly
                                      physical
explain ((SP1 SP2 SP3) SP5)          ;explaining physical is not wholly
                                      phenomenal
explain ((SP2 SP7) SP3)              ;explaining physical - phenomenal
                                      relation
explain ((SP2 SP3 SP7) SP8)          ;explaining supervenience
explain ((SP2 SP3 SP8) SP6)          ;explaining global identity
explain ((SP2 SP7) SP9)              ;explaining fundamental ontology
explain ((SP2 SP9) SP10)             ;explaining smallism
explain ((SP2 SP3 SP8) SP11)         ;explaining causation of phenomenal
                                      consciousness
explain ((SP2 SP11) SP3)             ;explaining phenomenal reality
explain ((SP2) C2)                   ;explaining pineal gland by physical reality
explain ((SP2 SP8) C1)               ;explaining differences by supervenience
explain ((SP7 SP8) C3)               ;explaining observed correlation
explain ((SP9) C4)                   ;explaining no mental properties at
                                      ultimates
explain ((SP2) E)                    ;explaining the remainder
```

Negative constraints:

```
contradict ((SP6) I1)                ;global supervenience is incompatible
                                      with I1
contradict ((SP6) I2)                ;global supervenience is incompatible
                                      with I2
contradict ((SP6) I3)                ;global supervenience is incompatible
                                      with I3
contradict ((SP6) I4)                ;global supervenience is incompatible
                                      with I4
contradict ((SP7) I1)                ;Physical necessitation is incompatible
                                      with I1
contradict ((SP7) I2)                ;Physical necessitation is incompatible
                                      with I2
contradict ((SP7) I3)                ;Physical necessitation is incompatible
                                      with I3
contradict ((SP7) I4)                ;Physical necessitation is incompatible
                                      with I4
contradict ((SP8 )I1)                ;Supervenience is incompatible with I1
contradict ((SP8) I2)                ;Supervenience is incompatible with I2
contradict ((SP8) I3)                ;Supervenience is incompatible with I3
contradict ((SP8) I4)                ;Supervenience is incompatible with I4
```

The network settled by cycle 128 with the following activation results:

SP1	-0.5035596	rejected	E	0.26019254
SP2	-0.80544275	rejected		
SP3	-0.78254783	rejected	I1	0.79184914
SP4	-0.35415682	rejected	I2	0.79184914
SP5	-0.35415682	rejected	I3	0.79184914
SP6	-0.8083738	rejected	I4	0.79184914
SP7	-0.8393121	rejected		
SP8	-0.8491379	rejected	C1	0.25049716
SP9	-0.43901715	rejected	C2	0.26019254
SP10	-0.32820684	rejected	C3	0.2426854
SP11	-0.21391654	rejected	C4	0.39025262
Mean goodness-of-fit:		0.043261714		

Table4 - Echo results for Supervenience.

Supervenience does not fit with the intuitions at all, therefore it gets rejected. Notice that the mean goodness-of-fit is relatively high, because the activations are adequate to the weights. One might suspect that the triple of SP6-8 inflates the computation because they have they are so closely related that they just add unnecessary negative constraints. However, if we remove all contradictions involving SP7 or SP8, the result is pretty much the same.

Supervenience	Consistency (with sciences)	✓
	No Rejections	x
	Mean goodness-of-fit > 0.03	✓

2.3.4 Supervenience Physicalism with Ability Hypothesis

SP1 - There is a fundamental sense in which there is only one kind of stuff in reality.
SP2 - There is physical reality.
SP3 - There is phenomenal reality.
SP4 - Phenomenal reality is not wholly physical reality.
SP5 - Physical reality is not wholly phenomenal reality.
SP6 - In the actual world w it is true that: for any possible world w* if w* is a physical duplicate of w, then w* is a duplicate of w simpliciter.
SP7 - Any instantiated property is necessitated by some instantiated physical properties.
SP8 - Any phenomenal property is supervenient on physical properties.
SP9 - On the fundamental level there are only physical properties.
SP10 - All facts are (fully) determined by facts about ultimates.

SP11 - Experiential (phenomenal) events are caused by and supervene on physical events.
Ab1 - Theories about physical reality cannot wholly teach us about phenomenal reality.
Ab2 - In experiencing phenomenal reality we acquire an ability to recognize, imagine and predict ones behaviour by means of imaginative experiments. This ability cannot be exhaustively taught by a physical theory.

Supervenience Physicalism with Ability Hypothesis, positive constraints:

explain ((SP1 SP2 SP3) SP4)	;explaining phenomenal is not wholly physical
explain ((SP1 SP2 SP3) SP5)	;explaining physical is not wholly phenomenal
explain ((SP2 SP7) SP3)	;explaining physical - phenomenal relation
explain ((SP2 SP3 SP7) SP8)	;explaining supervenience
explain ((SP2 SP3 SP8) SP6)	;explaining global identity
explain ((SP2 SP7) SP9)	;explaining fundamental ontology
explain ((SP2 SP9) SP10)	;explaining smallism
explain ((SP2 SP3 SP8) SP11)	;explaining causation of phenomenal consciousness
explain ((SP2 SP11) SP3)	;explaining phenomenal reality
explain ((SP7 SP8) C3)	;explaining observed correlation
explain ((SP9) C4)	;explaining no mental properties at ultimates
explain ((SP4 SP5) Ab1)	;connecting ability hypothesis
explain ((SP4 SP5) Ab2)	;connecting ability hypothesis
explain ((SP6 Ab1 Ab2) I1)	;explaining Mary's learning
explain ((SP7 Ab1 Ab2) I1)	;explaining Mary's learning
explain ((SP8 Ab1 Ab2) I1)	;explaining Mary's learning
explain ((SP2) E)	;explaining the remainder

Negative constraints:

contradict ((SP6) I2)	;global supervenience is incompatible with I2
contradict ((SP6) I3)	;global supervenience is incompatible with I3
contradict ((SP6) I4)	;global supervenience is incompatible with I4
contradict ((SP7) I2)	;Physical necessitation is incompatible with I2
contradict ((SP7) I3)	;Physical necessitation is incompatible with I3
contradict ((SP7) I4)	;Physical necessitation is incompatible with I4
contradict ((SP8) I2)	;Supervenience is incompatible with I2
contradict ((SP8) I3)	;Supervenience is incompatible with I3

contradict ((SP8) I4) ;Supervenience is incompatible with I4

The network settled by cycle 156 with the following activation results:

SP1	-0.5346966	rejected	E	0.25856605	
SP2	-0.8089542	rejected			
SP3	-0.7878002	rejected	I1	-0.40047786	rejected
SP4	-0.60583526	rejected	I2	0.78587073	
SP5	-0.60583526	rejected	I3	0.78587073	
SP6	-0.74115896	rejected	I4	0.78587073	
SP7	-0.7966035	rejected			
SP8	-0.83716977	rejected	C1	0.25231025	
SP9	-0.43439215	rejected	C2	0.25856605	
SP10	-0.32815376	rejected	C3	0.2548625	
SP11	-0.2459655	rejected	C4	0.39149737	
			Ab1	-0.65571696	rejected
			Ab2	-0.65571696	rejected
Mean goodness-of-fit:		0.03684268			

Table5 - Echo results for Supervenience with ability hypothesis.

Supervenience with the ability hypothesis still falls short of having positive activations. The incompatibility with three out of four intuitions weights too heavily.

Supervenience + Ability Hyp.	Consistency (with sciences)	✓
	No Rejections	x
	Mean goodness-of-fit > 0.03	✓

2.3.5 Supervenience Physicalism with Ability Hypothesis and Conceivability (Intuition) Production

SP1 - There is a fundamental sense in which there is only one kind of stuff in reality.
SP2 - There is physical reality.
SP3 - There is phenomenal reality.
SP4 - Phenomenal reality is not wholly physical reality.
SP5 - Physical reality is not wholly phenomenal reality.
SP6 - In the actual world w it is true that: for any possible world w* if w* is a physical duplicate of w, then w* is a duplicate of w simpliciter.
SP7 - Any instantiated property is necessitated by some instantiated physical properties.
SP8 - Any phenomenal property is supervenient on physical properties.
SP9 - On the fundamental level there are only physical properties.
SP10 - All facts are (fully) determined by facts about ultimates.

SP11 - Experiential (phenomenal) events are caused by and supervene on physical events.
Ab1 - Theories about physical reality cannot wholly teach us about phenomenal reality.
Ab2 - In experiencing phenomenal reality we acquire an ability to recognize, imagine and predict ones behaviour by means of imaginative experiments. This ability cannot be exhaustively taught by a physical theory.
Lb1 - "If there are no substantive a priori ties between two concepts, then it is possible to conjoin either of the concepts with the negation of the other without producing an inconsistency. That is to say, it is possible to use the concepts to conceive coherently of situations […] in which there are particulars that fall under one of the concepts but do not fall under the other concept. If it is within a subject's power to conceive coherently of its being the case that p, then, unless the subject has a good a posteriori reason to think that the proposition that p could turn out to be inconsistent with a set of propositions that have been shown independently to be necessary, the subject will come to believe that it is possible - or at any rate, come to believe that it is prima facie possible - for it to be the case that p." (Hill, 1997, p. 75)
Lb2 - "If it is within a subject's power to conceive coherently of its being the case that p, then, unless the subject has a good a posteriori reason to think that the proposition that p could turn out to be inconsistent with a set of propositions that have been shown independently to be necessary, the subject will come to believe that it is possible - or at any rate, come to believe that it is prima facie possible - for it to be the case that p." (Hill, 1997, p. 75)
Lb3 - There is no substantive a priori tie between physical reality and phenomenal reality.
Lb4 - We lack a posteriori reasons that those concepts are necessarily coextensive.

Supervenience Physicalism with Ability Hypothesis and conceivability production, positive constraints:

explain ((SP1 SP2 SP3) SP4)	;explaining phenomenal is not wholly physical
explain ((SP1 SP2 SP3) SP5)	;explaining physical is not wholly phenomenal
explain ((SP2 SP7) SP3)	;explaining physical - phenomenal relation
explain ((SP2 SP3 SP7) SP8)	;explaining supervenience
explain ((SP2 SP3 SP8) SP6)	;explaining global identity
explain ((SP2 SP7) SP9)	;explaining fundamental ontology
explain ((SP2 SP9) SP10)	;explaining smallism
explain ((SP2 SP3 SP8) SP11)	;explaining causation of phenomenal consciousness
explain ((SP2 SP11) SP3)	;explaining phenomenal reality
explain ((SP7 SP8) C3)	;explaining observed correlation
explain ((SP9) C4)	;explaining no mental properties at ultimates
explain ((SP4 SP5) Ab1)	;connecting ability hypothesis
explain ((SP4 SP5) Ab2)	;connecting ability hypothesis
explain ((SP6 Ab1 Ab2) I1)	;explaining Mary's learning
explain ((SP7 Ab1 Ab2) I1)	;explaining Mary's learning

explain ((SP8 Ab1 Ab2) I1)	;explaining Mary's learning
explain ((SP2 SP8) C1)	;explaining differences by supervenience
explain ((SP2) C2)	;explaining pineal gland by physical reality
explain ((SP2 SP3 SP4 SP5) Lb3)	;explaining our lack of a posteriori reason for coextensiveness
explain ((SP2 SP3 SP4 SP5) Lb4)	;explaining our lack of a priori reason for coextensiveness
explain ((Lb1 Lb2 Lb3 Lb4) I2)	;explaining I2
explain ((Lb1 Lb2 Lb3 Lb4) I2)	;explaining I3
explain ((Lb1 Lb2 Lb3 Lb4) I2)	;explaining I4
explain ((SP2) E)	;explaining the remainder

Negative constraints: None.

The network settled by cycle 102 with the following activation results:

SP1	0.56015694		E	0.6231477
SP2	0.87332094			
SP3	0.8287173		I1	0.71057546
SP4	0.7417643		I2	0.7111159
SP5	0.7417643		I3	0.49498945
SP6	0.2709629		I4	0.49498945
SP7	0.69050485			
SP8	0.76357687		C1	0.6170801
SP9	0.6085291		C2	0.6231477
SP10	0.3679386		C3	0.60653234
SP11	0.33364034		C4	0.5918739
Ip1	0.610666		Ab1	0.6770242
Ip2	0.610666		Ab2	0.6770242
Ip3	0.68109864			
Ip4	0.68109864			
Mean goodness-of-fit:		0.03868274		

Table6 - Echo results for Supervenience with ability hypothesis and conceivability production.

Supervenience with the ability hypothesis and an explanation for the false intuition fares well. No contradictions and rejections and an ok goodness-of-fit.

Supervenience + Ability Hyp. + Conceiv.	Consistency (with sciences)	✓
	No Rejections	✓
	Mean goodness-of-fit > 0.03	✓

2.3.6 Fundamental Duality Monism

FDM1 - There is a fundamental sense in which there is only one kind of stuff in reality.
FDM2 - There is experiential reality, experiential being.
FDM3 - There cannot be experience without a subject of experience.
FDM4 - There cannot be a subject of experience without experience.
FDM5 - There is physical reality, physical being.
FDM6 - All reality (or being) is either experiential or physical.
FDM7 - There is only one kind of stuff and it has both experiential reality and physical reality and no other kind of reality.
FDM8 - There is only physical reality.
FDM9 - All experiential reality is physical reality.
FDM10 - All physical being involves experiential being (given that Physicalism is true).
FDM11 - At least some ultimates are experience-involving.
FDM12 - All physical reality is (at bottom) the same kind of stuff.
FDM13 - Experiential reality cannot be wholly non-experiential reality.
FDM14 - Non-experiential reality cannot be wholly experiential reality.
FDM15 - Experiential reality cannot possibly emerge from wholly and utterly non-experiential reality.
FDM16 - Non-experiential reality cannot possibly emerge from wholly and utterly experiential reality.
FDM17 - All facts are (fully) determined by facts about ultimates.
FDM18 - There is only one fundamental kind of stuff. It (all of it) has both experiential reality and non-experiential reality. There is no other kind of reality.
FDM19 - Experiential (phenomenal) events are caused by events involving the experiential reality of ultimates.

Fundamental Duality Monism, positive constraints:

```
explain ((FDM2 FDM5) FDM6)                                  ;explaining reality as mental or physical
explain ((FDM1 FDM2 FDM5 FDM6 FDM13 FDM14) FDM7) ;explaining mental property
                                                            dualism
explain ((FDM1 FDM2 FDM5 FDM8) FDM9)                        ;explaining experential as part of the
physical natural kind
explain ((FDM2 FDM8 FDM7 FDM9) FDM10)                       ;explaining panpsychism
explain ((FDM2) FDM3)                                       ;explaining necessary subject
explain ((FDM2) FDM4)                                       ;explaining necessary experience
explain ((FDM9 FDM10) FDM11)                                ;explaining experience at micro-level
explain ((FDM1) FDM12)                                      ;explaining physical(natural kind) monism
explain ((FDM2 FDM5) FDM13)                                 ;explaining mental =/= physical
explain ((FDM2 FDM5) FDM14)                                 ;explaining physical =/= mental
explain ((FDM2 FDM13) FDM15)                                ;explaining the impossibility of brute
                                                            emergence of experience
explain ((FDM2 FDM14) FDM16)                                ;explaining the impossibility of brute
                                                            emergence of non-experience
```

explain ((FDM1 FDM7) FDM18) ;explaining duality monism on fundamental level
explain ((FDM11 FDM17 FDM15 FDM18) FDM19) ;explaining causation of experience by ultimates
explain ((FDM2 FDM5 FDM13 FDM14 FDM19) I1) ;explaining Mary's learning
explain ((FDM2 FDM5 FDM13 FDM14) I2) ;explaining physical but not phenomenal duplicates
explain ((FDM2 FDM5 FDM13 FDM14) I3) ;explaining phenomenal but not physical duplicates
explain ((FDM2 FDM5 FDM13 FDM14) I4) ;explaining behavioural and functional but not phenomenal duplicates
explain ((FDM5) E) ;explaining remainder by physical reality
explain ((FDM5) C2) ;explaining pineal gland functions by physical reality
explain ((FDM2 FDM5 FDM13 FDM14) C1) ;explaining differences in physics, without differences in types of mental states
explain ((FDM17 FDM18 FDM19) C3) ;explaining mental-physical correlation

Negative constraints:

contradict ((FDM11) C4) ;no evidence for mental properties of ultimates

The network settled by cycle 207 with the following activation results:

FDM1	0.65012485		E	0.4051888
FDM2	0.8827709			
FDM3	0.4097677		I1	0.61450386
FDM4	0.4097677		I2	0.6205313
FDM5	0.8660698		I3	0.6205313
FDM6	0.56329626		I4	0.6205313
FDM7	0.60662013			
FDM8	0.51764		C1	0.6205313
FDM9	0.5695923		C2	0.6223601
FDM10	0.4524424		C3	0.58154154
FDM11	0.24819244		C4	0.40910965
FDM12	0.33871648			
FDM13	0.82496095			
FDM14	0.8238266			
FDM15	0.5071331			
FDM16	0.4016376			
FDM17	0.4510452			
FDM18	0.5576695			
FDM19	0.5806501			
Mean goodness-of-fit:		0.033380587		

Table7 - Echo results for Fundamental Duality Monism.

Fundamental Duality Monism fits all criteria but pragmatic consistency with science-based data in C4.

Fundamental Duality Monism	Consistency (with sciences)	x
	No Rejections	✓
	Mean goodness-of-fit > 0.03	✓

2.4 What we achieved: Improving the Filter, 'Supertheories' and Inference to the Loveliest Explanation

The method demonstrated in these examples can rule out inadequate theories and reduces the pool of viable explanation candidates. However, we will still be left with more than a single solution. We would like to find one privileged explanation candidate – the one that has the best shot at being an actual explanation. Following this aim, a natural step is to find additional science based propositions, or agreed upon intuitions and filter again. This definitely rules out further theories. Unfortunately, this is only possible under certain circumstances. Explanation candidates are, like explanations in general, sensitive to their description. Two of them can be alike in their principles, but vary in vagueness. The interactive dualism in Descartes's fashion which we filtered had concrete claims about the mind and body interaction, but there is an explanation candidate that is silent on the nature of this very interaction. Such a vague interactive dualism has less apparent contradiction. The fundamental duality monism of Strawson involves claims about smallism. Nevertheless, one could posit a version without smallism that fares better in the coherence tests because it does not contradict C4. Furthermore, we find a general pattern that for any theory which passes our filter mechanism, a more general or vague version passes as well.

A second and related problem comes with most general explanation candidates. Consider the hypothesis that some physical fact causes phenomenal consciousness, with no specifications how it does. This sort of explanation candidate will be consistent with nearly any evidence and therefore cannot be ruled out by our filter mechanism. Hence, any improvement of the filter is not sufficient to select a single privileged theory.

One way to cope with the plurality of theories is to combine them into a single one. Let us suppose a version of property dualism passes our filter.[54] Now take this theory and supervenience physicalism (with ability hypothesis and conceivability production) together and build a disjunction out of them. We may claim that the solution to the hard problem of consciousness in that case would be:

$$((\text{Property Dualism}) \vee (\text{Supervenience Physicalism} + \text{Abil.} + \text{Conceiv.}))$$

This is not completely satisfying, because our explanation interest is barely met with such a diverse disjunction. To give an analogy: if we want to know which way to go at a crossroads, we will not be accepting an answer “left or right”.

The more promising approach is the consideration of more explanatory virtues. We can succeed with a further inference to the loveliest explanation. For this purpose, we pick the one in the remaining pool of explanation candidates, which fits our explanation interest best and explains in the most detailed way. The loveliest explanation is one that, if it is true, provides the most understanding. (Lipton, 2004, p. 59) It gives the most information about the causal history or mechanism. It is different from the likeliest explanation and rarely coextensive. The most vague and general statements are very likely to be true, but the concrete explanations, that may provide deeper understanding, are easier falsified. Remember the concrete dualism failing in circumstances in which a vague one would have been fine. Inference to the loveliest explanation is frequently employed, but their relation to likelihood is not apparent.[55] Any such connection would be a miracle. This is known as ‘Voltaire’s objection’. (Lipton, 2004, p. 142)

One strategy to argue for a reliability of such an inference is to point to this frequency of using it. If it is true that our inductive practices are guided by loveliness and our inductive practices are reliable in safe contexts[56], then inference to the loveliest explanation is reliable. Lipton (2004) adopts this strategy, when he argues that no other description fits our

[54] One could, for example, add propositions that explain why mental properties do not show up in experiments of natural sciences to fundamental duality monism.
[55] Coherence criteria like we used in the filter mechanism are in a sense already consideration of loveliness.
[56] An unsafe context is one that makes a normally reliable belief-forming process unreliable. E.g. Darkness for visual perception of objects.

inductive practises as well or better and "[…] as you believe that our actual practices are reliable, you will tend to discount Voltaire's objection" (Lipton, 2004, p. 148).

Even with the relation from loveliness to likelihood granted, it is difficult to compare how much explanation candidates would explain if they were true. A general rule should be to privilege more detailed versions of similar theories. If these theories are true, they certainly provide more understanding. Similar theories should be defined as sharing a set of proposition and differ only in ways that themselves imply those propositions. An example with interactive dualism:

One version uses ID1-ID12, the other only ID1-ID10. ID11 with ID12 implies ID9 and ID10.

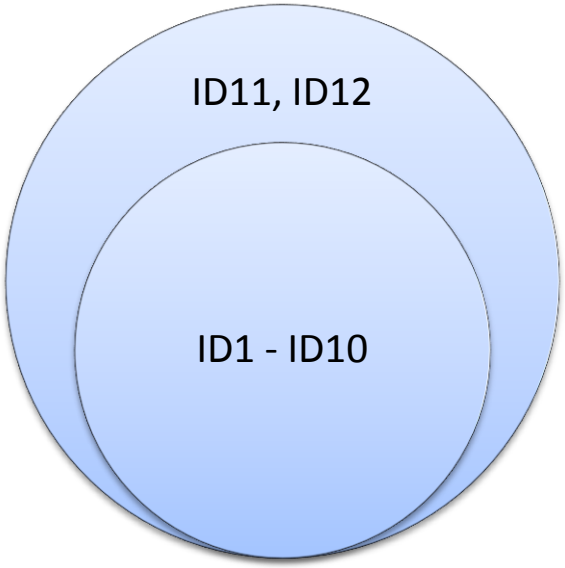

Figure 2 - Vague and Concrete Interactive Dualism.

This rule is not sufficient to privilege a single theory, but it removes redundant ones. We cannot provide a way to single out an explanation candidate, but the methods exemplified are able to reduce the pool of explanation candidates substantially. This is not a total success for us, but nevertheless, it is a success. We may achieve an even better result if more suitable ways to compare loveliness are found. We can also advance the method if the data used for filtering is enhanced in a way that makes our preference of the more precise theories sufficient for singling out an explanation candidate. However, this is beyond the

scope of this thesis and left for future work. For this inquiry, we have to settle on a disjunction of most concrete versions of explanation candidates.

For the rest of this thesis, we will rather investigate the epistemic status achieved by our method.

3 Epistemic Status

In the beginning of the thesis, we set the bar for a proper explanation very high. Now is the time to evaluate our method with regard to this demanding account of explanation. We want knowledge about the causal history of phenomenal consciousness and therefore need belief, justification, truth, no defeaters and no gettier cases. The belief condition needs no discussion.

3.1 Justification as Having Reasons

With an internalist account of justification, we need to evaluate whether we fulfilled our epistemic duties with our inference. Supposed the method privileges one explanation candidate: is it rational to believe this explanation candidate? There are good reasons for claiming it is. The theory fulfils our explanation interest, it coheres with our beliefs and there is no accessible theory that fares better. The data for our coherence measurement is well-established and produced, considering our epistemic duties. We made no mistake regarding our intuitions[57] and rely on scientific criteria. We have no defeaters for our evidence. The only potential problem for the internalist might be the access to the coherence measurement by neural networks. Nevertheless, it is in principle possible to go into every step of calculation done by the network, if one demands this kind of access. In general, internalist justification seems to be easily matched by our method.

3.2 Justification as Reliability

Externalist justification might be more challenging. We will limit ourselves to reliabilism for the sake of simplicity. Is our inference to the best explanation employed reliable?

The first way to answer this question is by analogy with other inferences to the best explanation. We can provide arguments that such inferences are in general reliable. We already mentioned Thagard's argument in favour of scientific inference to the best explanation (Thagard, 2012, p. 46). This is the most common way to argue for such

[57] In any case it is very difficult to imagine how such a mistake would look like.

inferences. We use them in general and we find them successful. Hence, they are most likely reliable and inferred theories probably true. Otherwise it would be a miracle how they would predict accurately, as Hilary Putnam (1975) phrased it. The common objection is a meta-induction[58] that tries to show how past theories, with somewhat successful predictions, were simply false. Therefore we should not infer our successful theories nowadays as true. On the same token, the method producing those theories is not reliable.

The more severe problem for a no-miracle argument is the circularity involved. The claim that our inferences to the best explanation are reliable is itself an inference to the best explanation (Lipton, 2004, p. 186). A skeptic to our inferences may object it solely on this reasoning.

Both those objections, the circularity and the existence of false theories, point to an interesting similarity. Those are common arguments for a skeptic regarding induction in general, not only for the inductive practice of inferring to the best explanation. Induction can be justified by inductive inferences and such inferences can also lead into false conclusions. The success of inductive inferences in general is hardly explicable. That leaves inference to the best explanation in no worse position than any other account of induction. The miracle of reliable inference to the best explanation reduces to Hume's problem, as Lipton (2004) observed. Whoever denies its reliability has to let any induction go and is left with very few justified beliefs. This, however, does not prove our method's reliability. Nevertheless, it shows that we need to ascribe reliability (ceteris paribus) if we believe that induction can provide justification and if we believe that sciences can produce justification, we ought to accept reliable induction.

The second option to explain the reliability is the following: Suppose we start with a complete pool of all explanation candidates. One of them is true, the others are false. Then we rule out inadequate theories by our method. With this process, we increase the probability of each single remaining explanation candidate to be true. If we reach the point of a single explanation candidate left, we found our true explanation. Now this procedure has a couple of problems that need to be addressed. First, we have to make sure that we

[58] Most famous by Larry Laudon (1984).

cannot rule out a true explanation. This requirement is only met if our data is true. The bad news is that we do not know whether our data is actually true, we may at best confirm that it is reliably formed. Nevertheless, this does not influence the reliability of our inference, it just emphasizes that it is a belief-dependent process. The more dangerous worry occurring when using coherence as a tool to rule out inadequate theories is the truth conduciveness of coherence.[59] A computational foundherentism like we used here fares better than most coherentims in this case. The pure coherentist cannot fall back to other sources of knowledge to establish a connection to truth. The data Thagard and we use in our coherence measurement, is independently, partially justified and starts with an initial activation level. The coherence is established in relation to some very likely propositions. Olsson (2005, p. 166) claims that this is not enough to produce true conclusions in general. What has to be added is at least the requirement to have such observational input and different independent evidence (data). Both are satisfied already in the basic examples we exemplified.

Secondly, we do not know whether we have a complete pool of all explanation candidates and we probably have not. This has at least two impacts:

1) We cannot ascribe certain truth to the single explanation candidate left (if we are able to single out a theory), because there may be other candidates not considered. However, we still provide justification because with every theory ruled out, the other explanation candidates gain higher probability of being true.
2) We might not have enough justification to pass the threshold for knowledge. Increased likelihood is one thing, but being probable enough for reliability is another. With the uncertainty of unacknowledged explanation candidate, we never know how probable a theory is. We only have justification relative to the pool of explanation candidates. However, we need to be satisfied with this kind of justification. The justificatory bar is set by the initial pool of candidates. Our inference to the best explanation is in no worse shape than any scientific theory regarding underconsideration. It is true for any scientific theory, that there may be more, not

[59] Peter Klein and Ted Warfield (1994) tried to show that it is not truth conducive and started the so-called 'Analysis debate'.

actually considered alternatives. If any scientific theory is able to pass the justification threshold for knowledge, then our inferred explanation-candidate does so as well.

We have given no positive argument in favour of justification provided by our method. Nevertheless, the justification of our inference to the best explanation is analogue to that of induction and general underconsideration. And while there are reasons to be skeptic towards both induction and scientific theories, we do rely on both of them in practice. That alone has a positive force in favour of using our inference for the hard problem of consciousness.

3.3 Removed Defeaters

In our definition of explanation, we employed a no-defeater condition, commonly attributed to knowledge. This requirement is not fulfilled solely with our method, but in the process we did something relevant for the removal of defeaters. Like a belief that not-p (or evidence supporting not-p) is a defeater for p, two incompatible explanation candidates are defeaters for each other. If any incompatible theory is a defeater for an explanation candidate, it lowers the credence we ought to give to that candidate. When our procedure rules out theories, we get rid of defeaters and thereby raise the justification of the remaining explanation candidates. In an ideal situation, we single out one privileged explanation candidate and thereby ensure that no other theory provides a defeater in competing with it.

3.4 Truth

Even if we find the one most promising explanation candidate based on our pool of theories, we do not know whether it is true or not. Although it is a reliable belief-forming process, this does not make the resulting belief true with certainty. Knowledge, and therefore explanation, requires truth, but it does not require that we know of its being true. The best explanation candidate is only a real explanation if it also happens to be true. It is always by the favour of nature that one really explains something, as we might adapt from Wittgenstein (1969, §505).

4 Conclusion

This inquiry into inference to the best explanation as a method for solving the hard problem of consciousness has been an attempt at both articulation of said method and defence. We started by fixating the problem of phenomenal consciousness and interrogate what is needed for an explanation of the existence of qualia. We found a difference between actual explanations and mere explanation candidates and stressed what conditions real explanations need to meet. We thereby chose a causal account of explanation in favour of a covering-law model.

We elucidated on the Mysterian Challenge as a major motivation for finding a suitable method to grasp an explanation of there being phenomenal consciousness in the world. It has been argued that inference to the best explanation is able to help us against the mysterian denial of any accessible answer to the hard problem.

After we defined our goal we proceeded to look at different ways to spell out the slogan of inference to the best explanation for the hard problem. We discussed the initial pool of explanation candidates and provided a couple of examples how a theory should enter this pool. We then hinged our procedure of ruling out inadequate explanation candidates on coherence. Using coherence as a pivotal source of justification, we had to identify different ways of capturing coherence. We dismissed Ewing's, C.I. Lewis' and Bonjour's accounts and opted for Thagard's computational account of coherence as constraint satisfaction. This enabled us to calculate coherence of an explanation candidate with additional data. The data is independently justified. We found that initially justified data in science-based propositions and intuitions (both being to some degree dependent on each other). The intuitions used were selected by predominant agreement in philosophical discussion. We showed that the only other alternative, experimental philosophy, is currently not viable.

We inquired how to set up the coherence-network with those data and exemplified the method by testing several explanation candidates: Interactive dualism; Type Identity Theory, Supervenience Physicalism; Supervenience Physicalism with Ability Hypothesis; Supervenience Physicalism with Ability Hypothesis and clauses for cases of mistaken conceivability; and Fundamental Duality Monism.

We provided a short analysis of the result of a method based on our examples. The method will not provide a single privileged explanation candidate, but rather a bundle of theories. Even in the best case there will be a variety of explanation candidates, each differing in vagueness. We claimed that it is rational to pick the most precise theory available.

The inquiry of the justification provided by our inference to the best explanation showed, that an internalist justification is easily met, whereas an externalist justification is more difficult. Building on an analysis by Lipton (2004), we pointed out that the reliability of the kind of inference in question is in no worse shape than induction in general. Both can only be argued for by reasoning circularly. Arguments for the reliability of inference to the best explanation usually employ the very same inference, as we saw in Putnam's miracle argument.

We showed that, in an ideal case, our way of inferring by ruling out inadequate explanation candidates is truth-conducive, but met a problem of underconsideration. We do not know whether our initial pool of theories is complete, therefore we cannot know whether we actually have a true explanation candidate. This problem is similar to underconsideration of scientific theories in general. If we believe that scientific theories can nevertheless be sufficiently justified, our method justifies sufficiently as well.

In addition, we do not need to know the truth of the explanation candidate for it to be a real explanation. It just needs to fulfil the conditions of believe, justification, truth, no epistemic luck and no defeaters. If nature is favourable, our inference to the best explanation can lead us to a theory that fulfils all of those. Of course, there is still a lot to do for finding the answer to the hard problem with this method. We need a complete pool of explanation candidates and start the coherence measurement with full extent of data. The few examples we provided are barely the tip of the iceberg.

As usually with inferences to the best explanation, it will certainly not be a procedure to do once and be done with. Everytime a new explanation candidate is proposed we need to start all over again. Nevertheless, we found a very promising approach for solving the hard problem of consciousness.

5 Bibliography

Armstrong, D. M., 1961. *Perception and the physical world.* London: Routledge.

Armstrong, D. M., 1968. *A materialist theory of mind.* London: Routledge.

Bacchus, F. & van Beek, P., 1998. On the conversion between non-binary and binary constraint satisfaction problems. In: *Proceedings of the National Conference on Artificial Intelligence (AAAI-98).* Menlo Park: AAAI Press, pp. 311-318.

Bateson, G., 2000 (1970). Form, Substance, and Difference. In: *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* Chicago: The University of Chicago Press, pp. 454-471.

Bayne, T. & Montague, M., 2011. Cognitive Phenomenology: An Introduction. In: T. Bayne & M. Montague, Hrsg. *Cognitive Phenomenology.* New York: Oxford University Press, pp. 1-34.

Bilgrami, A. & Rovane, C., 2005. Mind, language, and the limits of inquiry. In: *The Cambridge Companion to Chomsky.* Cambridge: Cambridge University Press, pp. 181-203.

Blackburn, S., 1990. Filling in Space. *Analysis*, March, 50(2), pp. 62-65.

Block, N., 1990. Inverted Earth. *Philosophical Perspectives*, Issue 4, pp. 53-79.

Block, N., 1997. On a Confusion about a Function of Consciousness. In: N. Block, O. Flanagan & G. Güzeldere, Hrsg. Cambridge, MA: MIT Press, pp. 375-415.

BonJour, L., 1985. *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press.

Byrne, A., 2010. *"Inverted Qualia" ,The Stanford Encyclopedia of Philosophy (Spring 2010 Edition). Edward N. Zalta (ed.).* [Online]
Available at: http://plato.stanford.edu/archives/spr2010/entries/qualia-inverted
[Accessed April 15, 2013].

Campbell, N., 2004. Generalizing Qualia Inversion. *Erkenntnis*, January, 60(1), pp. 27-34.

Chalmers, D. J., 1995. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), pp. 200-219.

Chalmers, D. J., 1996. *The Conscious Mind.* Oxford: Oxford University Press..

Chalmers, D. J., 2002. Does Conceivability Entail Possibility?. In: T. S. Gendler & J. Hawthorne, Hrsg. *Conceivability and Possibility.* Oxford: Oxford University Press, pp. 145-200.

Chomsky, N., 1976. Problems and Mysteries in the Study of Human Language. In: A. Kasher, Hrsg. *Language in Focus: Foundations, Methos and Systems. Essays in Memory of Yehoshua Bar-Hillel.* Dordrecht: D. Reidel Publishing, pp. 281-357.

Chomsky, N., 2000. Naturalism and dualism in the study of language and mind. In: *New Horizons in the Study of Language and Mind.* Cambridge: Cambridge University Press, pp. 75-105.

Churchland, P., 1985. Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy*, January, 82(1), pp. 8-28.

Coleman, S., 2006. Being Realistic. In: A. Freeman, Hrsg. *Consciousness and its place in nature.* Exeter: Imprint Academic, pp. 40-52.

Conee, E., 1994. Phenomenal Knowledge. *Australasian Journal of Philosophy*, June, 72(2), pp. 136-150.

Crane, T. & Mellor, D. H., 1990. There is No Question of Physicalism. *Mind*, April, pp. 185-206.

Dennett, D. C., 1995. The Unimagined Preposterousness of Zombies: commentary on T. Moody, O. Flanagan and T. Polger. *Journal of Consciousness Studies*, 2(4), pp. 322-326.

Descartes, R., 1985 (1649). The Passions of the Soul. In: *The Philosophical Writings of Descartes Volume I.* Cambridge: Cambridge University Press, pp. 325-404.

Descartes, R., 2008 (1641). Meditations on First Philosophy. In: *Meditation on First Philosophy with selections from the objections and replies.* Oxford: Oxford University Press, pp. 1-64.

Eddington, A. S., 1929. *The nature of the physical world.* Cambridge: Cambridge University Press.

Ewing, A. C., 1961 (1934). *Idealism: A Critical Survey.* 3. Hrsg. London: Methuen & Co.

Fetzer, J. H., 1974. A Single Case Propensity Theory of Explanation. October, 28(2), pp. 171-198.

Fitch, F. B., 1963. A Logical Analysis of Some Value Concepts. *The Journal of Symbolic Logic*, June, 28(2), pp. 135-142.

Flanagan, O., 1992. *Consciousness Reconsidered.* Cambridge, MA: MIT Press.

Gettier, E. L., 1963. Is Justified True Belief Knowledge?. *Analysis*, June, 23(6), pp. 121-123.

Gödel, K., 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. *I. Monatshefte für Mathematik und Physik*, Issue 38, pp. 173-198.

Goff, P., 2010. Ghosts and Sparse Properties: Why Physicalists Have More Fear from Ghosts than Zombies. *Philosophy and Phenomenological Research*, July, 81(1), pp. 119-139.

Goff, P., 2012. Does Mary know I experience plus rather than quus? A new hard problem. *Philosophical Studies*, September, 160(2), pp. 223-235.

Goldman, A. I., 1987. Cognitive Science and Methaphysics. *Journal of Philosophy*, Issue 84, pp. 537-544.

Goldman, A. I., 1992. Cognition and Modal Metaphysics. In: *Liaisons.* Cambridge, MA: MIT Press, pp. 49-68.

Goldman, A. I., 1993. The Psychology of Folk Psychology. *Behavioral and Brain Sciences*, Issue 16, pp. 15-28.

Goldman, A. I., 2007. Philosophical Intuitions: Their Target, Their Source, and their Epistemic Status. *Grazer Philosophische Studien*, Issue 74, pp. 1-26.

Harman, G., 1977. *The Nature of Morality.* New York: Oxford University Press.

Harman, G., 1990. The Intrinsic Quality of Experience. *Philosophical Perspectives*, Band 4, pp. 31-52.

Hempel, C. G., 1942. The Function of General Laws in History. *The Journal of Philosophy*, 15 January, pp. 35-48.

Hempel, C. G., 2001a (1962). Deductive-Nomological versus Statistical Explanation. In: J. H. Fetzer, Hrsg. *The Philosophy of Carl G. Hempel.* New York: Oxford University Press, pp. 87-145.

Hempel, C. G., 2001b (1963). Explanation and Prediction by Covering Laws. In: J. H. Fetzer, Hrsg. *The Philosophy of Carl G. Hempel.* New York: Oxford University Press, pp. 69-86.

Hempel, C. G., 2001c (1977). Postscript 1976: More Recent Ideas on the Problem of Statistical Explanation. In: J. H. Fetzer, Hrsg. *The Philosophy of Carl G. Hempel.* New York: Oxford University Press, pp. 165-187.

Hempel, C. G. & Oppenheim, P., 1948. Studies in the Logic of Explanation. *Philosophy of Science*, April, 15(2), pp. 135-175.

Hilary, P., 1975. *Mathematics, Matter, and Method.* Cambridge: Cambridge University Press.

Hill, C. S., 1997. Imaginability, Conceivability, Possiblity and the Mind-Body Problem. *Philosophical Studies*, July, 87(1), pp. 61-85.

Hill, C. S., 2009. *Consciousness.* Cambridge: Cambridge University Press.

Hill, C. S. & Mclaughlin, B. P., 1999. There are Fewer Things in Reality Than are Dreamt of in Chalmers's Philosophy. *Philosophy and Phenomenological Research*, June, 59(2), pp. 445-454.

Hitchcock, C., 1999. Contrastive Explanation and the Demons of Determinism. *British Journal for the Philosophy of Science*, 50(4), pp. 585-612.

Horgan, T., 1984a. Functionalism, Qualia, and the Inverted Spectrum. *Philosophy and Phenomenological Research*, June, 44(4), pp. 453-469.

Horgan, T., 1984b. Jackson on Physical Information and Qualia. *The Philosophical Quarterly*, April, 34(135), pp. 147-152.

Horst, S., 2007. *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science.* New York: Oxford University Press.

Huebner, B., 2010. Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?. *Phenomenology and the Cognitive Sciences*, Issue 9, pp. 133-155.

Huebner, B., Bruno, M. & Sarkissian, H., 2010. What Does the Nation of China Think About Phenomenal States?. *Review of Philosophy and Psychology*, Issue 1, pp. 225-243.

Huemer, M., 2007. Compassionate Phenomenal Conservatism. *Philosophy and Phenomenological Research*, January, 74(1), pp. 30-55.

Hume, D., 2000 (1739). *A Treatise of Human Nature (ed. by Norton, David F. and Norton, Mary J.).* Oxford: Oxford University Press.

Jackson, F., 1982. Epiphenomenal Qualia. *The Philosophical Quarterly*, April, 32(127), pp. 127-136.

Kelley, C. M. & Lindsay, S. D., 1993. Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, Issue 32, pp. 1-24.

Klein, P., 2011. *"Skepticism", The Stanford Encyclopedia of Philosophy (Summer 2011 Edition).* [Online]
Available at: http://plato.stanford.edu/archives/sum2011/entries/skepticism/
[Accessed January 15, 2013].

Klein, P. & Warfield, T. A., 1994. What Price Coherence?. *Analysis*, July, 54(3), pp. 129-1323.

Knobe, J. & Prinz, J., 2008. Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, Issue 7, pp. 67-83.

Lackey, J., 2008. *Learning from Words: Testimony as a Source of Knowledge.* Oxford: Oxford University Press.

Laudan, L., 1984. A Confutation of Convergent Realism. In: J. Leplin, Hrsg. *Scientific Realism.* Berkeley: University of California Press, pp. 218-249.

Lerner, A. B., Case, J. D. & Takahashi, Y., 1960. Isolation of Melatonin and 5-Methoxyindole-3-acetic from Bovine Pineal Glands. *The Journal of Biological Chemistry*, July, 235(7), pp. 1992-1997.

Levine, J., 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly*, Issue 64, pp. 354-361.

Lewis, C. I., 1946. *An Analysis of Knowledge and Valuation.* LaSalle: Open Court.

Lewis, D., 1983. Postscript to "Mad Pain and Martian Pain". In: *Philosophical Papers Vol. I.* Oxford: Oxford University Press, pp. 130-132.

Lewis, D., 1986. Causal explanation. In: D. Lewis, ed. *Philosophical Papers Vol.II.* Oxford: Oxford University Press, pp. 214-240.

Lipton, P., 2004. *Inference to the best explanation.* 2. ed. London: Routledge.

Loar, B., 1997. Phenomenal States (second version). In: *The Nature of Consciousness: Philosophical Debates.* Cambridge, MA: MIT Press, pp. 597-617.

Loar, B., 1999. David Chalmers's The Conscious Mind: Review on The Conscious Mind by David Chalmers. *Philosophy and Phenomenological Research*, June, 59(2), pp. 465-472.

McGinn, C., 1993a (1989). Can We Solve the Mind-Body Problem?. In: *The Problem of Consciousness.* Oxford: Blackwell, pp. 1-22.

McGinn, C., 1993b. *The Problem of Consciousness.* Oxford: Blackwell.

McGinn, C., 1999. *The Mysterious Flame.* New York: Basic Books.

Melnyk, A., 2003. *A Physicalist Manifesto.* Cambridge: Cambridge University Press.

Millgram, E., 2000. The Price of the Ticket. *Journal of Philosophy*, February, 97(2), pp. 82-93.

Nagel, T., 1974. What is it like to be a bat?. *Philosophical Review*, October, 83(4), pp. 435-450.

Nagel, T., 1986. *The View from Nowhere.* New York: Oxford University Press.

Nemirow, L., 1980. Review of: Mortal Questions by Thomas Nagel. *The Philosophical Review*, July, 89(3), pp. 473-477.

Ney, A., 2008. Defining Physicalism. *Philosophy Compass*, 3(5), pp. 1033-1048.

Olsson, E. J., 2005. *Against Coherence: Truth, Probability, and Justification.* Oxford: Oxford University Press.

Pettit, P., 1993. A Definition of Physicalism. *Analysis*, October, pp. 213-223.

Pust, J., 2000. *Intuitions as Evidence.* New York: Garland Publishing.

Pust, J., 2012. *"Intuition", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition). Edward N. Zalta (ed.).* [Online]
Available at: http://plato.stanford.edu/archives/win2012/entries/intuition
[Accessed February 5, 2013].

Putnam, H., 1975. *Mathematics, Matter, and Method.* Cambridge: Cambridge University Press.

Railton, P., 1978. A Deductive-Nomological Model of Probalistic Explanation. *Philosophy of Science*, June, 45(2), pp. 206-226.

Reiter, R. J., 1980. The Pineal and Its Hormones in the Control of Reproduction in Mammals. *Endocrine Reviews*, 1(2), pp. 109-131.

Roseberg, G., 2004. *A Place for Consciousness.* Oxford: Oxford University Press.

Rosenthal, D. M., 1997. A Theory of Consaciousness. In: N. Block, O. Flanagan & G. Guzeldere, Hrsg. *The Nature of Consciousness: Philosophical Debates.* Cambridge, Mass.: MIT Press, pp. 129-145.

Rowlands, M., 2003. *The Nature of Consciousness.* Cambridge: Cambridge University Press (Virtual publishing).

Russell, B., 1910. Knowledge by acquaintance and knowledge by description. *Proceedings of the Aristotelian Society*, Issue 11, pp. 108-128.

Russell, B., 1912. *The Problems Of Philosophy.* Oxford: Clarendon Press.

Russell, B., 2007 (1927). *Analysis of Matter.* Nottingham: Spokesman.

Seager, W., 2006. The 'Intrinsic Nature' Argument for Panpsychism. In: A. Freeman, Hrsg. *Consciousness and its place in nature.* Exeter: Imprint Academic, pp. 129-145.

Shoemaker, D., 1999. On David Chalmers's the Conscious Mind: Review of The Conscious Mind by David Chalmers. *Philosophy and Phenomenological Research*, June, 59(2), pp. 439-444.

Smart, J. J. C., 1959. Sensations and Brain Processes. *The Philosophical Review*, April, 68(2), pp. 141-156.

Sosa, E., 2007. Intuitions: Their Nature and Epistemic Efficacy. *Grazer Philosophische Studien*, Issue 74, pp. 51-67.

Stich, S., 1975. Logical Form and Natural Language. *Philosophical Studies*, December, 28(6), pp. 397-418.

Stoljar, D., 2006. *Ignorance and Imagination.* Oxford: Oxford University Press.

Stoljar, D., 2010. *Physicalism.* London: Routledge.

Strawson, G., 2006a. Panpsychism? Reply to Commentators with a Celebration of Descartes. In: A. Freeman, Hrsg. *Consciousness and its place in nature.* Exeter: Imprint Academic, pp. 184-280.

Strawson, G., 2006b. Realistic Monism: Why Physicalism entails Panpsychism. In: A. Freeman, Hrsg. *Consciousness and its place in nature.* Exeter: Imprint Academic, pp. 3-31.

Strawson, G., 2008 (2003). Real Materialism. In: *Real Materialism: and other essays.* Oxford: Oxford University Press, pp. 19-51.

Strawson, G., 2010. *Mental Reality.* 2. ed. Cambridge, Massachusetts: The MIT Press.

Stroud, B., 1984. The Problem of the External World. In: *The Significance of Philosophical Scepticism.* Oxford: Oxford University Press, pp. 1-38.

Sytsma, J. & Machery, E., 2010. Two conceptions of subjective experience. *Philosophical Studies*, Issue 151, pp. 299-327.

Thagard, P., 1989. Explanatory coherence. *Behavioral and Brain Sciences*, Issue 12, pp. 435-502.

Thagard, P., 1992. *Conceptual Revolutions.* Princeton: Princeton University Press.

Thagard, P., 2000. *Coherence in Thought and Action.* London: MIT Press.

Thagard, P., 2012. Coherence: The Price Is Right. *The Southern Journal of Philosophy*, March, 50(1), pp. 42-49.

Thagard, P. & Verbeurgt, K., 1998. Coherence as Constraint Satisfaction. *Cognitive Science*, 22(1), pp. 1-24.

Tye, M., 1994. Qualia, Content, and the Inverted Spectrum. *Noûs*, June, 28(2), pp. 159-183.

Tye, M., 1995. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* Cambridge, MA: MIT Press.

Van Gulick, R., 2004. So Many Ways of Saying No to Mary. In: P. Ludlow, Y. Nagasawa & D. Stoljar, Hrsg. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument.* Cambridge, MA: MIT Press, pp. 365-406.

Vogt, K., 2011. *"Ancient Skepticism", The Stanford Encyclopedia of Philosophy (Winter 2011 Edition). Edward N. Zalta (ed.).* [Online]
Available at: http://plato.stanford.edu/archives/win2011/entries/skepticism-ancient
[Accessed: January 15, 2013].

Williamson, T., 2000. *Knowledge and its Limits.* Oxford: Oxford University Press.

Wittgenstein, L., 1969. *On Certainty.* New York: Harper and Row.

Woodward, J., 2003. *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Yablo, S., 1993. Is Conceivability a Guide to Possiblity. *Philosophy and Phenomenological Research,* LIII(1), pp. 1-42.

6 Appendix

Abstract

The thesis discusses inference to the best explanation as a method for finding an answer to the 'hard problem of consciousness' (Chalmers, 1995) - the problem of phenomenal experience. The existence of such mental states is coupled with a mysterian claim to be unexplainable for human beings. The question how we can or should inquiry the mental to be able to refute the mysterian claim arises. We need a procedure to grasp an explanation for the existence of phenomenal consciousness. It is argued that an inference to the best explanation is a method that fulfils this criterion.

Starting with a detailed interrogation of phenomenal consciousness, we elucidate the notion of explanation. We provide deductive-nomological and causal accounts, before we argue for choosing the causal notion. Following this, we discuss mysterian, skeptic claims, which are refuted by pointing to the advantages of inference to the best explanation.

In the next step we state our method of inference. We start with an initial pool of explanation candidates and rule out inadequate ones by consistency and coherence, with science based and intuition based data used as evidence. The discussion involves the problem how an explanation candidate is formulated and what theories should be part of the initial set of candidates. We provide examples of explanation candidates. Furthermore we elucidate the selection process by searching a suitable account of coherence, which is also computable. We show why and which data should be used for evaluation of theories. The selection-process is then exemplified.

Finally we evaluate the method from an epistemological point of view. We argue that it can provide justification and remove defeaters. Thereby we can fulfil important parts of the requirements for a real explanation and reach such an explanation if it happens to be true.

Zusammenfassung

In der vorliegenden Arbeit wird eine Methode zur Lösung des “schwierigen Problems des Bewusstseins” (Chalmers, 1995) untersucht. Dass Vorhandensein phänomenaler Zustände wird verbunden mit der Schwierigkeit diese Existenz zu erklären. Mysterianer behaupten, dass eine solche Erklärung für Menschen überhaupt nicht möglich ist. Damit stellt sich die Frage wie phänomenales Bewusstsein untersucht werden kann und soll. Wir brauchen eine Methode die den mysterianischen Einwänden trotzt und uns zu einer Erklärung leitet. In dieser Arbeit wird die These vertreten, dass ein Schluss auf die Beste Erklärung die geeignete Methode hierfür ist.

Nach einer anfänglichen Erläuterung von phänomenalem Bewusstsein wird der Begriff der Erklärung beleuchtet. Ein deduktiv-nomologisches Modell wird einem kausalen gegenübergestellt und für die Wahl von Letzterem argumentiert. Danach werden die mysterianischen Behauptungen dargelegt und gezeigt wie der Schluss auf die Beste Erklärung eine Antwort bieten kann.

Im Anschluss gilt es die Methode des Schlusses auf die beste Erklärung im Detail auszuführen. Von einer Ausgangsmenge an Erklärungskandidaten werden inadäquate ausgesondert. Dies geschieht über die Kriterien von Konsistenz und Kohärenz mit wissenschaftsbasierten und intuitiv gewonnenen Propositionen. Es wird dargelegt wie und welche Theorien als Erklärungskandidaten in Frage kommen. Ferner wird der Selektionsmechanismus erläutert. Ein angemessener Kohärenzbegriff wird untersucht und der Prozess zur Findung von Propositionen, mit denen die Erklärungskandidaten kohärieren sollen, expliziert. Zur Erfassung der Kohärenz werden neuronale Netzwerke auf Basis des Programms ECHO (Thagard, 1989) verwendet.

Schlussendlich wird die Methode erkenntnistheoretisch evaluiert. Es wird argumentiert, dass tatsächlich Rechtfertigung generiert wird. Zusätzlich werden in diesem Prozess Anfechter eliminiert. Damit sind wichtige Teile von Wissen und damit den Anforderungen einer wirklichen Erklärung erfüllt. Sofern der gerechtfertigte Erklärungskandidat auch tatsächlich wahr ist, erreichen wir eine wirkliche Erklärung.

Curriculum Vitae

Lukas Schwengerer

Education

2010 - 2013 (estimate)	**Master of Arts in Philosophy** at the *University of Vienna, Vienna, Austria* Master Thesis: *Inference to the Best Explanation as a Method for Solving the Hard Problem of Consciousness* (Advisor: Mag. Dr. habil. Sven Bernecker, PhD*)*
2007 - 2010	**Bachelor of Arts in Philosophy** at the *University of Vienna, Vienna, Austria* Bachelor Theses: Defending the Intrinsic Nature Argument (Advisor: Mag. Dr. Michael Blamauer*);* and *Eine Frage der Funktion – Über das Verhältnis Ernst Tugendhats sprachanalytischer Philosophie zu Wittgenstein (*Advisor: ao. Univ.-Prof. Mag. Dr. Elisabeth Nemeth)
1998 - 2006	**Matura** (A-Level) at BRG21/F26, Vienna, Austria

Professional Experience

10/2012 – 02/2013	**Tutor**, University of Vienna, Vienna, Austria *Assisting in a seminar on memory held by Mag. Dr. habil. Sven Bernecker, PhD*
03/2012 – 06/2012	**Student assistant**, University of Vienna, Vienna, Austria Assisting and teaching first year students in introductory courses held by *Mag. Dr. habil. Sven Bernecker, PhD*

10/2011 – 02/2012	**Tutor**, University of Vienna, Vienna, Austria *Assisting and teaching first year students in introductory courses held by ao. Univ.-Prof. Dr. Richard Heinrich*

Voluntary Activities

10/2010 – Current	Active Member of the Vienna Forum for Analytic Philosophy

Lebenslauf

Lukas Schwengerer

Studium und Ausbildung

2010 - 2013 (voraussichtlich)	**Master of Arts in Philosophie** *an der Universität Wien, Wien, Österreich* Masterarbeit: *„Inference to the Best Explanation as a Method for Solving the Hard Problem of Consciousness“* (Betreuer: Mag. Dr. habil. Sven Bernecker, PhD*)*
2007 - 2010	**Bachelor of Arts in Philosophie** *an der Universität Wien, Wien, Österreich* Bachelorarbeiten: „Defending the Intrinsic Nature Argument“ (Betreuer: Mag. Dr. Michael Blamauer*);* und“ *Eine Frage der Funktion – Über das Verhältnis Ernst Tugendhats sprachanalytischer Philosophie zu Wittgenstein“ (*Betreuerin: ao. Univ.-Prof. Mag. Dr. Elisabeth Nemeth)
1998 - 2006	**Matura** *am BRG21/F26, Wien, Österreich*

Berufliche Erfahrung

10/2012 – 02/2013	**Tutor**, Universität Wien, Wien, Österreich *Assistenz bei einem Seminar zum Thema Erinnerung, gehalten von Mag. Dr. habil. Sven Bernecker, PhD*
03/2012 – 06/2012	**Studienassistent**, Universität Wien, Wien, Österreich Unterstützung und Lehre zur STEOP Vorlesung „Einführung in die theoretische Philosophie“, gehalten von *Mag. Dr. habil. Sven Bernecker, PhD*

10/2011 – 02/2012	**Tutor**, Universität Wien, Wien, Österreich Unterstützung und Lehre zur STEOP Vorlesung „Einführung in die theoretische Philosophie“, gehalten von *ao. Univ.-Prof. Dr. Richard Heinrich*

Zusätzliche Aktivitäten

10/2010 – jetzt	Aktives Mitglied im Wiener Forum für analytische Philosophie